Addressing Missing Data and Scalable Optimization for Data-driven Decision Making

by

Dogyoon Song

B.S. in Electrical and Computer Engineering, B.S. in Mathematics, B.S. in Physics
Seoul National University (2013)

S.M. in Electrical Engineering and Computer Science
Massachusetts Institute of Technology (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

June 2021

Signature of Author: _____

Department of Electrical Engineering and Computer Science
May 19, 2021

Certified by: _____

Pablo A. Parrilo
Joseph F. and Nancy P. Keithley Professor of Electrical Engineering and Computer Science
Thesis Co-Supervisor

Certified by: _____

Devavrat Shah
Professor of Electrical Engineering and Computer Science
Thesis Co-Supervisor

Accepted by: _____

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Addressing Missing Data and Scalable Optimization for Data-driven Decision Making

by Dogyoon Song

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2021 in partial fulfillment of
the requirements for the degree of Doctor of Philosophy

Abstract

Data-driven decision making has become a necessary commodity in virtually every domain of human endeavor, fueled by the exponential growth in the availability of data and the rapid increase in our computing power. In principle, if the collected data contain sufficient information, it is possible to build a useful model for making decisions. Nevertheless, there are a few challenges to address to bring it into reality. First, the gathered data can be contaminated by noise, or even by missing values. Second, building a model from data usually involves solving an optimization problem, which may require prohibitively large computational resources. In this thesis, we explore two research directions, motivated by these two challenges.

In the first part of the thesis, we consider statistical learning problems with missing data, and discuss the efficacy of data imputation approaches in predictive modeling tasks. To this end, we first review low-rank matrix completion techniques and establish a novel error analysis for the matrix estimation beyond the traditional mean squared error (Frobenius norm), focusing on the singular value thresholding algorithm. Thereafter, we study two specific predictive problem settings – namely, errors-in-variables regression and Q-learning with thrifty exploration – and argue that the predictions based on the imputed data are typically nearly as accurate as the predictions made when the complete data were available.

In the second part of the thesis, we investigate the tradeoff between the scalability and the quality of optimal solutions in the context of approximate semidefinite programming. Specifically, we ask the question: "how closely can we approximate the set of unit-trace $n \times n$ positive semidefinite (PSD) matrices, denoted by D^n, using at most N number of $k \times k$ PSD constraints?" We show that any set S that approximates D^n within a constant approximation ratio must have superpolynomially large $\boldsymbol{S}_+^k$-extension complexity for all $k = o(n/\log n)$. Our results imply that it is impossible to globally approximate a large-scale PSD cone using only a few, smaller-sized PSD constraints. Therefore, we conclude that local, problem-adaptive techniques are essential to approximate SDPs for enhanced scalability.

Thesis Co-Supervisor: Pablo A. Parrilo
Title: Joseph F. and Nancy P. Keithley Professor of Electrical Engineering and Computer Science

Thesis Co-Supervisor: Devavrat Shah
Title: Professor of Electrical Engineering and Computer Science

Acknowledgements

I have been truly fortunate to have two great advisors, Pablo Parrilo and Devavrat Shah, and would like to thank them for their guidance and support throughout this long journey. I am extremely grateful to them for teaching me how to do research properly, for sharing with me their tremendous intellectual curiosity and enthusiasm, and for creating the perfect environment for me to pursue my interests during my Ph.D. studies. None of the works in this thesis would have been possible without their help, advice, encouragement, and boundless patience. Thanks, Devavrat and Pablo, for everything.

I would also like to express my gratitude to Gregory Wornell for his many words of advice, for his support at various stages during my time at MIT – sometimes as a graduate counselor, sometimes as the course instructor I TA-ed for – and for his service on my thesis committee.

My gratitude extends to my amazing collaborators. I enjoyed working with them, and I learned a lot from them even beyond research. In addition to Pablo and Devavrat, I want to thank Anish Agarwal, Yihua Li, Dennis Shen, Zhi Xu, Yuzhe Yang, and Christina Lee Yu.

More broadly, I am thankful to the people in LIDS, EECS, and MIT at large. I appreciate all of the administrative staff members – with special thanks to Rachel Cohen, Lynne Dell, Janet Fischer, Gracie Gao, Francisco Jaimes, Brian Jones, and Richard Lay – for their assistance in ensuring that everything is running smoothly. I am also grateful for the friendships I have built here at MIT; you guys have made my life here more enjoyable, and have certainly helped me a lot to stay in good mental (and physical) health as a graduate student.

Last but absolutely not least, my heart is full of gratitude for my family, who have always infused my life with unconditional love and support. I am beyond thankful to my grandparents, Hojin Song and Jeongsoon Yoo, and my parents, Younghan Song and Hyejeong Han, for filling my childhood with so much love and happy memories, and for always being my biggest supporters. My deepest gratitude goes to my dear wife and lifelong companion, Jouha Min, and to the most adorable baby girl, Jeanne Min Song, for making me a better person, and for giving me good reasons to finally write my thesis. This dissertation would not have been possible without their love, encouragement, and sacrifices. I owe them everything from the beginning, to the present, and beyond. I dedicate this dissertation to my beloved family.

Funding information: I appreciate the generous support from Samsung Scholarship for funding my research. This thesis research was also supported in part by NSF grants CNS-1523546, CMMI-1462158, CMMI-1634259, a joint project with KAIST (South Korea), and a project funded by KACST (Saudi Arabia).

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

With the exponential growth in the availability of data and the rapid increase in our computing power, data-driven decision-making (DDDM) is now a necessary commodity in virtually every domain of human endeavor. Companies collect and use data to build a predictive model for advantageous marketing strategies and human resource management. Robots and self-driving cars interactively gather and process data from the environment to make real-time decisions to achieve their objectives while fulfilling safety constraints.

Arguably, a typical DDDM procedure consists of three steps as illustrated in Figure 1.1: collect data that contain useful information; build a model from the data; and exploit the model to make decisions. For example, suppose that we are to decide whether we should invest in the stock market today or not. One possible strategy to make such a decision is to collect data about the stock price and economic indicators (GDP growth rate, interest rate, etc.) accumulated so far, build a model that predicts the stock price of tomorrow based on the indicators measured today, and then invest if and only if the predicted price is higher than the price of today. As another example, a self-driving car interactively senses its environment, forecasts the outcomes in its possible trajectories, and chooses one that is certifiably safe and leads to the destination fast. In principle, if the collected data contain sufficient information, it is possible to build a model that can help the decision.

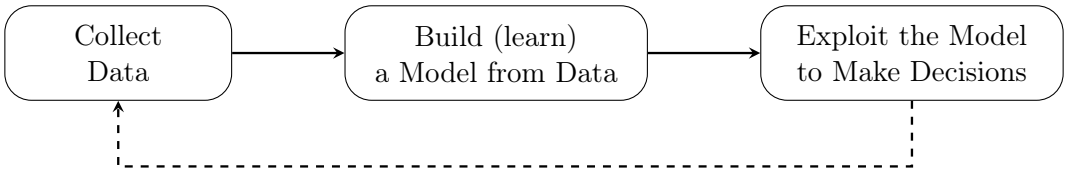


Figure 1.1: A typical process of data-driven decision making.

Nevertheless, there are two fundamental challenges in this process. First, the collected data can have some values missing, which can be caused by the unavailability of some data points, or by the budget constraints in data collection. Missing values in data inhibit the use of standard statistical tools that would be used in building a model if it were not for the

missing data. Second, building a model from data usually involves solving an optimization problem, which may require prohibitively large computational resources.

In this thesis, we explore two research directions motivated by these challenges. In Part II, we study data imputation with low-rank matrix completion as a possible remedy for the missing data in predictive problems. In Part III, we investigate the tradeoff between the computational scalability of optimization problems and the optimality of the approximate solutions in the context of semidefinite programming (SDP).

■ 1.1 Data Imputation for Predictive Modeling

In Part II of this thesis, we consider two types of decision-making problems – supervised learning and reinforcement learning – which we describe here with a toy example. Suppose that a restaurant owner wants to maximize the profit from running the restaurant. As a first step to achieve the goal, they may want to predict the number of guests who will visit tomorrow based on the information available today, e.g., the number of guests who visited, the feedback comments from customers, the weather outside, the number of Likes & reviews the restaurant received on social networking services, etc. They can use the data from the past to learn a regression function in the manner of supervised learning. Moving one step forward, the owner may want to take some actions to increase sales and boost profits, e.g., by running a promotion or changing the menu. These actions will not only affect the immediate reward (profit of tomorrow), but also drift the restaurant to enter into a new state (e.g., reputation) that will influence the long-term profit. Therefore, the owner will want to identify a policy for what decision to make (=action to take) at each state of the restaurant, to maximize the long-term profit. This is an example of reinforcement learning, and finding an optimal policy requires iterations of gathering data (trying action at a state) and updating the expectations for the long-term profits.

One of the common challenges encountered in these learning problems is that some values in the data may be unavailable for several reasons. For example, some of the customer feedback and reviews might have missing entries. These missing values make it harder to learn the regression function. Also, in the reinforcement learning setup, trying a new action to collect more data about the expected long-term profit can be costly and even risky sometimes, and thus, the owner may deliberately decide to gather data from only a small subset of all possible actions. Now the owner can cut some expenses related to data collection, but will need to figure out an optimal policy from the limited information.

As a matter of fact, the missing data problem is ubiquitous throughout various disciplines, and it is a nuisance as it disturbs learning the relationship between variables from data [95, 53]. For decades, various attempts have been made to 'fix' the data either by discarding incomplete

data points or by filling in the missing values with reasonably generated replacement values. The latter approach is referred to as *imputation*, and has comparative advantage over the former deletion approach (a.k.a. complete case analysis in the literature as it only handles the complete cases of data points). It is because imputation preserves all cases by replacing missing values with values estimated based on other available information, and hence, no information is wasted. Specifically, in this thesis, we restrict our discussion to the imputation methods that replace the missing values in a tabular dataset with the values estimated by low-rank matrix completion, which can be viewed as a refined instance of regression imputation.

A prudent reader might ask why we consider regression imputation instead of methods based on maximum likelihood estimation (MLE), such as the celebrated expectation-maximization (EM) algorithm proposed by Dempster, Laird, and Rubin [46]. It is true that MLE-based techniques have been very successful in both theory and practice and are regarded as the current state of the art [124, 53, 95]. Nevertheless, they require a probabilistic data generative model in parametric form (i.e., likelihood) and usually do not come with global convergence guarantees. In contrast, our proposed approach does not require such a probabilistic model, and we establish upper bounds on the bias arising from the imputation.

It should be also emphasized that the two learning problems of our interest involve statistical modeling for the purpose of prediction. That is, data and statistical tools are used primarily for building a predictive model that accurately forecasts new or future outcomes, following the tradition of predictive modeling. Note that predictive modeling should be distinguished from descriptive modeling that is aimed at succinctly summarizing the observed data, and also from explanatory modeling, which uses statistical models to test causal hypotheses. As a consequence, in Part II of this thesis, we exclusively focus on predictive modeling and attempt to find 'a' model that accurately predicts the response to new input variables, rather than attempting to estimate 'the' true underlying model. It is because descriptive models are not helpful in making decisions for the future as they are not aimed at forecasting possible outcomes, whereas causal explanations are not always necessary in making decisions although such explanations could be helpful sometimes. See the articles of Breiman [25] and Shmueli [128] and references therein for more discussions on this distinction.

In summary, in Part II of the thesis, we consider imputing missing values in tabular datasets with low-rank matrix completion for the purpose of supervised and reinforcement learning, and ask the following informally stated question:

Question 1.1 (central question of Part II, informal)**.** When we have only access to a partial dataset with missing values, can we solve the statistical learning problems with imputed datasets nearly as accurately as if the complete dataset were provided?

1.2 Global Approximation of a Semidefinite Program

A semidefinite program (SDP) is an optimization problem of the form

$$\begin{array}{ll} \text{minimize}_{X \in \boldsymbol{S}^n} & \mathrm{Tr}(CX) \\ \text{subject to} & \mathrm{Tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X \succeq 0, \end{array} \tag{1.1}$$

where $\boldsymbol{S}^n$ denotes the set of $n \times n$ real symmetric matrices, $\mathrm{Tr}(\cdot)$ is the trace of a matrix, $C, A_1, \ldots, A_m \in \boldsymbol{S}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$ are problem data, and $X \succeq 0$ denotes that the decision matrix variable X is constrained to be positive semidefinite.

SDPs have received much attention as they have numerous applications throughout applied and computational mathematics as well as engineering. In discrete optimization, SDP provides some of the most powerful convex relaxations for prominent hard problems such as the maximum cut [62] and the maximal stable set [97] problems in graph. In statistics and machine learning, SDPs also arise as computationally tractable relaxations of sparse recovery problems, e.g., rank minimization [116] and sparse PCA [41]. In polynomial optimization, the hierarchies of SDPs based on the notion of sum-of-squares (SOS) polynomials provide one of the most promising approaches for global optimization without convexity assumptions [86, 112]. Last but not least, various notions of the safety, stability, and robustness in control and robotics can be certified by searching a Lyapunov function that satisfies appropriate nonnegativity constraints, and this search can be automated by the SOS approach [111]; more recently, these techniques to search a certificate have been also actively studied in the context of algorithmic robust statistics [85, 47, 36].

In addition to the long list of applications, SDPs are well known for their strong theoretical and computational properties. For instance, for every SDP of the form (1.1) (called the primal problem), there exists another associated SDP, called the dual problem. Under mild assumptions (usually called constraint qualifications), e.g., if both primal and dual problems are strictly feasible, strong duality holds and the optimal costs are equal for both problems. Moreover, many algorithms for linear programs have been adapted to SDPs, and these algorithms can solve SDPs to arbitrary accuracy in polynomial time [155].

Despite these numerous virtues, when it comes to solving SDPs for practical purposes, they suffer from one serious problem, which is scalability. When the SDP formulated in the form (1.1) has large n or m, the computation time and memory required to solve it will be prohibitively large, as most solvers rely on interior-point methods. Such scalability issues remain as major challenges for researchers in the field, and various attempts have been made to overcome these impediments.

Recently, Ahmadi and Majumdar [6] introduced tractable alternatives of SDP based on linear-programming (LP) and second-order-cone-programming (SOCP), which they call diagonally dominant sum-of-squares (DSOS) and scaled diagonally dominant sum-of-squares (SDSOS) optimization, respectively. The underlying idea of their approaches is to replace the delicate $n \times n$ positive semidefinite (PSD) constraint $X \in \{X' \in \boldsymbol{S}^n : X' \succeq 0\}$ with simpler sufficient conditions that comprise $\binom{n}{2}$ number of linear (or quadratic) inequalities so that the resulting approximate programs can be solved faster. As a result, these approaches trade off scalability with conservatism – they are guaranteed to produce a feasible solution of the original SDP, although it may have suboptimal value – and therefore, have advantages in safety-critical applications in particular.

In spite of promising empirical evidences that DSOS and SDSOS approaches can solve SDPs much faster than the standard methods without sacrificing the optimality too much [6, Section 4], there is little theoretical understanding about the tradeoff between the computational gain and the quality of approximate solutions. Motivated by this gap in the knowledge, we study the problem of approximating a large-scale SDP with smaller-sized SDPs in Part III of this thesis. Specifically, we ask the following question:

Question 1.2. Given two positive integers $k \leq n$, how many $k \times k$-sized PSD constraints are required to approximate the feasible set of the SDP of the form (1.1) so that the optimality gap (difference in the optimal value) is less than ϵ?

In essence, Quesetion 2 asks how many $k \times k$-sized PSD constraints are necessary for approximating the expressive power of a single $n \times n$ PSD constraint. Note that the case with $k = 1$ corresponds to an LP approximation of SDP, and the case with $k = 2$ is an SOCP approximation of SDP.

■ 1.3 Outline and Contributions of the Thesis

Part I of this thesis provides some mathematical preliminaries that will be used in our analysis in later chapters. The remainder of this thesis is mainly divided into two parts, each of which contains two or three chapters.

In Part II of the thesis, we study the utility of data imputation based on low-rank matrix completion for predictive learning tasks. To this end, we begin our discussion by providing a brief literature review on data imputation and matrix completion, and then discuss error guarantees of low-rank matrix completion algorithms (Chapter 3). Chapters 4 and 5 exhibit case studies of the usefulness of imputed data in predictive learning tasks. Chapter 4 considers the prediction capability of imputed-data-based regression when the covariate data have missing values. Chapter 5 studies the sample complexity of model-free Q-learning algorithms.

Part III of the thesis is concerned with the approximability of the cone of $n \times n$ positive semidefinite (PSD) matrices by using only a small number of $k \times k$ PSD constraints for k much smaller than n. Chapter 6 is a preparatory chapter that provides additional background on the motivation of our study and technical machinery used in our proofs. Chapter 7 contains results on the hardness of approximating the $n \times n$ PSD cone with smaller-sized PSD constraints.

A summary of our contributions in each chapter is provided below.

Part II

Data Imputation with Matrix Completion (Chapter 3) The main contribution of this chapter is the novel analysis of singular value thresholding algorithm for matrix completion. To be more precise, we prove an upper bound for the matrix estimation error measured in the spectral norm that matches the optimal upper bound and the minimax lower bound reported by Koltchinskii et al. [82, Theorem 13], but with a conceptually more straightforward proof. Moreover, we establish an upper bound on the estimation error of matrix completion in the $\ell_{2,\infty}$-norm (and also in the ℓ_∞-norm) sense, which is a stronger guarantee than the widely used Frobenius norm error bound (=mean squared error).

Although our analysis in this chapter is focused on a specific algorithm (singular value thresholding) and the resulting matrix estimator for the concreteness of discussion, any matrix completion algorithm may be used for the applications to be discussed in later chapters as long as the resulting matrix estimator has small estimation error. Thus, we also provide a brief literature survey for existing algorithms for matrix completion and their error analysis.

Errors-in-variables Regression (Chapter 4) Our first contribution in this chapter is the analysis of the prediction performance of regression that use the covariate data imputed by low-rank matrix completion. Specifically, we study the errors-in-variables regression problem, but our aim is at finding a good predictor for the response given covariate data, instead of identifying the true underlying regression model parameter. We argue that if the objective of regression analysis is in prediction rather than model identification, then it is possible to achieve prediction performance nearly as good as the predictor based on the full covariate data. Moreover, we describe a procedure that predicts the response given a *new* covariate that is not from the training set, which is not considered in other works in the literature. We discuss an informal error analysis for this out-of-sample prediction, and validate its performance with numerical experiments.

As a byproduct or our analysis, we observe that the well-established technique of principal component regression (PCR) algorithm can be interpreted as an instance of 'regression-after-imputation' algorithm (Algorithm 2) considered in this chapter. This observation enables extending our analysis to yield finite-sample analysis for PCR.

Sample-efficient Q Learning (Chapter 5) We examine reinforcement learning in the context of model-free Q-learning, and discuss how the low-rank structure of the optimal Q-function can be utilized to reduce the burden of exploration and improve the sample efficiency. The main contributions in this chapter are the methodological proposal of Q-learning algorithm that exploits low-rank structure in Q^* (Algorithm 3) and its theoretical analysis. We argue that if the Q^* function admits a low-rank structure, then that structure can be utilized by means of low-rank matrix completion to improve sample efficiency in value iteration updates. That is, it suffices to explore only a small fraction of state-action pairs instead of probing all pairs in order to update the current guess about Q^*.

As a main theoretical contribution of the chapter, we show that the proposed Q-learning procedure converges to Q^* with the desired improvement in sample complexity. Without any structural assumptions, the number of samples (explorations of state-action pairs) required to learn Q^* within an accuracy of ϵ scales as $|\mathcal{S}||\mathcal{A}|/\epsilon^2$ where $\mathcal{S}, \mathcal{A}$ denote the state space and the action space, respectively. In contrast, if the matrix completion oracle in use satisfies a certain ℓ_∞-contraction property, then it is possible to learn Q^* from only $\mathrm{rank}(Q^*) \cdot (|\mathcal{S}| \vee |\mathcal{A}|)/\epsilon^2$ number of samples. We complement our analysis with numerical experiments on simple stochastic control tasks.

Part III

Hardness of Approximating the PSD Cone by Smaller-sized PSD Cones (Chapter 7) This chapter is devoted to hardness results about approximating the cone of $n \times n$ positive semidefinite (PSD) matrices by using only a small number of $k \times k$ PSD constraints for k much smaller than n. The main contributions in this chapter are impossibility theorems of approximating the PSD cone. These impossibility theorems state the hardness of *globally* approximating the $n \times n$ PSD cone using a small number of $k \times k$ PSD constraints. Here we remark that our negative results do not contradict the empirical success of DSOS/SDSOS approaches because the optimal value of those approximate programs can be still close to that of the original SDP as long as the feasible set of the original SDP (1.1) is *locally* approximated, e.g., in the direction of C.

We begin this chapter by specifying three notions of approximation for sets and cones that we use. In essence, these notions of approximation are defined by comparing the width of the sets (bases of the cones) in slightly varying manners. They are natural notions to measure approximability of a set by the other, and closely related to quantifying the difference in the optimal value (optimality gap) induced by relaxing the feasible set in (1.1).

With the notions of approximation identified, we consider two schemes to construct a set that approximates the $n \times n$ PSD cone. First, we consider a specific construction of

approximating cone, which is defined as the set of $n \times n$ symmetric matrices that are PSD when restricted to a given fixed set of k-dimension al subspaces in $\mathbb{R}^n$. We show that when k is much smaller than n, it is necessary to impose PSD constraints on at least exponentially many subspaces to produce a cone that approximates $\boldsymbol{S}_+^n$ well. Second, we study a more general construction of approximating cones through the lens of extended formulations, and prove the hardness of approximating $\boldsymbol{S}_+^n$ by showing lower bounds on the $\boldsymbol{S}_+^k$-extension complexity that holds for any set that approximates $\boldsymbol{S}_+^n$.

It is noteworthy that the extension complexity lower bounds have a strong implication about the hardness of approximating $\boldsymbol{S}_+^n$ with $k \times k$ PSD constraints. Because they apply to *any* constructions of the approximating set, the lower bounds completely refute the possibility of *globally* approximating the large PSD cone $\boldsymbol{S}_+^n$ with a few, small-sized $k \times k$ PSD constraints. Therefore, the results from this chapter seem to suggest that local and/or problem-data-adaptive techniques are essential to approximate SDPs.

1.3.1 Bibliographic Note on Related Publications

Preliminary versions of the results in Chapter 4 and their applications in econometrics (synthetic control method) can be found in joint work with Anish Agarwal, Dennis Shen, and Devavrat Shah [3, 4]. Most of the materials in Chapter 5 are based on joint work with Devavrat Shah, Zhi Xu, and Yuzhe Yang [127]. Finally, the materials in Chapter 7 are based on joint work with Pablo A. Parrilo [134].

1.4 Notation

Before proceeding, we define notation and terminology that we commonly use in this thesis; the notation is mostly standard.

Set notation Throughout, we let $\mathbb{R}$ denote the set of real numbers and $\mathbb{N}$ denote the set of natural numbers. For a positive integer $n \in \mathbb{N}$, we denote with $[n]$ the set of integers $\{1, \ldots, n\}$. We let $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ denote the $(n-1)$-dimensional unit sphere in ℓ_2 norm. We use $\boldsymbol{S}^n$ to denote the set of $n \times n$ real symmetric matrices. Let $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$ and $\boldsymbol{S}_+^n = \{A \in \boldsymbol{S}^n : A \text{ is positive semidefinite}\}$. Given a set $S \subseteq \mathbb{R}^n$, we let $\mathrm{cl}(S)$, $\mathrm{conv}(S)$, and $\mathrm{cone}(S)$ denote the closure, the convex hull, and the conical hull of S, respectively.

Vector and matrix notation For any vector space V, we let $\mathit{I}_V : V \to V$ denote the identity map in V. We let $\mathbb{R}^n$ denote the n-dimensional vector space over $\mathbb{R}$ and $\mathbb{R}^{n_1 \times n_2}$ denote the set of $n_1 \times n_2$ real matrices. We may identify a vector $x \in \mathbb{R}^n$ with a $n \times 1$ matrix through the natural vector space isomorphism. For a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we let $\mathrm{diag}(x) \in \mathbb{R}^{n \times n}$ denote the $n \times n$ diagonal matrix such that $\mathrm{diag}(x)_{ii} = x_i$, $\forall i$. For $A \in \mathbb{R}^{n_1 \times n_2}$, we denote

with A^T its transpose. For $A \in \mathbb{R}^{n\times n}$, its trace $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$. For $I \subseteq [n_1], J \subseteq [n_2]$, let $X(I,J) \in \mathbb{R}^{|I|\times|J|}$ denote the $|I| \times |J|$ submatrix of X obtained by deleting the rows not in I, and the columns not in J. We reserve $\mathbf{1}_n \in \mathbb{R}^n$ and $\mathbf{1}_{n_1\times n_2} \in \mathbb{R}^{n_1\times n_2}$ to denote the vector/matrix with all elements equal to 1. We let I_n denote the $n \times n$ identity matrix.

We let $\langle \cdot, \cdot \rangle$ denote the standard inner product. For example, $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$ for $x, y \in \mathbb{R}^d$ and $\langle A, B\rangle = \mathrm{Tr}(A^T B)$ for $A, B \in \mathbb{R}^{n_1\times n_2}$. For a subspace $\mathcal{X} \subseteq \mathbb{R}^n$, we let $\mathcal{X}^\perp = \{y \in \mathbb{R}^n : \langle x, y\rangle = 0,\ \forall x \in \mathcal{X}\}$ denote the orthogonal complement of $\mathcal{X}$ in $\mathbb{R}^n$. A matrix $U \in \mathbb{R}^{n\times d}$ (with $n \geq d$) is called semi-orthogonal if the columns of U are orthonormal vectors. A square semi-orthogonal matrix is called orthogonal.

We use $s_i(A)$ to denote the i-th largest singular value of A. When $A \in \boldsymbol{S}^n$, we use $\lambda_i(A)$ to denote the i-th largest eigenvalue of A. A matrix $A \in \boldsymbol{S}^n$ is positive semidefinite (PSD), denoted by $A \succeq 0$, if $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$, or equivalently, if $\lambda_n(A) \geq 0$. Given two matrices $A, B \in \boldsymbol{S}^n$, we write $A \succeq B$ or $B \preceq A$ if $A - B \in \boldsymbol{S}^n_+$.

For vector $x \in \mathbb{R}^n$ and $p \geq 1$, we use ℓ_p to denote the usual ℓ_p-norms $\|x\|_p = \big(\sum_{i=1}^n |x_i|^p\big)^{1/p}$, where $\|x\|_\infty = \max_i |x_i|$. For a matrix $A \in \mathbb{R}^{n_1\times n_2}$ and two positive integers $p, q \geq 1$, we let $\ell_{p,q}$ denote the $\ell_p \to \ell_q$ operator norms (a.k.a. subordinate matrix norms) of A, i.e., $|||A|||_{pq} = \sup_{x\in\mathbb{R}^{n_2}:\|x\|_p\leq 1} \|Ax\|_q$. As shorthand, we use $|||A|||_p = |||A|||_{pp}$. Also, we let $\|A\|_p$ be the ℓ_p-norm of matrix $A \in \mathbb{R}^{n_1\times n_2}$ viewed as a vector in $\mathbb{R}^{n_1 n_2}$. Note that $|||A|||_p \neq \|A\|_p$ in general. We let $\|A\|_*$ and $\|A\|_F$ denote the nuclear norm (a.k.a. trace norm), and the Frobenius norm of A, respectively. Note that $\|A\|_F = \|A\|_2$ and that the operator norm of A induced by ℓ_2-norm satisfies $|||A|||_2 = s_1(A)$, and is also referred to as the spectral norm of A.

Asymptotic notation We also use standard asymptotic notation throughout this thesis. In particular, we use $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot), o(\cdot)$, and $\omega(\cdot)$ notation. Formally, given two real-valued sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, we write $a_n = \mathcal{O}(b_n)$ if there exists a constant $c < \infty$ and $N \in \mathbb{N}$ such that $a_n \leq c b_n$ for all $n \geq N$. Likewise, we write $a_n = \Omega(b_n)$ if $b_n = \mathcal{O}(a_n)$, and $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. Lastly, we say $a_n = o(b_n)$ if $|a_n|/|b_n| \to 0$ as $n \to \infty$, and $a_n = \omega(b_n)$ if $b_n = o(a_n)$. We use the notation $a_n \lesssim b_n$ to denote $a_n = \mathcal{O}(b_n)$, and $a_n \ll b_n$ to denote $a_n = o(b_n)$. Also, we let $a_n \asymp b_n$ denote $a_n = \Theta(b_n)$.

Probabilistic notation When a random variable (random vector) X has distribution D, we denote it by $X \sim D$. We let $X \sim D$, $\mathbb{P}_{X\sim D}(X \in S)$ denote the probability of the event $X \in S$ and $\mathbb{E}_{X\sim D}[f(X)]$ be the expected value of $f(X)$ when $X \sim D$. When the reference random variable and/or distribution is clear from the context, we will omit the subscript and simply write $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ for brevity. With a slight abuse of notation, we also write $X \sim \mu$ for a probability measure if $\mathbb{P}(X \in S) = \mu(S)$ for all measurable sets S. We use $N(\mu, \sigma^2)$ to denote univariate Gaussian distribution with mean μ and variance σ^2, and $N(\mu, \Sigma)$ to denote multivariate Gaussian distribution with mean μ and covariance $\Sigma \succeq 0$.

Miscellaneous notation We denote by $\mathbb{1}\{E\}$ the indicator function of an event E, which takes value 1 if E is true and 0 otherwise. We let $\vee$ and $\wedge$ denote maximum and minimum so that $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, $[x]_+ = x \vee 0$ for $x \in \mathbb{R}$.

Chapter 2

Mathematical Preliminaries

In this chapter, we review some relevant mathematical preliminaries that are used in our analysis. Our exposition is brief as we only provide the basic technical background, but it is mostly self-contained. We refer the interested readers to the classic texts provided at the beginning of each section for more details. Expert readers may want to skip this chapter and continue reading from Chapter 3.

2.1 Linear Algebra

Rank Let A be an $n_1 \times n_2$ matrix over $\mathbb{R}$. The rows of A span a subspace of $\mathbb{R}^{n_2}$ known as the row space of A, denoted by $\mathrm{rsp}(A)$. Similarly, the columns of A span a subspace of $\mathbb{R}^{n_1}$ known as the column space of A, denoted by $\mathrm{csp}(A)$. The dimensions of these spaces are respectively called the row rank and column rank of A, namely, $\mathrm{rrank}(A) = \dim \mathrm{rsp}(A)$ and $\mathrm{crank}(A) = \dim \mathrm{csp}(A)$. The row rank of a matrix is always equal to its column rank, and we call it the rank of the matrix. We denote the rank of A by $\mathrm{rank}(A)$.

For any matrix $A \in \mathbb{R}^{n_1 \times n_2}$, there exist $U \in \mathbb{R}^{n_1 \times n_1}$, $V \in \mathbb{R}^{n_2 \times n_2}$ and $\Sigma \in \mathbb{R}^{n_1 \times n_2}$ such that (i) $A = U\Sigma V^T$; (ii) U and V are orthogonal; and (iii) $\Sigma_{11} \geq \cdots \geq \Sigma_{nn} \geq 0$ where $n = n_1 \wedge n_2$, and $\Sigma_{ij} = 0$ if $i \neq j$. This is called a singular value decomposition (SVD) of matrix A. Note that the singular values $\{\Sigma_{11}, \ldots, \Sigma_{nn}\}$ are uniquely defined, whereas singular vectors (the columns of U and V) may have some ambiguity (e.g., sign).

When $\mathrm{rank}(A) = r$, $\Sigma_{ii} > 0$ if and only if $i \leq r$. Thus, $U\Sigma V^T = U_r \Sigma_r V_r^T$ where $U_r \in \mathbb{R}^{n_1 \times r}$, $V_r \in \mathbb{R}^{n_1 \times r}$ are submatrices of U, V formed by taking only the left r columns, and $\Sigma_r \in \mathbb{R}^{r \times r}$ is the top left $r \times r$ principal submatrix of Σ. We call $U_r \Sigma_r V_r^T$ a compact SVD of A.

Singular values lead to a generalization of the matrix inverse that applies to all matrices (i.e., all linear transformations), so called the Moore-Penrose pseudoinverse. When A admits a compact singular value decomposition, $A = U\Sigma V^T$, its Moore-Penrose pseudoinverse, denoted by $A^\dagger$, is defined to be $A^\dagger = V\Sigma^{-1}U^T$. Note that $A^\dagger = A^{-1}$ when A is invertible.

Orthogonal Invariant Norm We say a matrix norm $\|\cdot\|$ is orthogonally invariant if $\|UAV^T\| = \|A\|$ for all $A \in \mathbb{R}^{d_1 \times d_2}$ and for all semi-orthogonal matrices $U \in \mathbb{R}^{n \times d_1}$ and $V \in \mathbb{R}^{n \times d_2}$ ($n \geq d_1 \vee d_2$). Note that any norms that are defined in terms of the singular values – e.g., the spectral norm and the Frobenius norm – are orthogonally invariant.

Orthogonal Projection Let $\mathcal{X}$ be a vector subspace of $\mathbb{R}^n$. For any $v \in \mathbb{R}^n$, there exists a unique minimizer $x^*(v) = \arg\min_{x \in \mathcal{X}} \|v - x\|_2$, which is called the orthogonal projection of v onto $\mathcal{X}$. The orthogonal projection $\Pi_{\mathcal{X}} : v \mapsto x^*(v)$ is a linear transformation from $\mathbb{R}^n$ to $\mathcal{X}$. Moreover, $\Pi_{\mathcal{X}} + \Pi_{\mathcal{X}^\perp} = I_{\mathbb{R}^n}$ for any subspace $\mathcal{X} \subseteq \mathbb{R}^n$.

Identifying $\mathbb{R}^n$ with $\mathbb{R}^{n \times 1}$, we can write $\Pi_{\mathcal{X}}(v) = P_U v$ where $P_U = UU^T \in \boldsymbol{S}^n$ for a semi-orthogonal matrix U such that $\mathrm{csp}(U) = \mathcal{X}$. Indeed, $P_U = P_{U'}$ for any semi-orthogonal matrices U, U' such that $\mathrm{csp}(U) = \mathrm{csp}(U') = \mathcal{X}$. Thus, we let $P_{\mathcal{X}} \in \boldsymbol{S}^n$ denote the projection matrix onto the vector subspace $\mathcal{X} \subseteq \mathbb{R}^n$.

■ 2.2 Geometry

■ 2.2.1 Principal Angles and Matrix Perturbation

Principal Angles between Subspaces The concept of principal angles between subspaces was introduced by Jordan [76], and it provides a way to measure how closely two subspaces are aligned to each other. In statistics, Hotelling defines the principal angles in the form of canonical correlations [71]. Recall that an acute angle between two unit vectors $x, y \in \mathbb{R}^n$, i.e., $\|x\|_2 = \|y\|_2 = 1$, is defined as $\theta(x, y) = \arccos|\langle x, y \rangle|$. This definition can be extended to the principal angles between subspaces as follows.

Definition 2.2.1 (Principal angles between subspaces)**.** Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ be subspaces with $\dim \mathcal{X} = d_1$ and $\dim \mathcal{Y} = d_2$. Let $d = d_1 \wedge d_2$. The i-th principal angles between $\mathcal{X}$ and $\mathcal{Y}$ are recursively defined by

$$\theta_i(\mathcal{X}, \mathcal{Y}) = \arccos\langle x_i, y_i \rangle \quad \text{where} \quad (x_i, y_i) \in \underset{\substack{(x,y) \in \mathcal{X} \times \mathcal{Y}: \\ \|x\|_2 = \|y\|_2 = 1, \\ \langle x, x_j \rangle = \langle y, y_j \rangle = 0, \ \forall j < i}}{\arg\max} \langle x, y \rangle.$$

The principal angle matrix between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\Theta(\mathcal{X}, \mathcal{Y}) = \mathrm{diag}\big(\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \theta_d(\mathcal{X}, \mathcal{Y})\big).$$

The vectors $\{x_1, \ldots, x_d\}$ and $\{y_1, \ldots, y_d\}$ are called the principal vectors.

Note that the principal angles are well defined, although the principal vectors are not

uniquely defined. To make this point clear, we present an alternative definition of the principal angles in the following theorem, which is based on the singular value decomposition [20, 65].

Theorem 2.2.2 (Alternative definition of principal angles)**.** *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ be subspaces with $\dim \mathcal{X} = d_1$ and $\dim \mathcal{Y} = d_2$. Let $X \in \mathbb{R}^{n\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$ be semi-orthogonal matrices such that $\operatorname{csp}(X) = \mathcal{X}$ and $\operatorname{csp}(Y) = \mathcal{Y}$. Let $U\Sigma V^T$ be a SVD of X^TY, where $\Sigma \in \mathbb{R}^{d_1\times d_2}$ has main diagonal elements $\Sigma_{11} \geq \ldots \geq \Sigma_{dd} \geq 0$ with $d = d_1 \wedge d_2$. Then $\theta_i(\mathcal{X}, \mathcal{Y}) = \arccos \Sigma_{ii}$. Moreover, the associated principal vectors are given by the first d columns of XU and YV.*

The sine of the principal angles between subspaces is a natural measure of their difference, partly due to its connection to the orthogonal projections. We refer the interested readers to [43, 139] for more details.

Lemma 2.2.3. *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ be subspaces with $\dim \mathcal{X} = \dim \mathcal{Y} = d$. Let $\sin\Theta(\mathcal{X}, \mathcal{Y}) = \operatorname{diag}\big(\sin\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \sin\theta_d(\mathcal{X}, \mathcal{Y})\big)$. Then*

$$\|\sin\Theta(\mathcal{X}, \mathcal{Y})\| = \|P_{\mathcal{Y}^\perp} P_{\mathcal{X}}\| = \|P_{\mathcal{X}^\perp} P_{\mathcal{Y}}\|. \tag{2.1}$$

for any orthogonally invariant norm $\|\cdot\|$.

Proof. Let $X, Y \in \mathbb{R}^{n\times d}$ be semi-orthogonal matrices such that $\operatorname{csp}(X) = \mathcal{X}$ and $\operatorname{csp}(Y) = \mathcal{Y}$. Let $U\Sigma V^T$ be a SVD of X^TY. Then $X^TYY^TX = U\Sigma\Sigma^TU^T = U\cos^2\Theta(\mathcal{X}, \mathcal{Y})U^T$ by Theorem 2.2.2. Now let $Y_\perp \in \mathbb{R}^{n\times(n-d)}$ be a semi-orthogonal matrix such that $\operatorname{csp}(Y_\perp) = \mathcal{Y}^\perp$. Then $Y_\perp Y_\perp^T = I_n - YY^T$ because $YY^T = P_{\mathcal{Y}}$ and $Y_\perp Y_\perp^T = P_{\mathcal{Y}^\perp}$. Therefore,

$$X^TY_\perp Y_\perp^T X = X^T\big(I_n - YY^T\big)X = U\sin^2\Theta(\mathcal{X}, \mathcal{Y})U^T.$$

In other words, $X^TY_\perp = U\sin\Theta(\mathcal{X}, \mathcal{Y})\widetilde{V}^T$ for some semi-orthogonal $\widetilde{V} \in \mathbb{R}^{n\times(n-d)}$.

■

Matrix Perturbation Theory Let $A, E \in \mathbb{R}^{n_1\times n_2}$ and $\hat{A} = A + E$. Matrix perturbation theory is concerned with how the singular values and the singular vectors of the perturbed matrix $\hat{A}$ differ from those of the original matrix A. In this thesis, it suffices for us to have two basic inequalities – Weyl's inequality and Davis-Kahan $\sin\Theta$ theorem.

Theorem 2.2.4 (Weyl [161])**.** *Let $A, E \in \mathbb{R}^{n_1\times n_2}$ and $\hat{A} = A + E$. Then*

$$\big|s_i(\hat{A}) - s_i(A)\big| \leq |||E|||_2, \quad \forall i \in [n_1, n_2].$$

Suppose that $A \in \mathbb{R}^{n_1\times n_2}$ approximately has rank r in the sense that there is a significant gap between $s_r(A)$ and $s_{r+1}(A)$. We may partition a SVD of X into a part that contains the

top r singular values and the rest as

$$A = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix},$$

where $U \in \mathbb{R}^{n_1 \times r}$, $U_\perp \in \mathbb{R}^{n_1 \times (n_1 - r)}$, $\Sigma_1 = \operatorname{diag}\big(s_1(A), \ldots, s_r(A)\big)$, $\Sigma_2 \in \mathbb{R}^{(n_1-r)\times(n_2-r)}$, $V \in \mathbb{R}^{n_2 \times r}$, and $V_\perp \in \mathbb{R}^{n_2 \times (n_2 - r)}$. Similarly, we can partition a SVD of $\hat{A} = A + E$ as

$$\hat{A} = \begin{bmatrix} \hat{U} & \hat{U}_\perp \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \hat{V}^T \\ \hat{V}_\perp^T \end{bmatrix},$$

where $\hat{U}, \hat{U}_\perp, \hat{\Sigma}_1, \hat{\Sigma}_2, \hat{V}, \hat{V}_\perp$ have similar structures as above.

Davis and Kahan [43] established the celebrated $\sin\Theta$ theorem that provides an upper bound for $\big\| \sin\Theta\big(\operatorname{csp}(U), \operatorname{csp}(\hat{U})\big) \big\|$ in terms of $\|E\|$, where $\|\cdot\|$ can be any unitary invariant norm. They focused on the eigenvectors of symmetric (indeed, Hermitian) matrices and their results are extended by Wedin [160] to singular vectors of general asymmetric matrices. Here, we state the Wedin's version of $\sin\Theta$ theorem.

Theorem 2.2.5 (Davis-Kahan $\sin\Theta$ theorem, Wedin's version [160])**.** *Let $A, E \in \mathbb{R}^{n_1 \times n_2}$ and $\hat{A} = A + E$. If $s_r(A) - s_{r+1}(\hat{A}) > 0$, then for any orthogonally invariant norm $\|\cdot\|$, the following inequality holds:*

$$\Big\| \sin\Theta\big(\operatorname{csp}(U), \operatorname{csp}(\hat{U})\big) \Big\| \vee \Big\| \sin\Theta\big(\operatorname{csp}(V), \operatorname{csp}(\hat{V})\big) \Big\| \leq \frac{\|U^T E\| \vee \|EV\|}{s_r(A) - s_{r+1}(\hat{A})}.$$

■ 2.2.2 Low-rank Variety and its Tangent Space

An algebraic variety is defined as the set of zeros of a system of polynomial equations [69]. The variety of matrices that has rank at most r is defined as

$$\mathcal{V}(r) := \{M \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(M) \leq r\}. \tag{2.2}$$

Note that this is an algebraic variety because it can be defined as the set of matrices with vanishing $(r+1) \times (r+1)$ minors. The dimension of this variety is $r(n_1 + n_2 - r)$, and this variety is nonsingular at matrices with rank exactly r. For any matrix $M \in \mathbb{R}^{n_1 \times n_2}$, the tangent space of $\mathcal{V}\big(\operatorname{rank}(M)\big)$ at M is defined as

$$T(M) = \operatorname{span}\big\{X \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rsp}(X) \subseteq \operatorname{rsp}(M) \text{ or } \operatorname{csp}(X) \subseteq \operatorname{csp}(M)\big\}.$$

When $M = U\Sigma V^T$ is a compact SVD of M, and $\mathrm{rank}(M) = r$, $T(M)$ can be written as

$$T(M) = \{UX^T + YV^T : X \in \mathbb{R}^{n_2 \times r}, Y \in \mathbb{R}^{n_1 \times r}\}. \tag{2.3}$$

Note that $\dim T(M) = r(n_1 + n_2 - r)$ and that $M \in T(M)$. In Chapter 3, we will treat $T(M)$ as a vector subspace of $\mathbb{R}^{n_1 \times n_2}$, and use it for our analysis of the SVT algorithm.

2.2.3 ϵ-net and Covering Number

Let (T, d) be a metric space, $K \subseteq T$ be a subset, and $\epsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ϵ-net of K if

$$\forall \in K,\ \exists x_0 \in \mathcal{N} \text{ such that } d(x, x_0) \leq \epsilon.$$

Note that for each ϵ, there may exist multiple ϵ-nets of K. The smallest possible cardinality of an ϵ-net of K is called the covering number of K, and is denoted by $N_{\mathrm{cov}}(K, d, \epsilon)$.

In this thesis, we only consider the case $T = \mathbb{R}^n$, $d(x, y) = d_{\mathrm{Euc}}(x, y) = \|x - y\|_2$, and $K = \mathbb{S}^{n-1}$. For this case, the following upper and lower bounds are known.

Lemma 2.2.6 ([156], Corollary 4.2.13)**.** *The covering number of the unit Euclidean sphere $\mathbb{S}^{n-1}$ satisfies the following inequality for any $\epsilon > 0$:*

$$N_{\mathrm{cov}}\big(\mathbb{S}^{n-1}, d_{\mathrm{Euc}}, \epsilon\big) \leq \Big(\frac{2}{\epsilon} + 1\Big)^n.$$

2.3 Convex Analysis

We recall some basic concepts and results in convex analysis. The materials in this section are standard and can be found in classic references, e.g. [118] and [12].

Duality If $S \subseteq \mathbb{R}^d$, the polar of S (in $\mathbb{R}^d$) is the closed convex set

$$S^\circ := \{y \in \mathbb{R}^d : \langle x, y\rangle \leq 1 \text{ for all } x \in S\}. \tag{2.4}$$

We observe a few properties involving the polars. First of all, if $S \subseteq T$, then $S^\circ \supseteq T^\circ$. Next, it is useful to note that $(S \cup T)^\circ = S^\circ \cap T^\circ$ for any $S, T \subseteq \mathbb{R}^d$, and that $(S \cap T)^\circ = \mathrm{cl}\,\mathrm{conv}(S^\circ \cup T^\circ)$ if S, T are closed, convex, and contain the origin. Lastly, if S is a closed, convex set that contains the origin, then $(S^\circ)^\circ = S$; this is known as the bipolar theorem (Lemma 2.3.3).

A nonempty closed convex set $\mathcal{K} \subset \mathbb{R}^d$ is called a cone if $\mathcal{K}$ is invariant under positive scaling, i.e., whenever $x \in \mathcal{K}$ and $t \geq 0$, then $tx \in \mathcal{K}$. Given a cone $\mathcal{K}$, its dual cone $\mathcal{K}^*$ (in $\mathbb{R}^d$) is defined via

$$\mathcal{K}^* := \{y \in \mathbb{R}^d : \langle x, y\rangle \geq 0 \text{ for all } x \in \mathcal{K}\}. \tag{2.5}$$

Note that $\mathcal{K}^* = -\mathcal{K}^\circ$. Thus, it follows from the properties of polars that (i) $(\mathcal{K}^*)^* = \mathcal{K}$; (ii) if $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $\mathcal{K}_1^* \supseteq \mathcal{K}_2^*$; and (iii) $\big(\mathcal{K}_1 \cap \mathcal{K}_2\big)^* = \operatorname{cl}\operatorname{cone}(\mathcal{K}_1 \cup \mathcal{K}_2)$ for two cones $\mathcal{K}_1, \mathcal{K}_2$.

The notion of conic duality is closely related to that of set polarity. To clarify the link, we first define a base of a closed convex cone $\mathcal{K}$. Fix a nonzero vector $e \in \mathbb{R}^d$ and the corresponding affine hyperplane

$$H_e := \{x \in \mathbb{R}^d : \langle e, x\rangle = \langle e, e\rangle\}.$$

If $e \in \mathcal{K}^* \setminus \mathcal{K}^\perp$ where $\mathcal{K}^\perp = \{v \in \mathbb{R}^d : \langle v, x\rangle = 0, \forall x \in \mathcal{K}\}$, then we call the set $\mathcal{K}_e^b := \mathcal{K} \cap H_e$ as the base of $\mathcal{K}$ with respect to e. The duality of cones carries over to the duality of bases.

Lemma 2.3.1 ([12], Lemma 1.6)**.** *Let $\mathcal{K} \subset \mathbb{R}^d$ be a closed convex cone and $e \in \mathcal{K} \cap \mathcal{K}^*$ be a nonzero vector. Then*

$$(\mathcal{K}^*)_e^b = \big\{y \in H_e : \langle -(y-e), x-e\rangle \le \langle e, e\rangle \text{ for all } x \in \mathcal{K}_e^b\big\}.$$

In other words, if we translate H_e so that e becomes the origin, and consider $\mathcal{K}_e^b$ and $(\mathcal{K}^*)_e^b$ as subsets of that vector space, then $(\mathcal{K}^*)_e^b = -\langle e, e\rangle(\mathcal{K}_e^b)^\circ$.

Remark 2.3.2. In this paper, we are concerned with cones $\mathcal{K}$ such that $\boldsymbol{S}_+^n \subseteq \mathcal{K} \subseteq \boldsymbol{S}^n$ and the unit-trace subspace H. Note that $H = H_e$ with $e = \frac{1}{n} I_n$. We let $B_H(\mathcal{K}) = \mathcal{K}_e^b - \frac{1}{n} I_n$ denote the base of $\mathcal{K}$ with respect to $e = \frac{1}{n} I_n$, translated by $-\frac{1}{n} I_n$ to contain 0.

Minkowski Functional and support function Let S be a nonempty subset of $\mathbb{R}^d$. The Minkowski functional (or gauge function) of S is defined to be the function $p_S : \mathbb{R}^d \to [0, \infty]$ valued in the extended real numbers such that

$$p_S(x) := \inf\{\lambda \in \mathbb{R} : \lambda > 0 \text{ and } x \in \lambda S\}. \tag{2.6}$$

We follow the convention that the infimum of the empty set is the positive infinity ∞. The support function of S is defined as $h_S : \mathbb{R}^d \to [0, \infty]$ such that

$$h_S(x) := \sup_{z \in S} \langle x, z\rangle. \tag{2.7}$$

There is a duality between the gauge function and the support function. In words, the gauge function of a convex set is the support function of its polar, and vice versa.

Lemma 2.3.3 ([118], Theorem 14.5)**.** *Let S be a closed convex set containing the origin. Then the polar S° is another closed convex set containing the origin, and $(S^\circ)^\circ = S$. Moreover,*

$$p_S(x) = h_{S^\circ}(x) \qquad \text{and} \qquad p_{S^\circ}(x) = h_S(x).$$

Mean Width Given a nonempty, bounded set $S \subset \mathbb{R}^d$, we define the mean width of S as the average of $h_S(u)$ with u distributed uniformly over the unit sphere in the ambient space:

$$w(S) := \int_{\mathbb{S}^{d-1}} h_S(u) d\sigma(u),$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ and σ is the normalized Haar measure on $\mathbb{S}^{d-1}$ (uniform probability measure on $\mathbb{S}^{d-1}$). It is often convenient to consider the Gaussian variant of the mean width because its value does not depend on the ambient dimension.

Definition 2.3.4 (Gaussian width)**.** For any nonempty bounded set $S \subset \mathbb{R}^d$, the Gaussian (mean) width of S is defined as

$$w_G(S) := \mathbb{E}_g h_S(g) = \mathbb{E}_g \left[\sup_{x \in S} \langle g, x \rangle\right] = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \sup_{x \in S} \langle z, x \rangle \exp(-\|z\|^2/2) dz.$$

where g is a standard Gaussian random vector in $\mathbb{R}^d$.

It is easy to verify that $w_G(S) = \kappa_d \cdot w(S)$ where $\kappa_d := \mathbb{E}_g \|g\|_2 = \frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)}$. Note that κ_d depends only on d and is of order $\sqrt{d}$ (it is known that $\sqrt{d - 1/2} \le \kappa_d \le \sqrt{d - d/(2d+1)}$).

The Gaussian width has many nice properties. Here we list a few of them that we use in later sections.

1. The Gaussian width does not depend on the ambient dimension.
2. The Gaussian width is invariant under translation and rotation.
3. If $S \subseteq S'$, then $w_G(S) \le w_G(S')$.

Urysohn's Inequality Given a bounded measurable set $S \subset \mathbb{R}^d$, its volume radius is defined as

$$\mathrm{vrad}(S) := \left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(B_d^2)}\right)^{1/d}$$

where B_d^2 is the unit d-dimensional Euclidean ball. The volume radius of S is the radius of the Euclidean ball that has the same volume as S.

A set $K \subset \mathbb{R}^d$ is a convex body if it is a convex, compact set with nonempty interior. The following inequality, known as Urysohn's inequality, states that the mean width is minimized for Euclidean balls, among the sets that have the same volume.

Lemma 2.3.5 (Urysohn's inequality; [12], Propositions 4.15 & 4.16)**.** *Let $K \subset \mathbb{R}^d$ be a convex body containing the origin in its interior. Then*

$$\frac{1}{w(K^\circ)} \le \mathrm{vrad}(K) \le w(K).$$

2.4 Concentration Inequalities

Here we collect a few facts about concentration of light-tailed random variables such as Gaussians and sub-Gaussians. These are standard results and more details can be found in references such as [24], [156] and [12].

2.4.1 Basic Concepts

Moments, Moment Generating Function, and Tail Bounds Let X be a random variable. Recall that $\mathbb{E}[X^k]$ and $\mathbb{E}\big[(X-\mathbb{E}X)^k\big]$ for $k \in \mathbb{N}$ are called the k-th moment and the k-th central moment of X, respectively. The moment generating function (MGF) of X is defined as

$$M_X(\lambda) = \mathbb{E}\big[e^{\lambda X}\big], \quad \lambda \in \mathbb{R},$$

whenever this expectation exists. The name of MGF originates from the property that if $M_X(\lambda)$ exists on an open interval around $\lambda = 0$, then $\mathbb{E}[X^k] = \frac{d^k}{d\lambda^k} M_X(\lambda)\big|_{\lambda=0}$.

In this thesis, we frequently use the fact that moments of a random variable capture useful information about its tail behavior, i.e., $\mathbb{P}(X - \mathbb{E}X \geq \tau)$ and $\mathbb{P}(X - \mathbb{E}X \leq -\tau)$ for $\tau > 0$. For instance, an elementary, yet powerful, tool to bound such tail probabilities is based on Markov's inequality.

Lemma 2.4.1 (Markov's inequality)**.** *For any nonnegative random variable X and $\tau > 0$,*

$$\mathbb{P}(X \geq \tau) \leq \frac{\mathbb{E}[X]}{\tau}.$$

Proof. Observe that $X \geq X \cdot \mathbb{1}\{X \geq \tau\} \geq \tau \cdot \mathbb{1}\{X \geq \tau\}$ for all $\tau > 0$. Taking expectation of both sides, we get $\mathbb{E}[X] \geq \tau \cdot \mathbb{P}(X \geq \tau)$. ■

Markov's inequality might seem crude at first glance, however, it can be boosted with a simple trick to yield sharper estimates. Let ϕ be a nondecreasing nonnegative function defined on a possibly infinite interval $I \subseteq \mathbb{R}$. Then for any $\tau > 0$,

$$\mathbb{P}(X \geq \tau) \leq \mathbb{P}\left(\phi(X) \geq \phi(\tau)\right) \leq \frac{\mathbb{E}\phi(X)}{\phi(\tau)}.$$

If we choose $\phi(\tau) = \tau^2$, then we obtain Chebyshev's inequality. More importantly, if we choose $\phi(\tau) = e^{\lambda\tau}$ for $\lambda > 0$, we obtain an exponential tail bound. This observation forms the basis of the Cramér-Chernoff bounding method, which optimizes $\lambda \in I$ after exponentiation to acquire the tightest upper bound.

Gaussian Random Variables A real random variable X is called a Gaussian random variable with mean μ and variance σ^2 if it follows a distribution whose density with respect to the Lebesgue measure on $\mathbb{R}$ is $f(x) = (2\pi\sigma^2)^{-1/2} \exp\big\{-(x-\mu)^2/2\sigma^2\big\}$ for all $x \in \mathbb{R}$. The MGF of $X \sim N(\mu, \sigma^2)$ is $M_X(\lambda) = \exp\big(\mu\lambda + \frac{1}{2}\sigma^2\lambda^2\big)$ for all $\lambda \in \mathbb{R}$. In particular, X is called a standard Gaussian random variable if $X \sim N(0,1)$. A real random vector $X = (X_1, \ldots, X_n)$ is called a standard Gaussian random vector if all of its components are independent standard Gaussian random variables. Equivalently, X is a standard Gaussian random vector if $\langle a, X\rangle$ is a standard Gaussian random variable for all deterministic vector $a \in \mathbb{R}^n$.

Recall that the space $\boldsymbol{S}^n$ of real symmetric $n \times n$ matrices equipped with the trace inner product is isometrically isomorphic to the real Euclidean space of dimension $\binom{n+1}{2}$ equipped with the usual inner product. We define the standard Gaussian distribution in $\boldsymbol{S}^n$ via the natural isomorphism between $\boldsymbol{S}^n$ and $\mathbb{R}^{\binom{n+1}{2}}$.

Definition 2.4.2. A random matrix $A \in \boldsymbol{S}^n$ is standard Gaussian if the elements $(a_{ij})_{1\le i\le j\le n}$ are independent random variables such that $a_{ii} \sim N(0,1)$ and $a_{ij} \sim N(0, 1/2)$ for $i < j$.

Note that A is a standard Gaussian vector in the space $\boldsymbol{S}^n$ if and only if $\sqrt{2}A$ is a GOE(n) (Gaussian Orthogonal Ensemble) matrix, cf. [12, Section 6.2]. The GOE has the property of orthogonal invariance, i.e., if $A \in \boldsymbol{S}^n$ is a GOE(n) matrix, then for any fixed orthogonal matrix $U \in O(n)$, the random matrix UAU^T is also a GOE(n) matrix.

Sub-Gaussian and Sub-exponential Random Variables Many interesting properties of Gaussian random variables are due to the fast decaying tail probabilities. Such properties are shared by some of non-Gaussian random variables, so called the class of sub-Gaussian random variables. This notion can be formalized based on the moment-generating function $M_X(\lambda)$ as follows.

Definition 2.4.3. A random variable X is sub-Gaussian with parameter $v > 0$, or v-sub-Gaussian, if $\mathbb{E}[X] = 0$ and

$$M_X(\lambda) \le \exp\Big(\frac{1}{2}v\lambda^2\Big), \quad \forall \lambda \in \mathbb{R}. \tag{2.8}$$

Definition 2.4.4. A random variable X is sub-exponential with parameters (v, c), or (v, c)-sub-exponential, where $v, c > 0$ if $\mathbb{E}[X] = 0$ and

$$M_X(\lambda) \le \exp\Big(\frac{1}{2}v\lambda^2\Big), \quad \forall \lambda \text{ such that } |\lambda| \le \frac{1}{c}.$$

For example, exponential and chi-squared random variables (with centering) are sub-exponential. Informally, a sub-gaussian random variable can be viewed as a sub-exponential random variable with c tending to 0.

Example 2.4.5 (Centered Bernoulli)**.** Let $\beta^0(p)$ denote the centered Bernoulli random variable with parameter $p \in [0,1]$ such that if $X \sim \beta^0(p)$, then $X = 1$ with probability p and $X = 0$ with probability $\overline{p} = 1 - p$. A centered Bernoulli random variable is trivially sub-Gaussian because it is bounded. Indeed, the sub-Gaussian norm of a centered random variable, i.e., the minimum σ^2 for which (2.8) holds, is known from [27], and it does not exceed $1/4$ for all $p \in [0,1]$. See Theorem 2.4.14 for the exact form of the sub-Gaussian norm of the centered Bernoulli random variable.

Example 2.4.6 (Squared sub-Gaussian)**.** If X is a sub-Gaussian random variable, then $Z = X^2 - \mathbb{E}[X^2]$ is a sub-exponential random variable. To be more specific, if X is v-sub-Gaussian, then Z is sub-exponential with parameters $(128v^2, 8v)$; see Lemma 2.4.15. We note that the constants 128 and 8 are used for concreteness, although they may not be the smallest possible.

A sub-exponential random variable exhibits sub-Gaussian tail behavior around its center and have exponentially decaying tail probabilities far away from 0. More precisely, the following tail probability bounds can be obtained by the Cramér-Chernoff method: if X is a sub-exponential random variable with parameters (v, c), then for every $\tau > 0$,

$$\mathbb{P}(X > \tau) \ \vee \ \mathbb{P}(X < -\tau) \leq \begin{cases} e^{-\tau^2/2v} & \text{if } 0 \leq \tau \leq \frac{v}{c}, \\ e^{-\tau/2c} & \text{if } \tau > \frac{v}{c}. \end{cases}$$

More generally, the notions of sub-Gaussian and sub-exponential random variables can be extended to random vectors and random matrices[1] as follows.

Definition 2.4.7. A random vector $X \in \mathbb{R}^n$ is sub-Gaussian with parameter $v > 0$ if $\langle u, X \rangle$ is sub-Gaussian with parameter v for all (deterministic) vector $u \in \mathbb{S}^{n-1}$. Similarly, a random vector $X \in \mathbb{R}^n$ is sub-exponential with parameters $v, c > 0$ if $\langle u, X \rangle$ is sub-exponential with parameter v, c for all vector $u \in \mathbb{S}^{n-1}$.

■ 2.4.2 Gaussian Inequalities

Gaussian Comparison Inequality The following fundamental inequality, known as Slepian's lemma, expresses that a Gaussian process can get farther (i.e., has a larger supremum) when it has weaker correlations. We refer the interested readers to [156, Theorem 7.2.1] and [12, Proposition 6.6] for more details.

Definition 2.4.8. A random process $(X_t)_{t \in T}$ is a Gaussian process if the random vector $(X_t)_{t \in T_0}$ has normal distribution for all finite subsets $T_0 \subset T$.

[1] Note that $X \in \mathbb{R}^{n_1 \times n_2}$ can be identified with a vector in $\mathbb{R}^{n_1 n_2}$ through a natural isometric isomorphism.

Lemma 2.4.9 (Slepian's lemma)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be Gaussian processes. Suppose that for all $t, s \in T$, the following three conditions hold: (i) $\mathbb{E}X_t = \mathbb{E}Y_t = 0$; (ii) $\mathbb{E}X_t^2 = \mathbb{E}Y_t^2$; and (iii) $\mathbb{E}X_tX_s \geq \mathbb{E}Y_tY_s$. Then for every $\tau \in \mathbb{R}$,*

$$\mathbb{P}\left[\sup_{t\in T} X_t \geq \tau\right] \leq \mathbb{P}\left[\sup_{t\in T} Y_t \geq \tau\right].$$

There is a well known upper bound for the expectation of the largest eigenvalue of a standard Gaussian random matrix in $\boldsymbol{S}^n$. Its proof is based on the Slepian's lemma and standard.

Lemma 2.4.10. *If a random matrix $G \in \boldsymbol{S}^n$ has the standard Gaussian distribution, then*

$$\mathbb{E}_G\big[\lambda_1(G)\big] = \mathbb{E}_G\Big[\sup_{v\in\mathbb{S}^{n-1}} \langle v, Gv\rangle\Big] \leq \sqrt{2n}.$$

Proof. We consider a Gaussian process $(X_v)_{v\in\mathbb{S}^{n-1}}$ defined over $\mathbb{S}^{n-1}$ such that $X_v = v^TGv+\gamma$ with G being standard Gaussian in $\boldsymbol{S}^n$ and $\gamma \sim N(0,1)$ independent of G. It is easy to verify that $\mathbb{E}\big[\sup_{v\in\mathbb{S}^{n-1}}\langle v, Gv\rangle\big] = \mathbb{E}_{G,\gamma}\big[\sup_{v\in\mathbb{S}^{n-1}} X_v\big]$. Now we introduce an auxiliary Gaussian process $(Y_v)_{v\in\mathbb{S}^{n-1}}$ such that $Y_v = g^Tv$ with $g \sim N(0, 2I_n)$. Observe that for all $u, v \in \mathbb{S}^{n-1}$, (1) $\mathbb{E}X_v = \mathbb{E}Y_v = 0$; (2) $\mathbb{E}X_v^2 = \mathbb{E}Y_v^2 = 2$; and (3) $\mathbb{E}X_uX_v - \mathbb{E}Y_uY_v = (1 - u^Tv)^2 \geq 0$. Thus, we can apply Slepian's lemma (Lemma 2.4.9) to obtain $\mathbb{E}_{G,\gamma}\big[\sup_{v\in\mathbb{S}^{n-1}} X_v\big] \leq \mathbb{E}_{g\sim N(0,2I_n)}\big[\sup_{v\in\mathbb{S}^{n-1}} Y_v\big] = \mathbb{E}_{g\sim N(0,2I_n)}\|g\|_2 \leq \big(\mathbb{E}_{g\sim N(0,2I_n)}\|g\|_2^2\big)^{1/2} = \sqrt{2n}$. ■

Remark 2.4.11. It is known that $\lim_{n\to\infty} \mathbb{E}_G\big[\lambda_1(G)\big]/\sqrt{2n} = 1$. Indeed, not only its expected value, but also its limiting distribution is known in the literature. The quantity $\lambda_1(G) - \sqrt{2n}$ is of order $n^{-1/6}$ and its distribution converges to the Tracy-Widom distribution after normalization [146].

Gaussian Concentration A smooth function of independent Gaussian random variables is sub-Gaussian. The following result is widely known as the Gaussian concentration inequality; see [24, Theorem 5.5] for example. Note that the sub-Gaussian parameter L^2 depends only on the smoothness of the function, and not on the number of Gaussian random variables.

Lemma 2.4.12 (Gaussian concentration)**.** *Let $X = (X_1, \ldots, X_n)$ be a vector of n independent standard Gaussian random variables. If $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz (with respect to the ℓ_2 norm), then $f(X) - \mathbb{E}f(X)$ is sub-Gaussian with sub-Gaussian parameter L^2.*

The following lemma states that the support function of a convex set concentrates around its mean. It can be proved applying Lemma 2.4.12 to the support function, which is Lipschitz with the Lipschitz constant being the diameter of the set.

Lemma 2.4.13. *Let $K \subset \mathbb{R}^d$ be a convex set containing 0. Let $w_G(K)$ denote the Gaussian width of K. Then for any $\alpha \geq 0$,*

$$\mathop{\mathbb{P}}_{g\sim N(0,I_d)}\left[\max_{x\in K}\langle g,x\rangle < (1-\alpha)w_G(K)\right] \vee \mathop{\mathbb{P}}_{g\sim N(0,I_d)}\left[\max_{x\in K}\langle g,x\rangle > (1+\alpha)w_G(K)\right] \leq \exp\left(-\frac{\alpha^2}{4\pi}\right).$$

Proof. Let $h_K(u) := \max_{x\in K}\langle u,x\rangle = \|u\|_{K^\circ}$ denote the support function of K. The function h_K is L-Lipschitz with $L = \sup_{x\in K}\|x\|_2$, the diameter of K, because for any $u, v \in \mathbb{R}^d$,

$$\left|h_K(u) - h_K(v)\right| = \left|\|u\|_{K^\circ} - \|v\|_{K^\circ}\right| \leq \|u-v\|_{K^\circ} \leq \sup_{x\in K}\|x\|_2\|u-v\|_2.$$

Moreover, we can show that $\sup_{x\in K}\|x\|_2 \leq \sqrt{2\pi}w_G(K)$. To see this, let $B(0,R)$ denote the Euclidean ball centered at 0 with radius R. It follows from [156, Proposition 7.5.2-(e)] that $\sup_{x,y\in K}\|x-y\|_2 \leq \sqrt{2\pi}w_G(K)$. Since $0 \in K$, this implies $K \subseteq B(0,\sqrt{2\pi}w_G(K))$. Applying Lemma 2.4.12 with $f = h_K$ and $\tau = \alpha w_G(K)$ completes the proof. ■

■ 2.4.3 Sub-Gaussian and Sub-exponential Inequalities

Examples of Sub-Gaussian and Sub-exponential Random Variables Recall that $\beta^0(p)$ denotes the centered Bernoulli random variable with parameter $p \in [0,1]$ such that if $X \sim \beta^0(p)$, then $X = 1$ with probability p and $X = 0$ with probability $\overline{p} = 1-p$. The following theorem from [27] identifies the smallest possible sub-Gaussian parameter, $v_{\beta^0}(p)$, for $\beta^0(p)$ as a function of p. In this thesis, we will refer to $v_{\beta^0}(p)$ as the 'sub-Gaussian norm' of $\beta^0(p)$. Observe that $v_{\beta^0}(p)$ is continuous in the interval $[0,1]$ and attains its maximum value $1/4$ at $p = 1/2$.

Theorem 2.4.14 ([27], Theorem 2.1)**.** *If $X \sim \beta^0(p)$, then X is v-sub-Gaussian if and only if $v \geq v_{\beta^0}(p)$ where*

$$v_{\beta^0}(p) = \begin{cases} 0, & p \in \{0,1\}, \\ 1/4, & p = 1/2, \\ \frac{p-\overline{p}}{2(\ln p - \ln \overline{p})}, & p \in (0,1)\setminus\{1/2\}. \end{cases} \tag{2.9}$$

A simple and useful fact about sub-Gaussian random variables is that if X is sub-Gaussian, then its centered square is a sub-exponential random variable.

Lemma 2.4.15. *Let X be a v-sub-Gaussian random variable and $Z = X^2 - \mathbb{E}[X^2]$. Then Z is sub-exponential with parameters $(128v^2, 8v)$.*

Proof. To prove the claim, we establish an upper bound for the MGF of Z. First of all, we express the MGF of Z as an infinite sum using the Taylor series expansion of exponential

function and the dominated convergence theorem:

$$M_Z(\lambda) = \mathbb{E}\big[e^{\lambda Z}\big] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}\big[\big(X^2 - \mathbb{E}[X^2]\big)^k\big]}{k!}.$$

Next, we note that for any real numbers $a, b \geq 0$, and positive integer k,

$$\begin{aligned}(a-b)^k \leq (a+b)^k &= \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \big(a^i b^{k-i} + a^{k-i} b^i\big) \leq \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \big(a^k + b^k\big) \\ &= 2^{k-1}(a^k + b^k).\end{aligned}$$

With $a = X^2$ and $b = \mathbb{E}[X^2]$, we get

$$\begin{aligned}M_Z(\lambda) &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^{k-1} \big\{\mathbb{E}[X^{2k}] + \big(\mathbb{E}[X^2]\big)^k\big\}}{k!} \\ &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^k \mathbb{E}[X^{2k}]}{k!} && \text{by Jensen's inequality} \\ &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k 2^k \cdot (2v)^k (2k)\Gamma(k)}{k!} && \text{by Lemma 2.4.16} \\ &= 1 + 2(4v\lambda)^2 \sum_{k=0}^{\infty} (4v\lambda)^k && \because \Gamma(k) = (k-1)! \\ &\leq 1 + 64v^2\lambda^2 && \text{for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{8v}.\end{aligned}$$

Note that $1 + 64v^2\lambda^2 \leq e^{64v^2\lambda^2}$, and therefore, Z is $(128v^2, 8v)$-sub-exponential. ■

Moments of Sub-Gaussian and Sub-exponential Random Variables One of the nice properties of sub-Gaussian random variables is that their k-th moments are finite for all $k \in \mathbb{N}$, and we have a simple upper bound for them that does not grow too rapidly as a function of k. Specifically, the upper bound involves the Gamma function, denoted by Γ. For any complex number z with positive real part, Γ can be defined via a convergent improper integral $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$. It is useful to note that $\Gamma(k) = (k-1)!$ for $k \in \mathbb{N}$, and moreover,

$$\Gamma(k/2) = C(k) \cdot \frac{(k-2)!!}{2^{\frac{k-1}{2}}} \quad \text{where} \quad C(k) = \begin{cases} \sqrt{\pi} & k \text{ is odd}, \\ 1 & k \text{ is even}, \end{cases} \tag{2.10}$$

for $k \in \mathbb{N}$. Here, $m!! = \prod_{l=0}^{\lceil m/2 \rceil - 1} (m - 2l)$ is the double factorial of m.

Lemma 2.4.16 (Moments of sub-Gaussian random variable)**.** *Let X be a v-sub-Gaussian random variable. Then for any $k \in \mathbb{N}$,*

$$\mathbb{E}\big[|X|^k\big] \le (2v)^{k/2} k\Gamma(k/2)$$

where Γ is the Gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$, which absolutely converges for any complex number z with strictly positive real part. In particular, this yields the following simpler upper bounds:

$$\big(\mathbb{E}\big[|X|^k\big]\big)^{1/k} \le \begin{cases} \sqrt{v}\cdot\sqrt{2\pi} & k = 1, \\ \sqrt{v}\cdot k^{1/2+1/k} & k \ge 2. \end{cases}$$

Proof. Since X is sub-Gaussian, its tail probabilities are bounded as

$$\mathbb{P}\left(|X| > \tau\right) \le 2e^{-\tau^2/2v} \qquad \text{for all } \tau > 0.$$

Then we observe that

$$\begin{aligned} \mathbb{E}\big[|X|^k\big] &= \int_0^\infty \mathbb{P}\left(|X|^k > \tau\right)d\tau = \int_0^\infty \mathbb{P}\left(|X| > \tau^{1/k}\right)d\tau \\ &\le 2\int_0^\infty e^{-\frac{\tau^{2/k}}{2v}}d\tau = (2v)^{k/2}k\int_0^\infty u^{k/2-1}e^{-u}du \qquad \text{where } u = \frac{\tau^{2/k}}{2v} \\ &= (2v)^{k/2}k\Gamma(k/2). \end{aligned}$$

To see the second statement, it suffices to observe that $\Gamma(k/2) \le (k/2)^{k/2}$ by Stirling's approximation. Especially when $k = 1$, we simply use the fact that $\Gamma(1/2) = \sqrt{\pi}$. ∎

There is a similar upper bound for the moments of sub-exponential random variables.

Lemma 2.4.17 (Moments of sub-exponential random variable)**.** *Let X be a sub-exponential random variable with parameter (v, c). Then for any $k \in \mathbb{N}$,*

$$\mathbb{E}\big[|X|^k\big] \le 2e^{v/2c^2}c^k k\Gamma(k).$$

Proof. Since X is (v, c)-sub-exponential, its tail probabilities are bounded as

$$\mathbb{P}\left(X > \tau\right) \le 2e^{v/2c^2}e^{-\tau/c} \qquad \text{for all } \tau > 0,$$

by applying the usual Cramér-Chernoff method, but with choosing $\lambda = 1/c$ fixed, instead of optimizing λ adaptively to τ. Then we observe that

$$\mathbb{E}\big[|X|^k\big] = \int_0^\infty \mathbb{P}\left(|X|^k > \tau\right)d\tau = \int_0^\infty \mathbb{P}\left(|X| > \tau^{1/k}\right)d\tau$$

$$
\begin{aligned}
&\leq 2e^{v/2c^2} \int_0^\infty e^{-\frac{\tau^{1/k}}{c}} dt \\
&= 2e^{v/2c^2} c^k \cdot k \int_0^\infty u^{k-1} e^{-u} du \qquad \text{where } u = \frac{\tau^{1/k}}{c} \\
&= 2e^{v/2c^2} c^k k\Gamma(k).
\end{aligned}
$$

■

Concentration of the Sum of Independent Random Variables

Theorem 2.4.18 (Hoeffding's inequality, [70])**.** *Let $X_1, \dots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ for all i almost surely. Then for any $\tau > 0$,*

$$
\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}X_i) > \tau \right) \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right).
$$

Spectral Norm of a Sub-Gaussian Random Matrix

Lemma 2.4.19. *Let $X \in \mathbb{R}^{n_1 \times n_2}$ be a v-sub-Gaussian random matrix. Then for any $\delta > 0$, the following inequality is true with probability at least $1 - \delta$:*

$$
|||X|||_2 \leq \sqrt{2v\Big\{(n_1 + n_2)\ln 9 + \log\big(1/\delta\big)\Big\}}.
$$

Proof. This proof follows the usual ϵ-net argument. First, we discretize the unit spheres $\mathbb{S}^{n_1-1}$, $\mathbb{S}^{n_2-1}$ with ϵ-nets. Let $\epsilon = 1/4$, and $\mathcal{N}_1, \mathcal{N}_2$ be ϵ-nets of $\mathbb{S}^{n_1-1}$ and $\mathbb{S}^{n_2-1}$, respectively. By Lemma 2.2.6, $|\mathcal{N}_1| \leq 9^{n_1}$ and $|\mathcal{N}_2| \leq 9^{n_2}$. Then we observe that

$$
\sup_{u \in (u,v) \in \mathcal{N}_1 \times \mathcal{N}_2} u^T X v \leq |||X|||_2 \leq \frac{1}{1 - 2\epsilon} \sup_{(u,v) \in \mathcal{N}_1 \times \mathcal{N}_2} u^T X v = 2 \sup_{(u,v) \in \mathcal{N}_1 \times \mathcal{N}_2} u^T X v.
$$

Since X is v-sub-Gaussian, $u^T X v$ is sub-Gaussian for all $(u, v) \in \mathcal{N}_1 \times \mathcal{N}_2$. Thus, the union bound yields that for any $\tau \geq 0$,

$$
\mathbb{P}\left(\sup_{(u,v) \in \mathcal{N}_1 \times \mathcal{N}_2} u^T X v > \tau \right) \leq |\mathcal{N}_1||\mathcal{N}_2| \exp\big(-\tau^2/2v\big) \leq 9^{n_1+n_2} \exp\big(-\tau^2/2v\big).
$$

To conclude the proof, it suffices to notice that

$$
9^{n_1+n_2} \exp\big(-\tau^2/2v\big) \leq \delta \quad \text{if and only if} \quad \tau \geq \sqrt{2v\Big\{(n_1 + n_2)\ln 9 + \log\big(1/\delta\big)\Big\}}.
$$

■

MGF of Sub-Gaussian Chaos of Order 2 We review the concentration of quadratic forms of the type

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j = X^T A X$$

where $A = (a_{ij})$ is an $n \times n$ matrix of coefficients, and $X = (X_1, \ldots, X_n)$ is a random vector with independent coordinates. Such a quadratic form is known as a chaos (of order 2) in probability theory.

When X_i's are sub-Gaussian random variables (e.g., Gaussian or Rademacher), the quadratic form $X^T A X$ is sub-exponential. The following upper bound is well known, and can be used to derive a Bernstein-type exponential concentration results (e.g., Hanson-Wright inequality) for $X^T A X$. Its proof is based on standard techniques such as decoupling and comparison to Gaussian chaos. We omit the proof here and refer the interested readers to [156, Sections 6.1 & 6.2] for more details.

Lemma 2.4.20 (MGF of sub-Gaussian chaos of order 2)**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian coordinates with sub-Gaussian parameter v, and let A be an $n \times n$ matrix with zero diagonal. Then $X^T A X$ is sub-exponential with parameters $(c_1 \|A\|_F^2 v,\ c_2 \|A\|_{op})$ for some absolute constants $c_1, c_2 > 0$, i.e.,*

$$\mathbb{E} \exp\left(\lambda X^T A X\right) \le \exp\Big(\frac{\lambda^2}{2} c_1 \|A\|_F^2 v\Big), \qquad \text{for all } \lambda \text{ s.t. } |\lambda| \le \frac{1}{c_2 \|A\|_{op}}.$$

Maximal Inequalities The following simple maximal inequality is well known, and it is asymptotically sharp if the random variables are i.i.d. Gaussian.

Lemma 2.4.21. *Let $X_1, \ldots, X_N$ be sub-exponential random variables with parameters (v, c). Then*

$$\mathbb{E}\Big[\max_{i \in [N]} X_i\Big] \le \max\big\{\sqrt{2v \log N},\ 2c \log N\big\}.$$

Proof. For any $\lambda \in (0, 1/c]$,

$$\begin{aligned}
\mathbb{E}\Big[\max_{i \in [N]} X_i\Big] &= \frac{1}{\lambda} \mathbb{E}\Big[\log \exp\Big(\lambda \max_{i \in [N]} X_i\Big)\Big] \le \frac{1}{\lambda} \log \mathbb{E}\Big[\exp\Big(\lambda \max_{i \in [N]} X_i\Big)\Big] \qquad \because \text{Jensen's inequality} \\
&= \frac{1}{\lambda} \log \mathbb{E}\Big[\max_{i \in [N]} \exp\big(\lambda X_i\big)\Big] \le \frac{1}{\lambda} \log\Big(\sum_{i=1}^{N} \mathbb{E}\Big[\exp\big(\lambda X_i\big)\Big]\Big) \\
&\le \frac{1}{\lambda} \log\Big(\sum_{i=1}^{N} e^{\frac{\lambda^2 v}{2}}\Big) \qquad \because \text{sub-exponential} \\
&= \frac{\log N}{\lambda} + \frac{\lambda v}{2}.
\end{aligned}$$

It remains to choose λ in the interval $(0, 1/c]$ to optimize the upper bound. If $\sqrt{2\log N/v} \leq 1/c$, then we choose $\lambda = \sqrt{2\log N/v}$ to get $\mathbb{E}\big[\max_{i\in[N]} X_i\big] \leq \sqrt{2v\log N}$. On the other hand, if $\sqrt{2\log N/v} \leq 1/c$, then we choose $\lambda = 1/c$ to get $\mathbb{E}\big[\max_{i\in[N]} X_i\big] \leq 2c\log N$ since $v/2c \leq \sqrt{2\log N/v} \leq c\log N$. ■

Matrix Bernstein Inequalities The self-adjoint[2] dilation of a rectangular matrix $M \in \mathbb{R}^{n_1\times n_2}$ is

$$\mathscr{S}(M) = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \in \boldsymbol{S}^{n_1+n_2}. \tag{2.11}$$

It is easy to verify that $\lambda_1\big(\mathscr{S}(M)\big) = |||\mathscr{S}(M)|||_2 = |||M|||_2$. We can use the self-adjoint dilations to extend the following results for self-adjoint matrices to rectangular matrices.

The following concentration inequalities, commonly referred to as matrix Bernstein inequalities, exhibit that the spectral norm of a sum of independent random matrices is well controlled. We present two versions that appear in Tropp [147] – Theorem 2.4.22 is easier to comprehend and more commonly used in the literature, however, we will need Theorem 2.4.23 in our analysis in Chapter 3 (in the proof of Theorem 3.4.3). Here we state both versions for potential future reference.

Theorem 2.4.22 (Matrix Bernstein, bounded case; [147], Theorem 6.1)**.** *Let $X_1, \ldots, X_N \in \boldsymbol{S}^n$ be independent random matrices such that $\mathbb{E}\, X_i = 0$ and $|||X_i|||_2 \leq L$ almost surely for all $i \in [N]$. Let $Z = \sum_{i=1}^N X_i$, and $v(Z) = \Big|\Big|\Big|\sum_{i=1}^N \mathbb{E}[X_i^2]\Big|\Big|\Big|_2$. Then for all $\tau \geq 0$,*

$$\begin{aligned}\mathbb{P}\left(|||Z|||_2 \geq \tau\right) &\leq n\cdot\exp\left(-\frac{v}{L^2}\cdot h\Big(\frac{L\tau}{v}\Big)\right)\\ &\leq n\cdot\exp\left(\frac{-\tau^2/2}{v(Z)+L\tau/3}\right),\end{aligned}$$

where the function $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

Theorem 2.4.23 (Matrix Bernstein, sub-exponential case; [147], Theorem 6.2)**.** *Let $X_1, \ldots, X_N \in \boldsymbol{S}^n$ be independent random matrices and $A_1, \ldots, A_N \in \boldsymbol{S}^n$ be fixed matrices such that for each $i \in [N]$,*

$$\mathbb{E}\, X_i = 0 \qquad \text{and} \qquad \mathbb{E}[X_i^k] \preceq \frac{k!}{2}L^{k-2}A_i^2 \quad \text{for } k = 2, 3, 4, \ldots$$

Let $Z = \sum_{i=1}^N X_i$, and $v = \Big|\Big|\Big|\sum_{i=1}^N A_i^2\Big|\Big|\Big|_2$. Then for all $\tau \geq 0$,

$$\mathbb{P}\left(|||Z|||_2 \geq \tau\right) \leq n\cdot\exp\left(\frac{-\tau^2/2}{v+L\tau}\right).$$

[2] In this thesis, we only consider real matrices, and thus, $M^* = M^T$.

Part II

Imputation with Matrix Completion for Predictive Modeling

Chapter 3

Data Imputation with Matrix Completion

3.1 Introduction to Part II

In this part of the thesis, we consider two types of decision-making problems in the presence of missing data and study the effectiveness of data imputation in solving these problems. Specifically, we consider imputing tabular datasets using low-rank matrix completion methods. To this end, in this chapter (Chapter 3), we begin our discussion by providing a brief literature review on data imputation and matrix completion. Then we discuss provable guarantees for the estimation error of matrix completion algorithms, including some novel results. Our analysis in this chapter is focused on a specific algorithm (singular value thresholding) for concreteness, however, any matrix completion algorithm may be used for the applications to be discussed in later chapters as long as their resulting matrix estimator has small error.

In the subsequent chapters, we consider two specific learning tasks, and argue the utility of data imputation based on low-rank matrix completion. In Chapter 4, we study errors-in-variables regression problem, and argue that missing data does not threaten the predictive performance of regression in a typical scenario. More precisely, we establish an upper bound on the prediction error of the ridgeless regression estimator when the imputed data is used. In Chapter 5, we examine model-free reinforcement learning in the setup of Q-learning, and discuss how the low-rank structure of the optimal Q-function can be utilized to reduce the burden of exploration and improve the sample efficiency. The key idea is that one can foresee the values of the Q-function at unexplored state-action pairs by means of imputing Q-values based on the observed data at a few explored state-action pairs.

3.1.1 A Brief History of Handling Missing Data

Missing data are ubiquitous across many scientific disciplines and they can occur for various reasons. For example, in typical survey data, a considerable fraction of respondents refuse to answer some sensitive questions like income, or they might just overlook some questions. In longitudinal studies where a measurement is repeated for a certain period of time, the

participants may drop out before the study ends. Missing data pose serious challenges in statistical analysis of the data because nearly all standard statistical methods presume that every case in the dataset has information on all the variables to be included in the analysis [8, 95].

For decades, a wide variety of attempts have been made to 'fix' the data. Perhaps the simplest solution is to exclude all the cases that have any missing data for any of the variables in the intended analysis. This strategy is commonly known under the name of *listwise deletion* or *complete case analysis*. Despite its simplicity, listwise deletion has an apparent disadvantage that some of available information is wasted, and such deletion may lead to biased conclusions unless the pattern of missing data is statistically independent of their values. Moreover, it may get worse when the dataset and the analysis involve more variables because all the cases that have any missing data need to be discarded in this approach.

Many alternative methods to listwise deletion have been proposed, and some of them share a principle to generate a synthetic value to replace the missing value in the data prior to analysis. These approaches are called *imputation*[1] methods. Imputation is an attractive strategy because it 'pre-processes' data to generate a complete dataset so that standard tools for statistical analysis can be applied without any adjustments. Now the remaining question is: "what are plausible values to replace the missing records in data?" There are several techniques for imputation such as hot-deck imputation, mean imputation, and regression imputation [53, Section 2] to name a few. These heuristic imputation methods and more sophisticated variants continue to enjoy widespread use in practice for their simplicity and flexibility [114, 153] although they are theoretically less understood.

The major concerns raised about imputation methods are related to the potential bias which the imputation might cause. For example, the imputed values are likely biased estimates of the 'true unobserved value' unless data are missing completely at random (MCAR) [121]. Moreover, imputing missing data with deterministic values could either underestimate (mean imputation) or overestimate (regression imputation) the correlations between variables, thereby leading to potentially incorrect conclusions in inference [8, 95]. As a result, it is widely acknowledged by statisticians that imputation is a fascinating idea to handle missing data, however, it also has pitfalls of distorting the conclusion of statistical analysis. In the words of Dempster and Rubin [45]:

> "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."

[1] Sometimes called *single* imputation, in contrast with the multiple imputation methods to be described.

Major breakthroughs beyond imputation came in 1970s and 80s with the advent of more advanced methods based on maximum likelihood estimation (MLE) [15, 46] and multiple imputation methods based on Bayesian statistics [122, 123]. Relative to traditional approaches, these techniques are less prone to bias in parameter estimation, and are regarded as the current state of the art [124, 53, 95].

Assuming a parametric model for the data, MLE-based methods attempt to find the best parameter configuration that describes the observed data, and then generate the most probable values for the missing entries. The celebrated expectation-maximization (EM) algorithm proposed by Dempster, Laird, and Rubin [46] is a prominent example of this approach. These MLE-based approaches have been very successful in both theory and practice, however, they require a probabilistic data generative model in parametric form (i.e., likelihood), and moreover, they usually do not come with global convergence guarantees.

Multiple imputation methods are variants of imputation that create multiple completed datasets rather than just one by drawing imputed values from a distribution (e.g., by Marcov chain Monte Carlo methods), and combine the analyses of each of the datasets to yield a conclusion. These approaches also require the access to a data generative model to simulate the multiple completed datasets.

■ 3.1.2 Imputing Tabular Data with Matrix Completion

In this thesis, we consider handling missing data in tabular datasets by imputing the missing values using low-rank matrix completion techniques. Tabular data are one of the most commonly encountered formats of data – a collection of data vectors, longitudinal (panel) data, and 2-dimensional images to name a few. Moreover, many real-world datasets tend to have low-rank structure and numerous techniques have been developed to exploit it in text analysis [44, 113], genomics [26, 162], link prediction in social networks [92], movie preference learning [17] and many other applications.

Indeed, low-rank approximations of data tables are already frequently employed in various fields of machine learning to impute missing data or extract low-dimensional features [83, 152, 157]. However, earlier works in this direction placed more emphasis on proposing methodologies for data imputation with low-rank matrix completion, rather than providing rigorous theoretical justifications. It was partly because the early works in the matrix completion literature were mostly concerned with analyzing the matrix estimation error measured in Frobenius norm (i.e., mean squared error of data imputation), which may not suffice to guarantee the success of the ultimate statistical task of interest. Nevertheless, as matrix completion techniques and the related low-rank heuristics are permeating through a broader research community and gaining increased popularity as a useful sub-routine, error guarantees

Table 3.1: Summary of our results and comparison with selected works from the literature. Here, n denotes the number of rows/columns, r denotes the rank, p is the fraction of observed entries in matrix M, and $\varphi_{\mathrm{MC}}(Z)$ is the estimator of M.

	Noise type	Algorithm	$\left\|\varphi_{\mathrm{MC}}(Z) - M\right\|_F$	$\|\varphi_{\mathrm{MC}}(Z) - M\|_{2,\infty}$
Candes-Plan [31]	deterministic	Nuc. norm min.	$n^{3/2}$	N/A
Keshavan et al. [80]	random	non-convex	$\sqrt{\frac{nr\log n}{p}}$	N/A
Koltchinskii et al. [82]	random	spectral (soft SVT)	$\sqrt{\frac{nr\log n}{p}}$	N/A
Negahban-Wainwright [106]	random	Nuc. norm min. (w/ addtl. constraints)	$\sqrt{\frac{nr\log n}{p}}$	N/A
Chen et al. [39]	random	Nuc. norm min. as well as non-convex	$\sqrt{\frac{nr}{p}}$	$\sqrt{\frac{r^2\log n}{p}}$
This thesis	**random**	**spectral (hard SVT)**	$\sqrt{\frac{nr\log n}{p}}$ (Theorem 3.4.3)	$\sqrt{\frac{r\log^2 n}{p^2}}$ (Theorem 3.4.5)

for matrix completion beyond Frobenius norm are being actively studied [38, 39, 125].

One of the main objectives in this part of the thesis is to argue that we are in *"situations where the missing data problem is sufficiently minor that it can be legitimately handled"* by means of low-rank-matrix-completion-based imputation, in typical settings of supervised and reinforcement learning. To this end, we begin our discussion by studying the estimation error of matrix completion algorithms measured in more sophisticated metrics than the mean squared error (=squared Frobenius norm), e.g. max row ℓ_2-norm and entrywise max norm.

■ 3.1.3 Contributions and Organization of the Chapter

The main contribution of this chapter is the novel analysis of singular value thresholding algorithm for matrix completion. More precisely, we prove an upper bound for the matrix estimation error measured in the spectral norm that matches the upper bound and the minimax lower bound reported by Koltchinskii et al. [82, Theorems 12 & 13], but with a conceptually more straightforward proof. Moreover, we analyze the estimation error in more stringent senses, using the $\ell_{2,\infty}$-norm and the ℓ_∞-norm by converting the aforementioned spectral norm analysis via Wedin's $\sin\Theta$ theorem. Our results for the error bounds of SVT algorithm are summarized in Table 3.1 along with the results from a few selected works from the literature for a quick comparison.

Although our analysis in this chapter is focused on the singular value thresholding algorithm and the resulting matrix estimator for concreteness, any matrix completion algorithm may be used for the applications to be discussed in later chapters as long as the resulting

matrix estimator has small estimation error. Thus, we also provide a brief overview of existing algorithms for low-rank matrix completion and their analysis to help the readers get a better sense of the typical level of estimation error to expect.

The rest of this chapter is organized as follows. In Section 3.2, we formally state the problem of low-rank matrix completion and review some well-known algorithms and their analyses selected from the literature. In Section 3.3, we describe the simple singular value thresholding (SVT) algorithm for matrix completion. Section 3.4 contains the main theoretical contribution of this chapter, i.e., the error analysis of the simple SVT algorithm. In Section 3.5, we present findings from numerical simulations that support our claims in the chapter. In Section 3.6, we summarize and discuss the error guarantees obtained in this chapter. Lastly, the proofs of the technical results from this chapter can be found in Section 3.7.

■ 3.2 Low-rank Matrix Completion

■ 3.2.1 Problem Statement

Let $M \in \mathbb{R}^{n_1 \times n_2}$ denote the matrix to estimate. Suppose that we have access to the matrix M through only a fraction of its elements, possibly corrupted by noise. Formally, let $\overline{\mathbb{R}} = \mathbb{R} \cup \{*\}$ where $*$ is a symbol to denote 'unknown,' and $Z \in \overline{\mathbb{R}}^{n_1 \times n_2}$ such that for $(i, j) \in [n_1] \times [n_2]$,

$$Z_{ij} = \begin{cases} M_{ij} + E_{ij} & \text{if } (i,j) \in \Omega, \\ * & \text{if } (i,j) \notin \Omega \end{cases} \tag{3.1}$$

where $\Omega \subseteq [n_1] \times [n_2]$ denotes the index set of observed elements and $E \in \mathbb{R}^{n_1 \times n_2}$ denotes the noise matrix. The noise E can be either random stochastic noise or deterministic error, depending on the specific problem setup. The objective of matrix completion problem is to estimate M from Z, which can be stated as follows.

Problem 3.1 (Matrix completion, exact recovery)**.** Given matrices M and Z generated as per (3.1), is there an estimator $\varphi_{\mathrm{MC}} : \overline{\mathbb{R}}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ such that $\varphi_{\mathrm{MC}}(Z) = M$?

Note that Problem 3.1 is ill-posed even in the noiseless case ($E = 0$), unless we make additional assumptions on M. It is because there are $n_1 n_2$ degrees of freedom in an $n_1 \times n_2$ matrix, whereas the constraint $\varphi_{\mathrm{MC}}(Z) = M$ consists of only $|\Omega| < n_1 n_2$ equations. A typical model assumption imposed on M is that $\mathrm{rank}(M) \ll n_1 \wedge n_2$. The matrix completion problem with low rank assumption is commonly referred to as the low-rank matrix completion. It is remarkable that under some assumptions that are standard by now (e.g., incoherence of the subspaces), exact recovery is possible as soon as $|\Omega|$ exceeds the degree of freedom in the model for M, and moreover, there are efficient algorithms to compute $\varphi_{\mathrm{MC}}(Z)$ [32, 116].

Nevertheless, exact recovery makes sense only when $E = 0$, which is likely far from reality in many statistical settings because of measurement error, model mismatch, etc. Therefore, our discussion in this thesis focuses on the approximate recovery problem stated next.

Problem 3.2 (Matrix completion, approximate recovery)**.** Is there an estimator $\varphi_{\mathrm{MC}} : \overline{\mathbb{R}}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ such that $\|\varphi_{\mathrm{MC}}(Z) - M\|$ is small in an appropriate matrix norm $\|\cdot\|$?

■ 3.2.2 Brief Literature Survey on Low-rank Matrix Completion

Low-rank matrix completion problem (Problem 3.1) can be formulated as the following rank minimization problem:

$$\begin{aligned} \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\text{minimize}} \quad & \operatorname{rank} X \\ \text{subject to} \quad & X_{ij} = M_{ij}, \quad \forall (i,j) \in \Omega. \end{aligned}$$

This problem has a long history as a mathematical problem [40], and also has received much attention in 1990s and 2000s in the context of collaborative filtering (also referred to as collaborative prediction or recommender systems) [63]. Algorithms proposed in early works rely on the low-rank approximation such as PCA [64], with heuristic variations including EM updates [135] and convex relaxation from low-rank to low-norm objective [136].

A major breakthrough was made with the seminal work of Candès and Recht [32], which was inspired by the success of compressed sensing [34, 48]. Candès and Recht reformulated the problem of low-rank matrix completion as a matrix-analogue of vector compressed sensing, influenced by the nuclear norm (a.k.a. trace-norm) relaxation of the rank objective function [59] and the conditions to guarantee the success of the nuclear-norm heuristic studied in [116]. Specifically, Candès and Recht argue that most low-rank matrices can be exactly estimated with high probability (with respect to the randomness in Ω) by solving the following convex program:

$$\begin{aligned} \underset{X \in \mathbb{R}^{n_1 \times n_2}}{\text{minimize}} \quad & \|X\|_* \\ \text{subject to} \quad & X_{ij} = M_{ij}, \quad \forall (i,j) \in \Omega. \end{aligned} \tag{3.2}$$

As discussed in [32, 116], the nuclear norm of a matrix can be characterized as the optimum of a semidefinite program (SDP), and thus, the problem (3.2) can be solved efficiently via semidefinite programming.

Thereafter, the problem of low-rank matrix completion has received much attention of researchers, and the theoretical efforts focused on two directions: (1) inventing a computationally faster algorithm to solve the low-rank matrix completion problem; and (2) understanding the statistical nature of the resulting matrix estimator $\hat{M} = \varphi_{\mathrm{MC}}(Z)$ in approximate recovery.

Algorithms On the algorithmic side, many attempts were made to approximately solve the Lagrangian form of (3.2), i.e.,

$$\underset{X\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij}-Z_{ij})^2+\lambda\|X\|_* \tag{3.3}$$

where $\lambda \geq 0$ is a regularization parameter (Lagrange multiplier). These efforts include the early attempts to solve (3.3) with faster first-order methods [29, 100], and the proposals based on the low-rank factorization of SDPs due to Burer and Monteiro [28]. The latter approaches start from representing the decision variable $X = UV^T$ as a product of two factor variables and proceed by solving the following program:

$$\underset{U\in\mathbb{R}^{n_1\times r},V\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \quad \frac{1}{2}\sum_{(i,j)\in\Omega}\left[(UV^T)_{ij}-Z_{ij}\right]^2+\lambda\Big(\frac{1}{2}\|U\|_F^2+\frac{1}{2}\|V\|_F^2\Big). \tag{3.4}$$

These include [79, 73], and are commonly referred to as 'non-convex optimization' because the problem (3.4) is not convex with respect to the new decision variables U, V. It is known that if the optimal solution of (3.3) has rank at most r, then the two formulations, (3.3) and (3.4), are equivalent [116, Lemma 5.1].

In addition to the attempts to solve the nuclear-norm minimization faster, some works proposed to use a different relaxation of rank, e.g., max-norm [136, 90], or to just use a spectral algorithm after imputing the missing values with 0 without solving an optimization problem [82, 37]. We refer the interested readers to the recent survey [107, Figure 4 & Table 5] for a more detailed exposition of existing algorithms from a practical viewpoint.

Error Analysis When it comes to the error analysis, Candès and Plan [31] first studied the stability of the convex program (3.3) in the presence of arbitrary bounded noise. However, their results yield a conservative upper bound on $\|\varphi_{\mathrm{MC}}(Z)-M\|_F$ when the noise is random. To elaborate on this point, let's suppose that each index pair (i,j) belongs to Ω independently with probability p, and that E_{ij} are independent and identically distributed η^2-sub-Gaussian random variables. We also assume $n_1 = n_2 = n$ for simplicity. Then the results of Candès and Plan yield $\|\varphi_{\mathrm{MC}}(Z)-M\|_F \lesssim \eta n^{3/2}$.

Subsequently, modified versions of the convex program (3.3) were studied to achieve sharper error bounds. Most notably, Negahban and Wainwright [106] considered the estimator obtained by solving (3.3) with an additional constraint $\|X\|_\infty \leq \alpha$ for some preset parameter $\alpha > 0$ that controls the 'spikiness' of the matrix entires; and Koltchinskii, Lounici, and Tsybakov [82] replaced the problem data $\{Z_{ij} : (i,j) \in \Omega\}$ with its expectation under the uniform sampling model, i.e., $\{pM_{ij} : (i,j) \in [n_1]\times[n_2]\}$. The reformulation of Koltchinskii et al. effectively leads to a spectral method – a single round of soft singular value thresholding

on a rescaled zero-padded matrix. The resulting error bounds from both works can be read as $\|\varphi_{\mathrm{MC}}(Z) - M\|_F \lesssim \{\|M\|_\infty \vee \eta\}\sqrt{nr/p}$ up to some logarithmic factors. This upper bound made an improvement from the results of Candès and Plan by a factor of $\sqrt{n^2p/r}$, and is known to be minimax rate-optimal for the class of matrices with bounded spikiness.

Despite the rate optimality, there are two drawbacks of the above results. First, it is somewhat counter-intuitive that the estimation error increases proportionally to the 'signal amplitude' $\|M\|_\infty$. Indeed, Keshavan, Montanari, and Oh reported that it is possible to achieve $\|\varphi_{\mathrm{MC}}(Z) - M\|_F \lesssim \eta\sqrt{nr/p}$ up to some logarithmic factors by solving (3.4) with spectral initialization followed by projected gradient descent on Grassmannian manifolds under some ad hoc conditions [80]. Recently, Chen et al. showed the convex-optimization-based estimator from (3.3) is very close to the estimator obtained from the nonconvex formulation (3.4), thereby proving a similar upper bound that holds for both the convex estimator and the nonconvex estimator, under some technical conditions [39].

Secondly and more importantly, the results mentioned above and many others found in the literature are primarily concerned with controlling the Frobenius norm and/or the spectral norm (the dual of the regularizing norm used) of the estimation error. An error bound in Frobenius norm might suffice in some cases, especially when estimating the matrix M is the eventual goal. However, if the matrix is imputed for the purpose of statistical analysis beyond just restoring the underlying matrix M, then controlling the average error might not suffice. This is the main motivation of the study in this chapter.

Related Work on Matrix Completion Not Covered in this Section There is a vast literature on matrix completion and it is not our intention to cover them all here. While we presented some of the major recent advances in low-rank matrix completion, our survey is focused on optimization-based approaches and is certainly far from exhaustive. Some of the interesting contributions that are not discussed in this section include the 'feature-based' and 'similarity-based' approaches toward matrix completion. These approaches view the matrix as a table of responses generated by an underlying bivariate function, and attempt to estimate the function based on the explicit feature information associated to row/column entities [105, 115], or surrogate features learned from the behavioral patterns of the response [2, 88, 91]. Actually, these types of methods precede the rank minimization approaches for matrix completion in time, and were extensively studied in the context of recommender systems. We refer the interested readers to a comprehensive handbook on recommender systems [119, Chapters 2 & 3] for more detailed expositions about the content-based and neighbor-based methods.

■ 3.3 Simple Singular Value Thresholding Algorithm

In this section, we describe a simple one-step singular value thresholding (SVT) procedure. In short, the simple SVT outputs the estimate $\hat{M}$ of M by thresholding the singular values of $\frac{n_1 n_2}{|\Omega|}\overline{Z}$ where $\overline{Z}$ is obtained from Z by replacing $*$ with 0. We call it 'simple' singular value thresholding to distinguish it from iterative thresholding methods [29, 100], which approximately solve the nuclear minimization problem with SVT.

Various forms of one-step thresholding methods have been studied in the literature, and strong theoretical guarantees were obtained for them. For example, Koltchinskii et al. [82] investigated a soft thresholding method pointing the equivalence between its resulting estimator and the minimizer of population version of nuclear-norm penalized problem. They also show that the soft SVT is minimax optimal up to a logarithmic factor. Klopp [81] considered a hard thresholding procedure in the same vein, with the nuclear norm penalty replaced with the rank penalty. Chatterjee [37] proposed a universal SVT – a hard SVT followed by rectification – that can be applied to a variety of matrix estimation problems including matrix completion. Our method is a hard SVT, and thus, closest to the method studied by Klopp.

■ 3.3.1 Best Low-rank Approximation

For $r \in \mathbb{N}$, we let $\mathcal{P}_r : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ denote the projection to the set of $n_1 \times n_2$ matrices that have rank at most r. More precisely,

$$\mathcal{P}_r(M) \in \underset{\substack{X \in \mathbb{R}^{n_1 \times n_2} \\ \mathrm{rank}(X) \le r}}{\arg\min} \|X - M\|_F. \tag{3.5}$$

It is known that $\mathcal{P}_r(M)$ can be explicitly written as a truncated SVD of M. If $M = U\Sigma V^T = \sum_{i=1}^{n_1 \wedge n_2} \sigma_i u_i v_i^T$ is a SVD of M such that $\sigma_1 \ge \cdots \ge \sigma_{n_1 \wedge n_2}$, where $\sigma_i = \Sigma_{ii}$ and u_i, v_i are the i-th column of U, V, respectively, then $\sum_{i=1}^{r} \sigma_i u_i v_i^T$ is a minimizer of the problem (3.5). Moreover, this is the unique minimizer if and only if $\sigma_r > \sigma_{r+1}$.

This construction was first given by Eckart and Young [51] for the Frobenius norm, and was later shown by Mirsky [102] to solve the problem with the norm replaced with an arbitrary unitarily invariant norm. It is especially important for us to note that

$$\mathcal{P}_r(M) = \underset{\substack{X \in \mathbb{R}^{n_1 \times n_2} \\ \mathrm{rank}(X) \le r}}{\arg\min} \|X - M\|_F = \underset{\substack{X \in \mathbb{R}^{n_1 \times n_2} \\ \mathrm{rank}(X) \le r}}{\arg\min} |||X - M|||_2 \tag{3.6}$$

as we will use the fact that $\mathcal{P}_r(M)$ is the best rank-r approximation of M in the spectral norm sense in our proof of Theorem 3.4.3.

3.3.2 Algorithm: Matrix Completion via Simple SVT

Now we describe the singular value thresholding (SVT) algorithm for matrix completion. Recall that $\Omega = \{(i,j) \in [n_1] \times [n_2] : Z_{ij} \neq *\}$ denote the set of matrix indices (i,j), for which Z_{ij} is observed. For the convenience of formal description of the SVT algorithm, we define the restriction operator

$$\mathcal{R}_\Omega : \overline{\mathbb{R}}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2} \quad \text{such that} \quad \mathcal{R}_\Omega(Z)_{ij} = \begin{cases} Z_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

The SVT estimator is described in Algorithm 1. Here, $p \in (0,1]$ and $r \in \mathbb{N}$ are some algorithmic parameters of the user's choice.

Algorithm 1: Simple singular value thresholding algorithm for matrix completion

Input: $Z \in \overline{\mathbb{R}}^{n_1 \times n_2}$, $p \in (0,1]$ and $r \in \mathbb{N}$
Output: $\hat{M} \in \mathbb{R}^{n_1 \times n_2}$

1. Estimate M as
$$\hat{M} \leftarrow \frac{1}{p} \mathcal{P}_r\big(\mathcal{R}_\Omega(Z)\big).$$

The resulting SVT estimator of M has two explicit forms that are useful in our analysis.

- (truncated SVD) Let $\mathcal{R}_\Omega(Z) = U \Sigma V^T$ be a SVD of $\mathcal{R}_\Omega(Z)$. Then $\hat{M}$ can be written as a truncated SVD that keeps only top-r singular components corresponding to the largest singular values:
$$\hat{M} = \frac{1}{p} U_r \Sigma_r V_r^T.$$

- (projection to tangent space) Additionally, $\hat{M}$ admits the following expression:
$$\hat{M} = \frac{1}{p} \Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z)\big) \tag{3.8}$$
where $T(\hat{M}) = \big\{U_r X^T + Y V_r^T : X \in \mathbb{R}^{n_2 \times r}, Y \in \mathbb{R}^{n_1 \times r}\big\}$ is the tangent space of the low-rank variety at $\hat{M}$, cf. (2.3), and $\Pi_{T(\hat{M})}$ denotes the projection onto $T(\hat{M})$.

3.4 Theoretical Guarantees for the Simple SVT Algorithm

Despite its simplicity, the simple SVT algorithm for matrix completion (Algorithm 1) turns out to produce a reasonable approximation of M. Although there are some theoretical guarantees for one-step SVT methods already available in the literature [82, 81, 37], these results are

limited to the Frobenius norm, and/or to the spectral norm. In this section, we discuss upper bounds on the matrix estimation error of the resulting estimator in three different norms – spectral norm, $\ell_{2,\infty}$-norm (i.e., $\ell_2 \to \ell_\infty$ operator norm), and ℓ_∞-norm.

To begin with, we reproduce a spectral norm upper bound on the estimation error that matches the optimal rate of the minimax lower bound reported in [82]. This rate of spectral norm bound was known in the literature, but we obtain a sharper constant with a refined analysis in the proof. Thereafter, we use the spectral norm upper bound to derive upper bounds on the matrix estimation error measured in more sophisticated norms, i.e., the $\ell_{2,\infty}$-norm (row-wise recovery) and the ℓ_∞-norm (entry-wise recovery), based on subspace perturbation argument. Here we note that our derivation of $\ell_{2,\infty}$-norm/ℓ_∞-norm bounds has a different flavor from the usual primal-dual analysis [39]. In particular, neither of our $\ell_{2,\infty}$-norm guarantee nor that of Chen et al. [39] dominates the other.

For our analysis of the simple SVT algorithm, we impose two assumptions on the model (3.1). These assumptions are standard in the literature.

Assumption 3.1. For each $(i,j) \in [n_1] \times [n_2]$, $(i,j) \in \Omega$ with probability p independently.

Assumption 3.2. The noise $E \in \mathbb{R}^{n_1 \times n_2}$ is sub-Gaussian with parameter η^2.

Additionally, we assume that the model parameters p and $r = \mathrm{rank}(M)$ are known a priori, and that they can be used as the algorithmic parameters in Algorithm 1. In practice, these values are not known a priori, and need to be estimated from data. Estimating p is easy because the empirical fraction of measured elements, $\hat{p} = |\Omega|/n_1 n_2$, sharply concentrates to p as $n_1, n_2 \to \infty$. Estimating r is a little trickier, but there are several heuristics to choose a thresholding value that separates the top r singular values from the rest [61, 37]. The key idea is that $s_i(M+E) \gtrsim \sqrt{n_1 n_2 / r}$ for all $i \in [r]$ when the top r singular values are of the same order, whereas $s_i(M+E) \lesssim \eta(\sqrt{n_1} + \sqrt{n_2})$ for $i > r$. From now on, we fix the algorithmic parameters p and r in Algorithm 1 to be the model parameters p in Assumption 3.1 and $\mathrm{rank}(M)$, respectively.

In Section 3.4.1, we present the key geometric insights in our analysis, as they will give a clear idea about what rate of the estimation error we should expect. Section 3.4.2 discusses an upper bound on the estimation error, $\hat{M} - M$ in spectral norm (Theorem 3.4.3). Thereafter, in Section 3.4.3 and Section 3.4.4, we convert this result to into the error bounds in $\ell_{2,\infty}$-norm and ℓ_∞-norm applying Wedin's $\sin\Theta$ theorem.

■ 3.4.1 Two Useful Lemmas

The key observation in our analysis is that the zero-padded data matrix $\overline{Z}$ (obtained by replacing $*$ in Z by 0) is a sub-Gaussian random matrix centered at pM. This is formally stated in the next lemma.

Lemma 3.4.1 (Key lemma)**.** *Suppose that Assumption 3.1 holds. Then the random matrix* $\mathcal{R}_\Omega(Z) - pM$ *is* $\big(v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2\big)$*-sub-Gaussian where* $v_{\beta^0}(p)$ *is the sub-Gaussian norm of a centered Bernoulli random variable; see* (2.9)*.*

Throughout this section (and also in Chapter 4), we will frequently need to control the squared ℓ_2 norm of a sub-Gaussian random vector. Thus, we state a generic upper bound in the following lemma for later use. Its proof is based on the moment bounds of sub-Gaussian random variables (Lemma 2.4.16) and the fact that the square of a sub-Gaussian random variable (after centering) is sub-exponential (Lemma 2.4.15).

Lemma 3.4.2. *Let* $X \in \mathbb{R}^n$ *be a* v*-sub-Gaussian random vector and let* $\mathcal{V} \subseteq \mathbb{R}^n$ *be a* d*-dimensional subspace. Then for any* $\delta > 0$*, the following inequality is true with probability at least* $1 - \delta$*:*

$$\|\Pi_{\mathcal{V}}(X)\|_2^2 \le \mathbb{E}\big[\|\Pi_{\mathcal{V}}(X)\|_2^2\big] + 16v\Big(\sqrt{\log\big(d/\delta\big)} \vee \log\big(d/\delta\big)\Big).$$

Moreover, $\mathbb{E}\big[\|\Pi_{\mathcal{V}}(X)\|_2^2\big] \le 4vd$*.*

The proofs of Lemma 3.4.1 and Lemma 3.4.2 can be found in Section 3.7.1.

■ 3.4.2 Error Guarantee in Spectral Norm

First of all, we argue that the simple SVT algorithm outputs an estimate $\hat{M}$ that is close to M in the spectral norm sense, as stated in the next theorem.

Theorem 3.4.3. *Let* $\hat{M}$ *be the estimate of* M *that is produced by the simple SVT algorithm (Algorithm 1). If Assumptions 3.1 and 3.2 hold, then for any* $\delta > 0$*, the following is true with probability at least* $1 - \delta$*:*

$$\big|\big|\big|\hat{M} - M\big|\big|\big|_2 \le \frac{4}{p} \cdot \left\{\sqrt{v_* \log\Big(\frac{n_1+n_2}{\delta}\Big)} \;\vee\; 2L_* \log\Big(\frac{n_1+n_2}{\delta}\Big)\right\} \tag{3.9}$$

where

$$v_* = 6\sqrt{2\pi} \cdot p \cdot (n_1 \vee n_2) \cdot \left[(1-p)\|M\|_\infty \cdot \left\{\left[1 - \Big(1 - \frac{e^2}{8\sqrt{2\pi}}\Big)p\right]\|M\|_\infty + 2\eta\right\} + \eta^2\right],$$

$$L_* = (1-p)\|M\|_\infty + \eta + \frac{e^2}{8\sqrt{2\pi}} \frac{p(1-p)\|M\|_\infty^2}{(1-p)\|M\|_\infty + \eta}.$$

The spectral norm error bound stated in Theorem 3.4.3 is obtained from the sub-exponential matrix Bernstein inequality (Theorem 2.4.23). The proof of Theorem 3.4.3 is mostly about verifying the matrix Bernstein condition, and it is deferred until Section 3.7.2.

Interpretation of Theorem 3.4.3 Now, we interpret the upper bound in Theorem 3.4.3 to better appreciate its implications. Suppose that $\|M\|_\infty, \eta = \Theta(1)$. Then $v_* \asymp p \cdot (n_1 \vee n_2)$ and $L_* \asymp 1$. Thus, we can see that

$$\sqrt{v_* \log\Big(\frac{n_1+n_2}{\delta}\Big)} \gtrsim 2L_* \log\Big(\frac{n_1+n_2}{\delta}\Big) \quad \text{if and only if} \quad p \gtrsim \frac{1}{n_1 \vee n_2} \log\Big(\frac{n_1+n_2}{\delta}\Big).$$

In fact, if $(n_1 \vee n_2) \cdot p \ll \operatorname{rank}(M) \cdot \log((n_1+n_2)/\delta)$, then estimating M is impossible even in the noiseless setting because there are infinitely many valid solutions, e.g., [33]. Therefore, we may restrict our discussion to the case where $p \gtrsim \frac{1}{n_1 \vee n_2} \log\big(\frac{n_1+n_2}{\delta}\big)$ and assume $\sqrt{v_* \log\big(\frac{n_1+n_2}{\delta}\big)}$ is the dominant term in the right-hand side of (3.9).

With the simplifications discussed above, the error bound in Theorem 3.4.3 reads as $\big|\!\big|\!\big|\hat{M} - M\big|\!\big|\!\big|_2 \lesssim \sqrt{\frac{n_1 \vee n_2}{p} \log(n_1 \vee n_2)}$, which has the same error rate as a function of n_1, n_2, p with the upper bound for soft SVT shown by Koltchinskii et al. [82, Theorem 12]. Our error bound also matches the rate of the minimax lower bound for estimating a low-rank matrix with bounded entries, reported by the same authors [82, Theorem 13].

Next, we consider the dependence of the "variance proxy[2]" of $\hat{M} - M$, i.e., v_*, on the fraction of observed elements, p. First of all, we observe that v_*/p is a decreasing function of p. Thus, we reconfirm that the error bound $\sqrt{v_*/p} \cdot \log\big((n_1+n_2)/\delta\big)$ diminishes as p increases. In particular, $\lim_{p \to 1} v_* = 6\sqrt{2\pi} \cdot (n_1 \vee n_2)\eta^2$, and thus, we get $\big|\!\big|\!\big|\hat{M} - M\big|\!\big|\!\big|_2 \lesssim \eta\sqrt{(n_1 \vee n_2)\log(n_1 \vee n_2)}$ from Theorem 3.4.3 in the limit where p converges to 1. Note that this upper bound is a function of η only, and is independent of $\|M\|_\infty$, as it should be.

Last but not least, we examine how the noise amplitude η affects our error bound. We observe that $v_* = O_\eta(\eta^2)$, and the error bound in (3.9) is asymptotically linear in η for large η. On the other hand, when η decreases to 0, v_* does not diminish to 0, but it converges to $6\sqrt{2\pi} \cdot p \cdot (n_1 \vee n_2) \cdot (1-p) \cdot \Big[1 - \big(1 - \frac{e^2}{8\sqrt{2\pi}}\big)p\Big]\|M\|_\infty^2$. Subsequently, the error bound $\frac{1}{p}\sqrt{v_* \log\big(\frac{n_1+n_2}{\delta}\big)} \approx \sqrt{\frac{n_1 \vee n_2}{p} \log\big(\frac{n_1+n_2}{\delta}\big)} \cdot (1-p)\|M\|_\infty$, which is strictly positive if and only if $p \neq 1$. As a matter of fact, this undersampling effect (error induced by missing entries) is not an artifact of our analysis, but the limitation of the SVT method itself. We will continue our discussion on this matter in Section 3.6.

Why Matrix Bernstein in Theorem 3.4.3? Before moving on, we comment on why we used matrix Bernstein inequality to only get a seemingly complicated upper bound as stated in Theorem 3.4.3. A prudent reader might observe that

$$\big|\!\big|\!\big|\hat{M} - M\big|\!\big|\!\big|_2 \le \Big|\!\Big|\!\Big|\hat{M} - \frac{1}{p}\mathcal{R}_\Omega(Z)\Big|\!\Big|\!\Big|_2 + \Big|\!\Big|\!\Big|\frac{1}{p}\mathcal{R}_\Omega(Z) - M\Big|\!\Big|\!\Big|_2 \le \frac{2}{p}\big|\!\big|\!\big|\mathcal{R}_\Omega(Z) - pM\big|\!\big|\!\big|_2.$$

[2]We call v_* the variance proxy of $\hat{M} - M$ because (3.9) suggests $\hat{M} - M$ behaves as a sub-exponential random matrix with variance v_*.

and ask why we don't attempt to directly control the spectral norm of the sub-Gaussian (as shown in Lemma 3.4.1) random matrix $\mathcal{R}_\Omega(Z) - pM$, e.g., via the usual ϵ-net argument (see Lemma 2.4.19 and its proof).

It is a sensible suggestion, however, that approach leads to an upper bound of the form

$$\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2 \lesssim \frac{1}{p}\sqrt{\left(v_{\beta^0}(p)\cdot\|M\|_\infty^2 + \eta^2\right)\Big\{(n_1+n_2) + \log\left(1/\delta\right)\Big\}},$$

which has suboptimal dependence on p: $1/p$ instead of $1/\sqrt{p}$. This is why we stick to the current version of Theorem 3.4.3 and its proof based on matrix Bernstein inequality.

Remark 3.4.4. Note that we did not assume anything about the rank of M in Theorem 3.4.3. However, if we derive an error bound in the Frobenius norm from it, then $r = \mathrm{rank}(M)$ will appear in the error bound because $\|\hat{M} - M\|_F \le \sqrt{2r}\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2$. Then we will get $\|\hat{M} - M\|_F = O(\|M\|_F)$ if and only if $p = \Omega\big(\frac{r}{n_1 \vee n_2}\log\frac{n_1+n_2}{\delta}\big)$.

■ 3.4.3 Error Guarantee in $\ell_{2,\infty}$ Norm (Row-wise Recovery)

Next, we argue that the output of the simple SVT algorithm is stable in the sense of $\ell_{2,\infty}$ norm, which is strictly stronger than the error bound in the spectral norm sense. That is, we show that $\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_{2,\infty} = \max_{i\in[n_1]}\|e_i^T(\hat{M} - M)\|_2$ is small. Before presenting the theorem statement, we recall from (3.8) that $\hat{M} = \frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z)\big)$ can be expressed as a projection form onto $T(\hat{M})$, the tangent space of the low-rank variety at $\hat{M}$.

Theorem 3.4.5. *Let $\hat{M}$ be the estimate of M that is produced by the simple SVT algorithm (Algorithm 1). Then for any $i \in [n_1]$,*

$$\left\|e_i^T(\hat{M} - M)\right\|_2^2 \le \left\|\frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)^T e_i\right\|_2^2 + \frac{s_1(M)^2}{s_r(M)^4}\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2^4.$$

Moreover, if Assumptions 3.1 and 3.2 hold, then for any $\delta_1, \delta_2 > 0$, the following is true with probability at least $1 - \delta_1 - \delta_2$:

$$\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_{2,\infty}^2 \le \frac{8r}{p^2}\cdot\left(v_{\beta^0}(p)\cdot\|M\|_\infty^2 + \eta^2\right) \tag{3.10}$$

$$+ \frac{16}{p^2}\left(v_{\beta^0}(p)\cdot\|M\|_\infty^2 + \eta^2\right)\cdot\left\{\sqrt{\log\left(\frac{2r}{\delta_1}\right)} \vee \log\left(\frac{2r}{\delta_1}\right)\right\} \tag{3.11}$$

$$+ \frac{s_1(M)^2}{s_r(M)^4}\frac{256}{p^4}\cdot\left\{\sqrt{v_*\log\left(\frac{n_1+n_2}{\delta_2}\right)} \;\vee\; 2L_*\log\left(\frac{n_1+n_2}{\delta_2}\right)\right\}^4 \tag{3.12}$$

where v_* and L_* have the same expressions as stated in Theorem 3.4.3.

Interpretation of Theorem 3.4.5 The three terms in the upper bound from Theorem 3.4.5 have intuitive meanings. The first term in (3.10) captures the typical size of the stochastic noise (undersampling + additive noise), the second term in (3.11) is the size of the confidence interval for the stochastic noise, and the third term in (3.12) is the bias in estimation, induced by misidentified subspace (tangent space). These meaning will become apparent from the proof of Theorem 3.4.5 that is included in Section 3.7.3.

Again, we suppose that $\|M\|_\infty, \eta = \Theta(1)$ to get a better sense of the magnitude of these three terms in the error upper bound. Note that $v_* \asymp p \cdot (n_1 \vee n_2)$ and $L_* \asymp 1$ as before. In addition, we suppose that $\kappa = \frac{\sigma_1(M)}{\sigma_r(M)} = \Theta(1)$. Then $s_i(M) \asymp \sqrt{\frac{n_1 n_2}{r}} \|M\|_\infty$ for all $i \in [r]$. As a result, if $p \gtrsim \frac{1}{n_1 \vee n_2} \log\left(\frac{n_1+n_2}{\delta_2}\right)$, then the order of the three terms are as follows (assuming δ_1 is sufficiently small):

$$(3.10) \asymp \frac{r}{p^2}, \qquad (3.11) \asymp \frac{1}{p^2} \cdot \log\left(\frac{r}{\delta_1}\right), \qquad (3.12) \asymp \frac{r}{p^2} \frac{n_1 \vee n_2}{n_1 \wedge n_2} \log^2\left(\frac{n_1+n_2}{\delta_2}\right).$$

Observe that the term in (3.12) is the dominant term among the three in the error bound. Therefore, we expect that the simple SVT algorithm reliably recovers every row of M in the sense that $\max_{i\in[n_1]} \|e_i^T(\hat{M}-M)\|_2 = \left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_{2,\infty} \lesssim \|M\|_{2,\infty}$ if $p \gtrsim \sqrt{\frac{r}{n_2} \frac{n_1 \vee n_2}{n_1 \wedge n_2}} \cdot \log\left(\frac{n_1+n_2}{\delta_2}\right)$.

Remark 3.4.6. Actually, the dependence of (3.12) on the aspect ratio $\frac{n_1 \vee n_2}{n_1 \wedge n_2}$ is an artifact of our analysis. We use Wedin's version of $\sin\Theta$ theorem (Theorem 2.2.5) in our analysis that yields the same upper bound for the perturbations of the row space and the column space. If we use a more refined version of subspace perturbation result that provide separate upper bounds for the row- and the column- spaces, e.g., [30, Theorem 3], then it is possible to get rid of the matrix aspect ratio and obtain the correct rate that (3.12) $\asymp \frac{r}{p^2} \frac{n_1 \vee n_2}{n_1} \log^2\left(\frac{n_1+n_2}{\delta_2}\right)$.

Remark 3.4.7. We conjecture that it might be possible to achieve improved upper bounds on the $\ell_{2,\infty}$-norm, e.g., for optimization-based methods. First, we believe $1/p$ is the right scaling with respect to p, instead of $1/p^2$. Second, the dependence on $\{\|X\|_\infty \vee \eta\}$ seems inevitable for the simple SVT, however, it could be possible to obtain $\left|\!\left|\!\left|\hat{X} - X\right|\!\right|\!\right|_{2,\infty}^2 \lesssim \eta^2 \frac{r}{p} \frac{n_1 \vee n_2}{n_1}$ for optimization-based matrix completion algorithms. These are summarized in Conjecture 3.4.8. Also, see Section 3.6 for the rationale behind these conjectures and more discussions.

However, we also conjecture that the analysis in this section is order-optimal for the SVT algorithm and cannot be improved, based on the empirical evidences, e.g., Figure 3.1b (there is nontrivial error even when $\eta \to 0$) and Figure 3.2b ($\left|\!\left|\!\left|\hat{X} - X\right|\!\right|\!\right|_{2,\infty}^2 \asymp \frac{r}{p^2}$, when $n_1 > n_2$).

Conjecture 3.4.8. *There exist matrix completion algorithms that achieve* $\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_{2,\infty}^2 \lesssim \eta^2 \frac{r}{p} \frac{n_1 \vee n_2}{n_1}$ *in the same setting as discussed in this section.*

Proof Sketch of Theorem 3.4.5 Although we provide a formal proof of Theorem 3.4.5 in Section 3.7.3, here we sketch the main ideas in the proof. Once we have an upper bound for

$\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2$, we can control the perturbation of the row and the column subspaces with it, using Wedin's $\sin\Theta$ theorem (Theorem 2.2.5):

$$\left|\!\left|\!\left| \sin\Theta\big(\mathrm{csp}(\hat{M}), \mathrm{csp}(M)\big)\right|\!\right|\!\right|_2 \vee \left|\!\left|\!\left| \sin\Theta\big(\mathrm{rsp}(\hat{M}), \mathrm{rsp}(M)\big)\right|\!\right|\!\right|_2 \leq \frac{\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2}{s_r(M)}. \tag{3.13}$$

We consider the following decomposition of the error matrix:

$$\hat{M} - M = \big(\hat{M} - \Pi_{T(\hat{M})}(M)\big) + \big(\Pi_{T(\hat{M})}(M) - M\big) \tag{3.14}$$

where $T(\hat{M})$ denotes the tangent space of the low-rank variety at $\hat{M}$, which can be viewed as a vector subspace of $\mathbb{R}^{n_1\times n_2}$; see (2.3). Note that the orthogonal projection $\Pi_{T(\hat{M})} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ can be written as

$$\Pi_{T(\hat{M})}(X) = P_{\mathrm{csp}(\hat{M})}X + XP_{\mathrm{rsp}(\hat{M})} - P_{\mathrm{csp}(\hat{M})}XP_{\mathrm{rsp}(\hat{M})}.$$

Now we observe that $\hat{M} = \frac{1}{p}\mathcal{P}_r\big(\mathcal{R}_\Omega(Z)\big) = \frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z)\big)$. Thus, the two terms on the right hand side of (3.14) are orthogonal to each other. Moreover, we can see that $\hat{M} - \Pi_{T(\hat{M})}(M) = \frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)$ is sub-Gaussian with parameter $\frac{1}{p}\big(v_{\beta^0}(p)\|M\|_\infty^2 + \eta^2\big)$ by Lemma 3.4.1. Lastly, we note that

$$\Pi_{T(\hat{M})}(M) - M = \big(P_{\mathrm{csp}(\hat{M})} - I_{n_1}\big)\cdot M \cdot \big(P_{\mathrm{rsp}(\hat{M})} - I_{n_2}\big) = P_{\mathrm{csp}(\hat{M})^\perp}P_{\mathrm{csp}(M)}\cdot M \cdot P_{\mathrm{rsp}(M)}P_{\mathrm{rsp}(\hat{M})^\perp}.$$

Therefore, for any $i \in [n_1]$,

$$\begin{aligned}
\left\|e_i^T(\hat{M} - M)\right\|_2^2 &= \left\|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)\right\|_2^2 + \left\|e_i^T\big(\Pi_{T(\hat{M})}(M) - M\big)\right\|_2^2 \\
&\leq \underbrace{\left\|\frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)^T e_i\right\|_2^2}_{\text{sub-Gaussian effect of noise+undersampling}} \qquad (3.15) \\
&\quad + \underbrace{\|M\|_2^2\left|\!\left|\!\left| \sin\Theta\big(\mathrm{csp}(\hat{M}), \mathrm{csp}(M)\big)\right|\!\right|\!\right|_2^2 \left|\!\left|\!\left| \sin\Theta\big(\mathrm{rsp}(\hat{M}), \mathrm{rsp}(M)\big)\right|\!\right|\!\right|_2^2}_{\text{error due to misspecified tangent space}}. \qquad (3.16)
\end{aligned}$$

We may interpret the term in (3.15) as the variance, and the term in (3.16) as the (squared) bias, in estimating the i-th row of M from Z. The variance term is bounded from above by the ℓ_2 norm upper bound in Lemma 3.4.2, because $\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)^T e_i$ is a sub-Gaussian vector contained in the span of $\mathrm{csp}(M)\cup\mathrm{csp}(\hat{M})$, which has rank at most $2r$. In addition, the subspace perturbation bound in (3.13) combined with the spectral norm upper bound from Theorem 3.4.3 yields an upper bound for the bias term.

■ 3.4.4 Error Guarantee in ℓ_∞ Norm (Entry-wise Recovery)

We obtain an upper bound on the estimation error $\hat{M} - M$ in the ℓ_∞ norm by the same argument as in Section 3.4.3. Specifically, we note that for any $(i, j) \in [n_1] \times [n_2]$,

$$\left\|e_i^T(\hat{M} - M)\right\|_2^2 = \underbrace{\left\|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)e_j\right\|_2^2}_{\text{'variance' due to sub-Gaussian 'noise'}} + \underbrace{\left\|e_i^T\big(\Pi_{T(\hat{M})}(M) - M\big)e_j\right\|_2^2}_{\text{'squared bias' due to subspace error}}. \tag{3.17}$$

This is the same decomposition of $\hat{M} - M$ as discussed in Section 3.4.3. We get upper bounds for each term based on the sub-Gaussianity of $\hat{M} - \Pi_{T(\hat{M})}$ and the subspace perturbation bound, which have similar forms to (3.15) and (3.16).

More precisely, $e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)e_j = \frac{1}{p}e_i\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)e_j$ is a sub-Gaussian random variable with parameter $\frac{1}{p}\big(v_{\beta^0}(p)\|M\|_\infty^2 + \eta^2\big)$ by Lemma 3.4.1. Thus, we get a similar upper bound for this 'variance' as the sum of expressions in (3.10) and (3.11), but with $2r$ (upper bound for $\dim \operatorname{span}\{\operatorname{csp}(\hat{M}) \cup \operatorname{csp}(M)\}$) in those replaced by 1. For the 'squared bias' term, we obtain the same upper bound as in (3.12).

Corollary 3.4.9. *Let $\hat{M}$ be the estimate of M that is produced by the simple SVT algorithm (Algorithm 1). Then for any $(i, j) \in [n_1] \times [n_2]$,*

$$\left\|e_i^T(\hat{M} - M)e_j\right\|_2^2 \le \left\|\frac{1}{p}e_i^T\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)e_j\right\|_2^2 + \frac{s_1(M)^2}{s_r(M)^4}\left\|\left|\hat{M} - M\right|\right\|_2^4.$$

Moreover, if Assumptions 3.1 and 3.2 hold, then for any $\delta_1, \delta_2 > 0$, the following is true with probability at least $1 - \delta_1 - \delta_2$:

$$\|\hat{M} - M\|_\infty^2 \le \frac{4}{p^2} \cdot \big(v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2\big) \tag{3.18}$$

$$+ \frac{16}{p^2}\big(v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2\big) \cdot \left\{\sqrt{\log\left(\frac{1}{\delta_1}\right)} \vee \log\left(\frac{1}{\delta_1}\right)\right\} \tag{3.19}$$

$$+ \frac{s_1(M)^2}{s_r(M)^4}\frac{256}{p^4} \cdot \left\{\sqrt{v_* \log\left(\frac{n_1 + n_2}{\delta_2}\right)} \vee 2L_* \log\left(\frac{n_1 + n_2}{\delta_2}\right)\right\}^4 \tag{3.20}$$

where v_* and L_* have the same expressions as stated in Theorem 3.4.3.

Discussion on Corollary 3.4.9 We discuss the three terms in the upper bound from Corollary 3.4.9 in the same manner as we did in Section 3.4.3. Again, we suppose that $\|M\|_\infty, \eta = \Theta(1)$ to get a better sense of the magnitude of these three terms in the error upper bound. Note that $v_* \asymp p \cdot (n_1 \vee n_2)$ and $L_* \asymp 1$ as before. In addition, we suppose that $\kappa = \frac{\sigma_1(M)}{\sigma_r(M)} = \Theta(1)$. Then $s_i(M) \asymp \sqrt{\frac{n_1 n_2}{r}}\|M\|_\infty$ for all $i \in [r]$. As a result, if $p \gtrsim \frac{1}{n_1 \vee n_2}\log\left(\frac{n_1+n_2}{\delta_2}\right)$, then the

order of the three terms are as follows (assuming δ_1 is sufficiently small):

$$(3.18) \asymp \frac{1}{p^2}, \qquad (3.19) \asymp \frac{1}{p^2} \cdot \log\Big(\frac{1}{\delta_1}\Big), \qquad (3.20) \asymp \frac{r}{p^2}\frac{n_1 \vee n_2}{n_1 \wedge n_2} \log^2\Big(\frac{n_1+n_2}{\delta_2}\Big).$$

Observe that the term in (3.20) is the dominant term among the three in the error bound, and has the same order with the term in (3.12)

It is likely that the ℓ_∞-error upper bound in Corollary 3.4.9 is suboptimal because we intuitively expect that $\|\hat{M}-M\|_\infty^2 \asymp \frac{1}{n_2}\|\hat{M}-M\|_{2,\infty}^2$. The suboptimality in our upper bound seems largely due to the limitation in our analysis. To further elaborate on this point and identify where our analysis falls short, let's briefly look into a part of our proof of Theorem 3.4.5, and see how we adapted it to prove Corollary 3.4.9. In the proof of Theorem 3.4.5, we attempt to establish an upper bound for $\big|\!\big|\!\big|P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp}\big|\!\big|\!\big|_2$, based on the observation that

$$\begin{aligned}\big\|e_i^T\big(\Pi_{T(\hat{M})}(M)-M\big)\big\|_2^2 &= e_i^T P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp} e_i \\ &\le \big|\!\big|\!\big|P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp}\big|\!\big|\!\big|_2. \qquad (3.21)\end{aligned}$$

Adapting this proof to get an upper bound for $\big|e_i^T\big(\Pi_{T(\hat{M})}(M)-M\big)e_j\big|_2^2$ as in Corollary 3.4.9, we need to control $\big|\!\big|\!\big|P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} e_j \cdot e_j^T P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp}\big|\!\big|\!\big|_2$. However, we were not able to derive any better bound than the same upper bound (3.21) used in the proof of Theorem 3.4.5.

One may ask if it could be helpful to impose additional assumptions such as the μ-incoherence of $\mathrm{csp}(M)$ to prove a better upper bound, but we do not believe so. Intuitively, when $\mathrm{csp}(M)$ is μ-incoherent and $\big|\!\big|\!\big|\hat{M}-M\big|\!\big|\!\big|_2$ is small, $\mathrm{csp}(\hat{M})$ would be $\hat{\mu}$-incoherent for some $\hat{\mu} \asymp \mu$. However, its orthogonal complement, $\mathrm{csp}(\hat{M})^\perp$, is not guaranteed to be incoherent, and actually, it is unlikely to be. Thus, it seems adding incoherence assumption would not be helpful to improve the ℓ_∞-norm bound, at least with the current proof techniques.

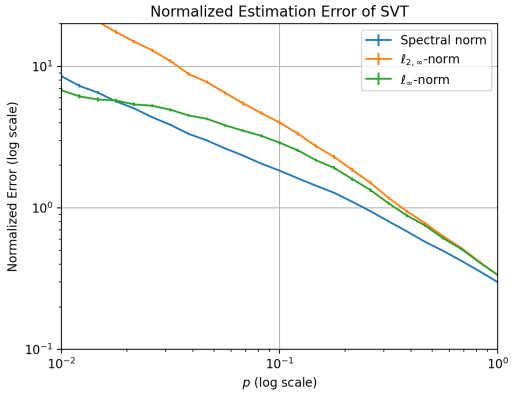
■ 3.5 Numerical Experiments

In this section, we confirm our theoretical findings from the chapter with simple numerical simulations.

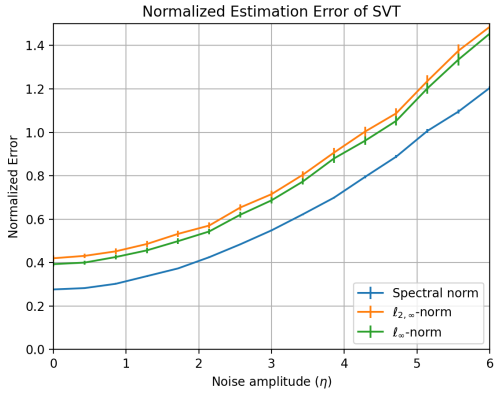
Model We construct a low-rank matrix $M = UV^T$ as the product of two random matrices $U \in \mathbb{R}^{n_1 \times r}, V \in \mathbb{R}^{n_2 \times r}$ such that $U_{ij}, V_{ij} \sim N(0,1)$ are independent and identically distributed Gaussian random variables. We let $E \in \mathbb{R}^{n_1 \times n_2}$ be a Gaussian random matrix with i.i.d. entries distributed as $N(0,\eta^2)$, and construct Ω by letting $(i,j) \in \Omega$ with probability p independently for all $(i,j) \in [n_1] \times [n_2]$.

Experiment 1 In our first experiment, we observe general trends of the estimation error $\hat{M} - M$ with respect to two model parameters, p and η. We measure the estimation error in three different norms, namely, spectral norm, $\ell_{2,\infty}$-norm, and ℓ_∞-norm. More precisely, for each parameter configuration, we generate 100 random instances of (M, E, Ω), and estimate $\hat{M}$ using the simple SVT method. Then we measure the normalized error, $\mathbb{E}_{\text{emp}}\|\hat{M} - M\|/\mathbb{E}_{\text{emp}}\|M\|$, where $\mathbb{E}_{\text{emp}}$ denotes the sample mean over the 100 instances, and $\|\cdot\|$ denotes a relevant norm. The normalized errors in the three different norms, along with confidence intervals (2 standard errors), are illustrated in Figure 3.1.

Recall that our analysis fo SVT algorithm in Section 3.4 suggests that (1) $\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2 \lesssim 1/\sqrt{p}$ whereas $\left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_{2,\infty}, \left\|\hat{M} - M\right\|_\infty \lesssim 1/p$; and (2) $\|\hat{M} - M\| \lesssim \sqrt{\|M\|_\infty^2 + \eta^2}$. These patterns are observed in Figure 3.1a (error vs p) and Figure 3.1b (error vs η), respectively.



(a) Normalized error of SVT vs p ($\eta = 3$).



(b) Normalized error of SVT vs η ($p = 0.5$).

Figure 3.1: Normalized estimation error of the simple SVT method for matrix completion. $n_1 = 200$, $n_2 = 100$, $r = 5$, and error bars represent 2 standard errors (100 runs).

Experiment 2 In our second experiment, we verify the simplified upper bounds obtained from our asymptotic analysis in Section 3.4.2 (spectral norm) and Section 3.4.3 ($\ell_{2,\infty}$-norm), which read as

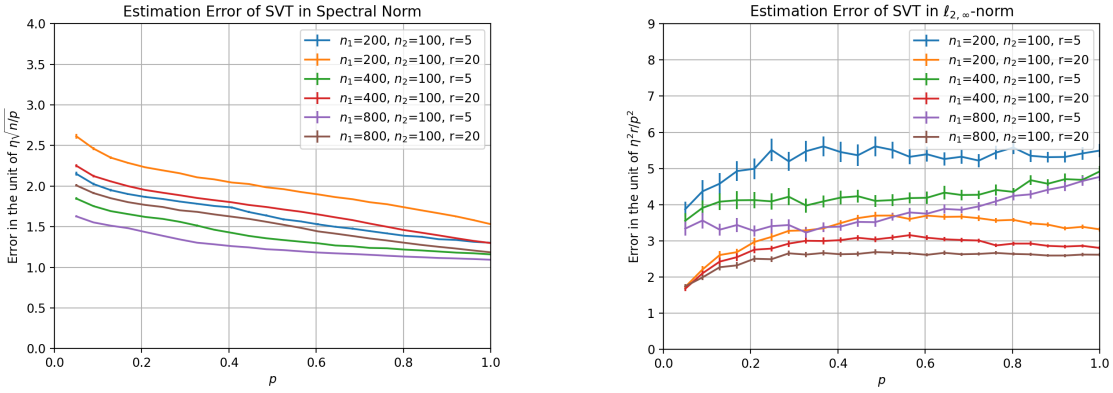
$$\begin{aligned} \left|\!\left|\!\left|\hat{M} - M\right|\!\right|\!\right|_2 &\lesssim (\|M\|_\infty \vee \eta)\sqrt{\frac{n_1 \vee n_2}{p}\log(n_1 + n_2)}, \\ \left\|\hat{M} - M\right\|_{2,\infty}^2 &\lesssim (\|M\|_\infty \vee \eta)^2 \frac{r}{p^2}\log^2(n_1 + n_2). \end{aligned} \tag{3.22}$$

We consider 6 different parameter configurations: $n_1 \in \{200, 400, 800\}, n_2 = 100, r \in \{5, 20\}$, and choose a sufficiently large η (we used $\eta = 5$) so that η dominates $\|M\|_\infty$. Again, for each parameter configuration, we generate 100 random instances of (M, E, Ω), and estimate $\hat{M}$ using the simple SVT method. For the convenience of comparison, we compare the normalized

estimation errors (motivated by (3.22), ignoring the log factors)

$$\frac{\|\!|\!|\hat{M} - M|\!|\!|_2}{\eta\sqrt{(n_1 \vee n_2)/p}}, \qquad \text{and} \qquad \frac{\|\hat{M} - M\|_{2,\infty}^2}{\eta^2 r/p^2}$$

for the 6 parameter configurations, in Figure 3.2. The numerical results suggest the correctness of the asymptotic upper bounds in (3.22).



(a) $\|\!|\!|\hat{M} - M|\!|\!|_2/\eta\sqrt{n_1 \vee n_2/p}$ vs p ($\eta = 5$). (b) $\|\hat{M} - M\|_{2,\infty}^2/(\eta^2 r/p^2)$ vs p ($\eta = 5$).

Figure 3.2: Estimation error of the simple SVT method, measured in the unit of $\eta\sqrt{n_1 \vee n_2/p}$ (spectral norm; left) and $\eta^2 r/p^2$ (squared $\ell_{2,\infty}$-norm; right).

Experiment 3 In this chapter, our theoretical analysis was focused on the SVT method. However, we believe a similar, or even stronger, error guarantees in $\ell_{2,\infty}$-norm and ℓ_∞-norm would also hold for optimization-based methods. In our third experiment, we gather empirical evidences that support this guess.

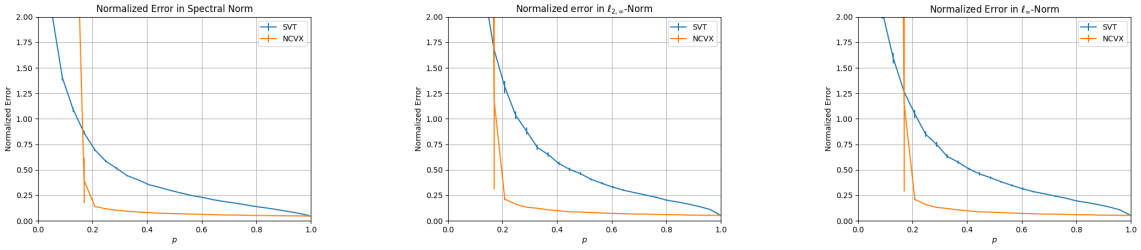
Using the estimate obtained by the SVT method as the initial point, we solve the non-convex formulation (3.4) via alternating minimization. More precisely, let $\hat{M}$ be the output of the SVT method, and $\hat{M} = \hat{U}\hat{\Sigma}\hat{V}^T$ be a compact SVD of $\hat{M}$. Starting with $U_{(0)} = \hat{U}\hat{\Sigma}^{1/2}$ and $V_{(0)} = \hat{V}\hat{\Sigma}^{1/2}$, we repeat the update

$$U_{(t)} \leftarrow \operatorname*{arg\,min}_{U \in \mathbb{R}^{n_1 \times r}} \frac{1}{2} \sum_{(i,j)\in\Omega} \left[(UV_{(t-1)}^T)_{ij} - Z_{ij}\right]^2 + \frac{\lambda}{2}\|U\|_F^2,$$

$$V_{(t)} \leftarrow \operatorname*{arg\,min}_{V \in \mathbb{R}^{n_2 \times r}} \frac{1}{2} \sum_{(i,j)\in\Omega} \left[(U_{(t)}V^T)_{ij} - Z_{ij}\right]^2 + \frac{\lambda}{2}\|V\|_F^2,$$

until convergence; let t^* denote the iteration number at which a stopping criterion is fulfilled. Eventually, we obtain the estimate $\hat{M}_{\text{ncvx}} = U_{(t^*)}V_{(t^*)}^T$.

We compare the estimation error of $\hat{M}_{\text{ncvx}}$ to that of $\hat{M}_{\text{SVT}}$ obtained by the SVT method, in the same setup as in the first experiment. That is, we let $n_1 = 200, n_2 = 100, r = 5, \eta = 1$, and generate 100 random instances of (M, E, Ω) for each value of p. Then we compare the normalized errors of the two estimators in the three different norms, along with confidence intervals (2 standard errors). The results are illustrated in Figure 3.3. Observe that the estimation errors of $\hat{M}_{\text{ncvx}}$ exhibit a radical transition around $p = 0.2$ and stay near 0 at all higher values of p, whereas the estimation errors of $\hat{M}_{\text{SVT}}$ change more gradually as p varies.



(a) Error in spectral norm. (b) Error in $\ell_{2,\infty}$-norm. (c) Error in ℓ_∞-norm.

Figure 3.3: Comparison of the SVT estimate and the solution of the nonconvex formulation (3.4), initialized with the SVT method. Here, $n_1 = 200, n_2 = 100, r = 5, \eta = 1$.

■ 3.6 Summary of the Chapter and Discussion

In this chapter, we reviewed the literature on data imputation and low-rank matrix completion. Motivated to study the estimation error caused by imputing the data matrix with matrix completion methods, we consider the simple singular value thresholding algorithm and establish error upper bounds for the resulting estimator. Specifically, we prove a spectral norm error bound that matches the minimax optimal rate. Although the strategy and the main tools used in our proof are similar to those of Koltchinskii et al. [82], we directly establish the sub-Gaussianity of $\mathcal{R}_\Omega(Z) - pM$ in Lemma 3.4.1 and then use it in the subsequent analysis. In particular, our proof does not require decoupling the under-sampling effect (p) from the additive noise (E), thereby leading to an improved constant in the resulting upper bound. Additionally, we use the spectral norm error bound as a precursor to derive estimation error bounds measured in more sophisticated norms, namely, in $\ell_{2,\infty}$- and ℓ_∞-norms, by means of classical results from subspace perturbation theory.

We conclude this chapter with making a few comments on our results.

- **(Beyond the Sub-Gaussian Noise)** Recall that we imposed Assumptions 3.1 and 3.2 for the convenience of our analysis. The assumption of i.i.d. measurement (Assumption 3.1) is essential in our analysis with the current techniques, however the assumption

of sub-Gaussian noise (Assumption 3.2) can be readily relaxed. Indeed, our analysis remains valid when the noise is sub-exponential.

- **(On the $\ell_{2,\infty}$- and ℓ_∞-Error Guarantees)** Notice that when $\|M\|_\infty, \eta = \Theta(1)$ and $p \to 1$, we get $\left|\left|\left|\hat{M} - M\right|\right|\right|_{2,\infty}^2 \lesssim r \cdot \frac{n_1 \vee n_2}{n_1 \wedge n_2} \log^2\left(\frac{n_1+n_2}{\delta}\right)$ with probability at least $1-\delta$. Observe that this is optimal up to the aspect ratio of M and a single logarithmic factor, because we expect $|||E|||_{2,\infty}^2 \asymp r \log\left(\frac{n_2}{\delta}\right)$, which sets a lower bound on $\left|\left|\left|\hat{M} - M\right|\right|\right|_{2,\infty}^2$. When $p < 1$, our upper bound is proportional to $1/p^2$. We conjecture that the right order of dependence on p is $1/p$, instead of $1/p^2$, for an 'optimal' matrix completion method. Based on the observation in Figure 3.2b, we guess this scaling of $1/p^2$ is due to the fundamental limitation of the SVT method. However, we do not know whether the scaling of $1/p$ is achievable by other matrix completion methods.

  Our ℓ_∞-norm error bound is effectively of the same order to the $\ell_{2,\infty}$-norm bound. It is because we cannot further reduce the 'bias' term that arises from the misidentification of the tangent space – see (3.17) – with the current analysis. Again, we do not know whether this is an artifact of our analysis, or it is due to the fundamental limitation of the SVT method. However, we conjecture that it might be possible to improve our analyses of simple SVT for these two norms to sharper upper bounds, e.g., with entrywise eigenvector perturbation bounds from recent advances in matrix perturbation theory [52, 1].

- **(Beyond Simple SVT Algorithm)** Our third experiment (Figure 3.3) in Section 3.5 suggests that optimization-based methods seem to possess less estimation error than the SVT method, once p surpasses a certain threshold. We believe the bias that originates from imputing the missing values with 0 in the SVT method (Algorithm 1) is the cause of this phenomenon; that is, once p is sufficiently large so that the underlying matrix can be recovered, optimization-based estimates are only prone to the measurement noise, whereas the estimate obtained by SVT additionally suffers from the noise incurred by zero-padding. Thus, establishing stronger error guarantees for optimization-based methods, e.g., error bounds in $\ell_{2,\infty}$-norm and ℓ_∞-norm, is an interesting open question.

  Recently, Chen et al. [39] proved upper bounds for entrywise error for the estimator obtained by solving the convex program (3.3) or the nonconvex program (3.4). Although their results require some additional technical assumptions on the condition number and the incoherence of singular subspaces, their results provide the best available upper bounds to the best of our knowledge. In particular, when all the nice assumptions are made, their ℓ_∞-error bound reads as $\|\varphi_{\mathrm{MC}}(Z) - M\|_\infty \lesssim \eta r/\sqrt{np}$ whereas our results lead to $\|\varphi_{\mathrm{MC}}(Z) - M\|_\infty \lesssim (\|M\|_\infty \vee \eta)\sqrt{r}/p$, ignoring the log factors. It would be an interesting open question to see if $\|\varphi_{\mathrm{MC}}(Z) - M\|_\infty \lesssim \eta\sqrt{r/np}$ is achievable.

Table 3.2: Summary of our results from this chapter, the results from Chen et al. [39], and the conjectures for best possible errors.

	$\lvert\lvert\lvert\varphi_{\mathrm{MC}}(Z)-M\rvert\rvert\rvert_2$	$\lVert\varphi_{\mathrm{MC}}(Z)-M\rVert_{2,\infty}$	$\lVert\varphi_{\mathrm{MC}}(Z)-M\rVert_{\infty}$
Our results for SVT	$\lesssim (\lVert M\rVert_\infty \vee \eta)\sqrt{\frac{n\log n}{p}}$	$\lesssim (\lVert M\rVert_\infty \vee \eta)\frac{\sqrt{r}\log n}{p}$	same as left
Results of Chen et al. [39, Theorem 1.4]	$\lesssim \eta\sqrt{\frac{n}{p}}$	$\lesssim \eta\frac{r}{\sqrt{p}}$ (obtained from right)	$\lesssim \eta\frac{r}{\sqrt{np}}$
Conjecture	$\lesssim \eta\sqrt{\frac{n}{p}}$	$\lesssim \eta\sqrt{\frac{r}{p}}$	$\lesssim \eta\sqrt{\frac{r}{np}}$

Our results, the results from [39], and our conjectures are summarized in Table 3.2.

3.7 Proofs

3.7.1 Proof of the Two Lemmas in Section 3.4.1

Proof of Lemma 3.4.1

Proof. Observe that $\mathcal{R}_\Omega(Z)-pM$ can be written as the sum of n_1n_2 independent random matrices. With $\Delta_{ij}=\mathbb{1}\{(i,j)\in\Omega\}$ being a Bernoulli random variable with parameter p,

$$\mathcal{R}_\Omega(Z)-pM=\sum_{(i,j)\in[n_1]\times[n_2]} R_{ij} \qquad \text{where} \quad R_{ij}=\big\{(M_{ij}+E_{ij})\cdot\Delta_{ij}-pM_{ij}\big\}\cdot e_ie_j^T\in\mathbb{R}^{n_1\times n_2}. \tag{3.23}$$

Observe that $\mathbb{E}[R_{ij}]=0$ for all $(i,j)\in[n_1]\times[n_2]$.

Thus, for any $U\in\mathbb{S}^{n_1\times n_2-1}=\{U\in\mathbb{R}^{n_1\times n_2}:\lVert U\rVert_F=1\}$,

$$\langle U,\mathcal{R}_\Omega(Z)-pM\rangle=\sum_{(i,j)\in[n_1]\times[n_2]}\big\{(M_{ij}+E_{ij})\cdot\Delta_{ij}-pM_{ij}\big\}\cdot U_{ij}.$$

We show that $\langle U,\mathcal{R}_\Omega(Z)-pM\rangle$ is $(v_{\beta^0}(p)\lVert M\rVert_\infty^2+\eta^2)$-sub-Gaussian in two steps.

First, we can observe that $(M_{ij}+E_{ij})\cdot\Delta_{ij}-pM_{ij}$ is sub-Gaussian with parameter $M_{ij}^2v_{\beta^0}(p)+\eta^2$ for all $(i,j)\in[n_1]\times[n_2]$ because of the tower property:

$$\begin{aligned}
\mathbb{E}\big[e^{\lambda((M_{ij}+E_{ij})\cdot\Delta_{ij}-pM_{ij})}\big]&=\mathbb{E}_{\Delta_{ij}}\Big[\mathbb{E}_{E_{ij}|\Delta_{ij}}\big[e^{\lambda((\Delta_{ij}-p)M_{ij}+E_{ij}\Delta_{ij}}\big|\Delta_{ij}\big]\Big]\\
&=\mathbb{E}_{\Delta_{ij}}\Big[e^{\lambda((\Delta_{ij}-p)M_{ij}}\cdot\mathbb{E}_{E_{ij}|\Delta_{ij}}\big[e^{\lambda E_{ij}\Delta_{ij}}\big|\Delta_{ij}\big]\Big]\\
&\overset{(*)}{\le}\mathbb{E}_{\Delta_{ij}}\big[e^{\lambda((\Delta_{ij}-p)M_{ij}}\big]\cdot\mathbb{E}_{E_{ij}}\big[e^{\lambda E_{ij}}\big]\\
&\le e^{v_{\beta^0}(p)\cdot M_{ij}^2\lambda^2/2}\cdot e^{\eta^2\lambda^2/2}.
\end{aligned}$$

Here, the inequality (*) follows from the observations that (i) $e^{\lambda((\Delta_{ij}-p)M_{ij}}$ is a nonnegative random variable; and (ii) $\mathbb{E}_{E_{ij}|\Delta_{ij}}\left[e^{\lambda E_{ij}\Delta_{ij}}\middle|\Delta_{ij}\right] \leq \mathbb{E}_{E_{ij}}\left[e^{\lambda E_{ij}}\right]$ by Jensen's inequality.

Second, we write $\mathbb{E}_{(i,j)}$ to denote the expectation conditioned on $\{\Delta_{ab}, E_{ab} : (a,b) \in [n_1]\times[n_2] \text{ and } \leq_{\text{lex}} (i,j)\}$ where $\leq_{\text{lex}}$ is the lexicographic order. Also, we let $\mathbb{E}_0$ denote the unconditional expectation. Letting $W_{ij} = (M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}$ for notational simplicity, we use the tower property again to get

$$\begin{aligned}
&\mathbb{E}e^{\sum_{(i,j)\in[n_1]\times[n_2]} W_{ij}\cdot u_{ij}} \\
&\quad= \mathbb{E}_0\mathbb{E}_{(1,1)}\mathbb{E}_{(1,2)},\dots\mathbb{E}_{(n_1,n_2-1)}e^{\sum_{(i,j)\in[n_1]\times[n_2]} W_{ij}\cdot u_{ij}} \\
&\quad\leq \mathbb{E}_0\mathbb{E}_{(1,1)}\mathbb{E}_{(1,2)},\dots\mathbb{E}_{(n_1,n_2-2)}e^{\sum_{(i,j)\in[n_1]\times[n_2]\setminus\{(n_1,n_2)\}} W_{ij}\cdot u_{ij}}\cdot e^{\frac{\lambda^2}{2}u_{n_1n_2}^2(M_{n_1n_2}^2 v_{\beta^0}(p)^2+\eta^2)} \\
&\quad\leq \dots \\
&\quad\leq e^{\frac{\lambda^2}{2}\sum_{(i,j)\in[n_1]\times[n_2]} u_{ij}^2(v_{\beta^0}(p)\cdot M_{ij}^2+\eta^2)}.
\end{aligned}$$

We conclude the proof by observing that for all $U \in \mathbb{S}^{n_1\times n_2-1}$,

$$\sum_{(i,j)\in[n_1]\times[n_2]} u_{ij}^2\big(v_{\beta^0}(p)\cdot M_{ij}^2+\eta^2\big) \leq v_{\beta^0}(p)\cdot\|M\|_\infty^2+\eta^2.$$

■

Proof of Lemma 3.4.2

Proof. Let $V \in \mathbb{R}^{n\times d}$ be a semi-orthogonal matrix such that $\text{csp}(V) = \mathcal{V}$. Observe that $\big\|\Pi_{\mathcal{V}}(X)\big\|_2^2 = \big\|V^TX\big\|_2^2 = \sum_{i=1}^d \big|\langle v_i, X\rangle\big|^2$ where v_i denotes the i-th column of V. Note that $\langle v_i, X\rangle$ is v-sub-Gaussian random variable for all $i \in [d]$. By Lemma 2.4.16, $\mathbb{E}\big[\big|\langle v_i, X\rangle\big|^2\big] \leq 4v$ for all $i \in [d]$, and therefore, $\mathbb{E}\big\|\Pi_{\mathcal{V}}(X)\big\|_2^2 \leq 4vd$.

Lastly, we can see that $\big|\langle v_i, X\rangle\big|^2 - \mathbb{E}\big[\big|\langle v_i, X\rangle\big|^2\big]$ is $(128v^2, 8v)$-sub-exponential by Lemma 2.4.15. Applying the union bound, we get

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^d\Big(|\langle v_i,X\rangle|^2 - \mathbb{E}\big[|\langle v_i,X\rangle|^2\big]\Big) > \tau\right) &\leq \sum_{i=1}^d \mathbb{P}\left(|\langle v_i,X\rangle|^2 - \mathbb{E}\big[|\langle v_i,X\rangle|^2\big] > \tau/d\right) \\
&\leq d\cdot\Big(e^{-\tau^2/256v^2}\vee e^{-\tau/16v}\Big).
\end{aligned}$$

To conclude the proof, it suffices to observe that

$$d\cdot\Big(e^{-\tau^2/256v^2}\vee e^{-\tau/16v}\Big) \leq \delta \quad \text{if and only if} \quad \tau \geq 16v\Big(\sqrt{\log\left(d/\delta\right)}\vee\log\left(d/\delta\right)\Big).$$

■

■ 3.7.2 Proof of Theorem 3.4.3

Proof. Recall that $\hat{M} = \frac{1}{p}\mathcal{P}_r\big(\mathcal{R}_\Omega(Z)\big)$ is the best rank-r approximation of $\frac{1}{p}\mathcal{R}_\Omega(Z)$ in the spectral norm, cf. (3.6). By triangle inequality and the optimality of $\hat{M}$, we get

$$\big|\!\big|\!\big|\hat{M} - M\big|\!\big|\!\big|_2 \le \left|\!\left|\!\left|\hat{M} - \frac{1}{p}\mathcal{R}_\Omega(Z)\right|\!\right|\!\right|_2 + \left|\!\left|\!\left|\frac{1}{p}\mathcal{R}_\Omega(Z) - M\right|\!\right|\!\right|_2 \le \frac{2}{p}\big|\!\big|\!\big|\mathcal{R}_\Omega(Z) - pM\big|\!\big|\!\big|_2. \tag{3.24}$$

Recall from (3.23) from the proof of Lemma 3.4.1 (or, directly observe) that $\mathcal{R}_\Omega(Z) - pM$ can be written as the sum of $n_1 n_2$ independent random matrices:

$$\mathcal{R}_\Omega(Z) - pM = \sum_{(i,j)\in[n_1]\times[n_2]} R_{ij} \qquad \text{where} \quad R_{ij} = \big\{(M_{ij} + E_{ij})\cdot\Delta_{ij} - pM_{ij}\big\}\cdot e_i e_j^T \in \mathbb{R}^{n_1\times n_2},$$

where $\Delta_{ij} = \mathbb{1}\{(i,j)\in\Omega\}$ is a Bernoulli random variable with parameter p. Observe that $\mathbb{E}[R_{ij}] = 0$ for all $(i,j)\in[n_1]\times[n_2]$.

Next, we consider the self-adjoint dilation of R_{ij}, cf. (2.11). Suppose that the so-called matrix Bernstein condition is satisfied for some fixed L and $A_{ij} \in \boldsymbol{S}_+^{n_1+n_2}$, i.e.,

$$\mathbb{E}\big[\mathscr{S}(R_{ij})^k\big] \preceq \frac{k!}{2}L^{k-2}A_{ij}^2 \qquad \text{for } k = 2,3,4,\ldots. \tag{3.25}$$

Then we can use the sub-exponential matrix Bernstein's inequality (Theorem 2.4.23) to establish an upper bound for $\big|\!\big|\!\big|\hat{M} - M\big|\!\big|\!\big|_2$.

In the rest of this proof, we verify the condition (3.25) for appropriate L and A_{ij}; see (3.33) for the expressions of those L and A_{ij}.

Verifying the matrix Bernstein condition For each $(i,j)\in[n_1]\times[n_2]$, let

$$X_{ij}^{\text{odd}} = \begin{bmatrix} 0 & e_i e_j^T \\ e_j e_i^T & 0 \end{bmatrix} \qquad \text{and} \qquad X_{ij}^{\text{even}} = \begin{bmatrix} e_i e_i^T & 0 \\ 0 & e_j e_j^T \end{bmatrix},$$

and then observe that $\mathscr{S}(R_{ij})^k = \big\{(M_{ij} + E_{ij})\cdot\Delta_{ij} - pM_{ij}\big\}^k \cdot X_{ij}$ where $X_{ij} = X_{ij}^{\text{odd}}$ if k is odd and $X_{ij} = X_{ij}^{\text{even}}$ if k is even. It is easy to see that

$$X_{ij}^{\text{odd}} \preceq X_{ij}^{\text{odd}} + 2X_{ij}^{\text{even}} \qquad \text{and} \qquad X_{ij}^{\text{even}} \preceq X_{ij}^{\text{odd}} + 2X_{ij}^{\text{even}}.$$

because

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \preceq \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \text{and} \qquad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \preceq \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

For each $(i,j) \in [n_1] \times [n_2]$, we let

$$\widetilde{A}_{ij} = \frac{\sqrt{3}}{2} \begin{bmatrix} e_i e_i^T & e_i e_j^T \\ e_j e_i^T & e_j e_j^T \end{bmatrix} + \frac{1}{2} \begin{bmatrix} e_i e_i^T & -e_i e_j^T \\ -e_j e_i^T & e_j e_j^T \end{bmatrix}. \tag{3.26}$$

so that[3] $\widetilde{A}_{ij}^2 = X_{ij}^{\text{odd}} + 2X_{ij}^{\text{even}}$. Thus,

$$\mathbb{E}\big[\mathscr{S}(R_{ij})^k\big] \preceq \mathbb{E}\Big[\big|(M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}\big|^k\Big] \cdot \widetilde{A}_{ij}^2. \tag{3.27}$$

Next, we show an upper bound for the k-th moment of $(M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}$. Note that we already know that this is sub-Gaussian random variable with parameter $v_{\beta^0}(p)\|M\|_\infty^2 + \eta^2$ by Lemma 3.4.1, where $v_{\beta^0}(p)$ is the smallest possible sub-Gaussian parameter of a centered Bernoulli random variable, defined in (2.9). However, instead of naïvely applying the generic moment upper bounds for sub-Gaussian (Lemma 2.4.16), we directly prove upper bounds for the k-th moments in this proof in order to capture the correct dependence on p. Eventually, we will get the upper bound in (3.32).

Observe that $\Delta_{ij}^l = \Delta_{ij} = \mathbb{1}\{(i,j) \in \Omega\}$ for all $l \in \mathbb{N}$ and $\Delta_{ij}^0 = 1$. Applying the binomial theorem, we have

$$\begin{aligned}
\big\{(M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}\big\}^k &= \sum_{l=0}^{k} \binom{k}{l} \cdot (M_{ij}+E_{ij})^l \cdot \Delta_{ij}^l \cdot (-pM_{ij})^{k-l} \\
&= \Delta_{ij} \cdot \sum_{l=0}^{k} \binom{k}{l} \cdot (M_{ij}+E_{ij})^l \cdot (-pM_{ij})^{k-l} + (1-\Delta_{ij})\cdot(-pM_{ij})^k \\
&= \Delta_{ij} \cdot \big\{(1-p)M_{ij} + E_{ij}\big\}^k + (1-\Delta_{ij})\cdot(-pM_{ij})^k.
\end{aligned}$$

By taking expectation of absolute value on both sides, we get

$$\begin{aligned}
\mathbb{E}\Big[\big|(M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}\big|^k\Big] &\le \mathbb{E}\Big[\Delta_{ij}\cdot\big|(1-p)M_{ij}+E_{ij}\big|^k\Big] + \mathbb{E}\Big[(1-\Delta_{ij})\cdot\big|-pM_{ij}\big|^k\Big] \\
&= p\cdot\mathbb{E}\Big[\big|(1-p)M_{ij}+E_{ij}\big|^k\Big] + (1-p)p^k|M_{ij}|^k.
\end{aligned} \tag{3.28}$$

To obtain the equality in the last line of (3.28), we used the tower property (law of total expectation) and the independence between E_{ij} and Δ_{ij} as follows:

$$\begin{aligned}
\mathbb{E}\Big[\Delta_{ij}\cdot\big|(1-p)M_{ij}+E_{ij}\big|^k\Big] &= \mathbb{E}_{\Delta_{ij}}\Big[\mathbb{E}_{E_{ij}|\Delta_{ij}}\Big[\Delta_{ij}\cdot\big|(1-p)M_{ij}+E_{ij}\big|^k\Big|\Delta_{ij}\Big]\Big] \\
&= \mathbb{E}_{\Delta_{ij}}\Big[\Delta_{ij}\cdot\mathbb{E}_{E_{ij}|\Delta_{ij}}\Big[\big|(1-p)M_{ij}+E_{ij}\big|^k\Big|\Delta_{ij}\Big]\Big]
\end{aligned}$$

[3] In fact, we defined $\widetilde{A}_{ij}$ by taking matrix square root of $X_{ij}^{\text{odd}} + 2X_{ij}^{\text{even}}$.

$$= \mathbb{E}\big[\Delta_{ij}\big] \cdot \mathbb{E}_{E_{ij}}\Big[\big|(1-p)M_{ij} + E_{ij}\big|^k\Big] \qquad \because \text{independence}$$
$$= p \cdot \mathbb{E}_{E_{ij}}\Big[\big|(1-p)M_{ij} + E_{ij}\big|^k\Big].$$

We apply the binomial theorem again to get the following upper bound on the expectation in the first term of (3.28):

$$\mathbb{E}\Big[\big|(1-p)M_{ij} + E_{ij}\big|^k\Big] = \mathbb{E}\left[\Big|\sum_{l=0}^{k}\binom{k}{l}\cdot(1-p)^l M_{ij}^l \cdot E_{ij}^{k-l}\Big|\right]$$
$$\leq \sum_{l=0}^{k}\binom{k}{l}\cdot(1-p)^l|M_{ij}|^l \cdot \mathbb{E}\big[|E_{ij}|^{k-l}\big]. \tag{3.29}$$

Recall from Assumption 3.2 that E_{ij} is η^2-sub-Gaussian. Therefore, it follows from Lemma 2.4.16 that

$$\mathbb{E}\big[|E_{ij}|^{k-l}\big] \leq (2\eta^2)^{(k-l)/2}(k-l)\Gamma\big((k-l)/2\big) \leq \sqrt{2\pi}\eta^{k-l}\cdot(k-l)!! \tag{3.30}$$

where $m!! = \prod_{l=0}^{\lceil m/2\rceil-1}(m-2l)$ is the double factorial of m. Note that we used the well known fact about the Gamma function that $\Gamma(m/2) \leq \sqrt{\pi}\frac{(m-2)!!}{2^{(m-1)/2}}$ for all $m \in \mathbb{N}$; see (2.10).

Inserting (3.30) to (3.29), we get

$$\mathbb{E}\Big[\big|(1-p)M_{ij} + E_{ij}\big|^k\Big] \leq \sum_{l=0}^{k}\binom{k}{l}\cdot(1-p)^l|M_{ij}|^l\cdot\sqrt{2\pi}\eta^{k-l}\cdot(k-l)!!$$
$$\leq \sqrt{2\pi}\cdot k!!\cdot\big\{(1-p)|M_{ij}| + \eta\big\}^k.$$

Combining this upper bound with (3.28), we obtain the following inequality:

$$\mathbb{E}\Big[\big|(M_{ij}+E_{ij})\cdot\Delta_{ij} - pM_{ij}\big|^k\Big]$$
$$\leq p\cdot\left[\sqrt{2\pi}\cdot k!!\cdot\big\{(1-p)|M_{ij}| + \eta\big\}^k + (1-p)p^{k-1}|M_{ij}|^k\right]$$
$$\leq p\sqrt{2\pi}\cdot k!\cdot\left[\big\{(1-p)|M_{ij}| + \eta\big\}^k + \frac{1}{\sqrt{2\pi}\cdot k!}(1-p)p^{k-1}|M_{ij}|^k\right]. \tag{3.31}$$

Here, we used the fact that $k!! \leq k!$ for all $k \in \mathbb{N}$. Furthermore, $1/k! \leq (e/k)^k/\sqrt{2\pi k}$ due to the Stirling's lower bound for factorials, namely, $k! \geq \sqrt{2\pi}k^{k+1/2}e^{-k}$, $\forall k \in \mathbb{N}$. We further control the sum in the square bracket from (3.31) using convexity.

For $k \geq 2$, the function $f_k : x \mapsto x^{k-1}$ is convex, and thus, $\alpha f_k(x) + \beta f_k(y) \leq (\alpha+\beta)\cdot$

$f_k\big((\alpha x+\beta y)/(\alpha+\beta)\big)$ for any $\alpha,\beta\geq 0$. Recall that $\|M\|_\infty=\max_{(i,j)\in[n]\times[d]}|M_{ij}|$. Letting

$$\alpha=(1-p)\|M\|_\infty+\eta,\quad \beta=\beta_k=\frac{e}{2\pi k^{3/2}}(1-p)\|M\|_\infty,\quad x=\alpha,\quad \text{and}\quad y=y_k=\frac{e}{k}p\|M\|_\infty,$$

we have

$$\begin{aligned}\alpha+\beta_k&=\Big(1+\frac{e}{2\pi k^{3/2}}\Big)(1-p)\|M\|_\infty+\eta,\\ \alpha x&=(1-p)^2\|M\|_\infty^2+2(1-p)\|M\|_\infty\eta+\eta^2,\\ \beta_k y_k&=\frac{e^2}{2\pi k^{5/2}}p(1-p)\|M\|_\infty^2.\end{aligned}$$

Thereafter, we continue with (3.31) to get

$$\begin{aligned}\mathbb{E}\Big[\big|(M_{ij}+E_{ij})\cdot\Delta_{ij}-pM_{ij}\big|^k\Big]&\leq p\sqrt{2\pi}\cdot k!\cdot\big(\alpha x^{k-1}+\beta_k y_k^{k-1}\big)&&\text{by (3.31)},\\ &\leq p\sqrt{2\pi}\cdot k!\cdot(\alpha+\beta_k)\Big(\frac{\alpha x+\beta_k y_k}{\alpha+\beta_k}\Big)^{k-1}.&&\because\text{convexity}\end{aligned}\tag{3.32}$$

This upper bound, together with (3.27) leads to certifying the desired matrix Bernstein condition (3.25); for $k=2,3,4,\ldots,$

$$\begin{aligned}\mathbb{E}\big[\mathcal{S}(R_{ij})^k\big]&\preceq\frac{k!}{2}\cdot\Big(\frac{\alpha x+\beta_k y_k}{\alpha+\beta_k}\Big)^{k-2}\cdot 2\sqrt{2\pi}\cdot p\cdot(\alpha x+\beta_k y_k)\cdot\widetilde{A}_{ij}^2\\ &\preceq\frac{k!}{2}\cdot\Big(\frac{\alpha x+\beta_2 y_2}{\alpha}\Big)^{k-2}\cdot 2\sqrt{2\pi}\cdot p\cdot(\alpha x+\beta_2 y_2)\cdot\widetilde{A}_{ij}^2\end{aligned}$$

because $\beta_k\geq 0$ and $\beta_k y_k\leq\beta_2 y_2$ for all $k\geq 2$. All in all, we verified the matrix Bernstein condition (3.25) with

$$\begin{aligned}L&=(1-p)\|M\|_\infty+\eta+\frac{e^2}{8\sqrt{2}\pi}\frac{p(1-p)\|M\|_\infty^2}{(1-p)\|M\|_\infty+\eta}\\ A_{ij}^2&=\Phi(\|M\|_\infty,p,\eta)\cdot\widetilde{A}_{ij}^2.\end{aligned}\tag{3.33}$$

where $\widetilde{A}_{ij}^2\in\boldsymbol{S}_+^{n+d}$ has a simple sparse form of an appropriately embedded $[2,1;1,2]$ matrix as described in (3.26), and

$$\Phi(\|M\|_\infty,p,\eta)=2\sqrt{2\pi}\cdot p\cdot\bigg[(1-p)\|M\|_\infty\cdot\bigg\{\bigg[1-\Big(1-\frac{e^2}{8\sqrt{2}\pi}\Big)p\bigg]\|M\|_\infty+2\eta\bigg\}+\eta^2\bigg].$$

Completing the proof of Theorem 3.4.3 Now that the matrix Bernstein (3.25) is verified with parameters L, A_{ij}^2 as described in (3.33), we complete the proof by applying the sub-

exponential matrix Bernstein inequality (Theorem 2.4.23). To this end, we first compute the matrix variance parameter

$$\begin{aligned} v &= \left|\!\left|\!\left| \sum_{(i,j)\in[n_1]\times[n_2]} A_{ij}^2 \right|\!\right|\!\right|_2 = \Phi(\|M\|_\infty, p, \eta) \cdot \left|\!\left|\!\left| \sum_{(i,j)\in[n_1]\times[n_2]} \widetilde{A}_{ij}^2 \right|\!\right|\!\right|_2 \\ &= \Phi(\|M\|_\infty, p, \eta) \cdot \left|\!\left|\!\left| \begin{bmatrix} 2n_2 I_{n_1} & \mathbf{1}_{n_1\times n_2} \\ \mathbf{1}_{n_2\times n_1} & 2n_1 I_{n_2} \end{bmatrix} \right|\!\right|\!\right|_2 \qquad \because \text{recall the expression of } \widetilde{A}_{ij}^2 \\ &\le 3\Phi(\|M\|_\infty, p, \eta) \cdot (n_1 \vee n_2). \qquad \because \text{Gershgorin circle theorem} \end{aligned}$$

Theorem 2.4.23 states that for all $\tau \ge 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\!\left|\!\left| \mathcal{R}_\Omega(Z) - pX \right|\!\right|\!\right|_2 \ge t \right) &\le (n_1+n_2) \cdot \exp\left(\frac{-\tau^2/2}{v + L\tau} \right) \\ &\le (n_1+n_2) \cdot \left\{ \exp\left(\frac{-\tau^2}{4v} \right) \ \vee \ \exp\left(\frac{-\tau}{4L} \right) \right\}. \end{aligned}$$

To conclude the proof, it suffices to recall from (3.24) that $\left|\!\left|\!\left| \hat{M} - X \right|\!\right|\!\right|_2 \le \frac{2}{p} \left|\!\left|\!\left| \mathcal{R}_\Omega(Z) - pX \right|\!\right|\!\right|_2$, and observe that $(n_1+n_2) \cdot \left\{ \exp(-\tau^2/4v) \vee \exp(-\tau/4L) \right\} \le \delta$ if and only if

$$\tau \ge \max\left\{ 2\sqrt{v \log\left(\frac{n_1+n_2}{\delta} \right)},\ 4L \log\left(\frac{n_1+n_2}{\delta} \right) \right\}.$$

■

■ 3.7.3 Proof of Theorem 3.4.5

Proof. Recall the definition of the tangent space $T(\hat{M})$ from (2.3). Note that $T(\hat{M})$ is a vector subspace of $\mathbb{R}^{n_1\times n_2}$ and the orthogonal projection onto $T(\hat{M})$, namely, $\Pi_{T(\hat{M})} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$, can be written as

$$\Pi_{T(\hat{M})}(X) = P_{\mathrm{csp}(\hat{M})} X + X P_{\mathrm{rsp}(\hat{M})} - P_{\mathrm{csp}(\hat{M})} X P_{\mathrm{rsp}(\hat{M})}. \tag{3.34}$$

Then we observe that

$$\hat{M} = \frac{1}{p} \mathcal{P}_r\big(\mathcal{R}_\Omega(Z) \big) = \frac{1}{p} \Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) \big).$$

Next, we consider the decomposition

$$\hat{M} - M = \big(\hat{M} - \Pi_{T(\hat{M})}(M) \big) + \big(\Pi_{T(\hat{M})}(M) - M \big).$$

We can immediately see that $\hat{M} - \Pi_{T(\hat{M})}(M) \in T(\hat{M})$ and $\Pi_{T(\hat{M})}(M) - M \in T(\hat{M})^\perp$, and

therefore, $\langle \hat{M} - \Pi_{T(\hat{M})}(M), \Pi_{T(\hat{M})}(M) - M \rangle = 0$. Thus, for every $i \in [n_1]$,

$$\|e_i^T(\hat{M} - M)\|_2^2 = \|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)\|_2^2 + \|e_i^T\big(\Pi_{T(\hat{M})}(M) - M\big)\|_2^2. \tag{3.35}$$

Now it suffices to show upper bounds for the two terms on the right hand side of (3.35).

The first term in (3.35) First of all, we observe that

$$\hat{M} - \Pi_{T(\hat{M})}(M) = \frac{1}{p}\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)$$

Recall that the random matrix $\mathcal{R}_\Omega(Z) - pM$ is sub-Gaussian with parameter $v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2$ by Lemma 3.4.1 where $v_{\beta^0}(p)$ is the sub-Gaussian norm of a centered Bernoulli random variable, defined in (2.9). Then we observe that $\operatorname{rrank}\Big(\Pi_{T(\hat{M})}\big(\mathcal{R}_\Omega(Z) - pM\big)\Big) \le \operatorname{rrank}(\hat{M}) + \operatorname{rrank}(M) \le 2r$. Applying Lemma 3.4.2, we can see that

$$\mathbb{E}\Big[\|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)\|_2^2\Big] \le \frac{8r}{p^2} \cdot \big(v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2\big).$$

and that for any $\delta > 0$, the following inequality is true with probability at least $1 - \delta$:

$$\begin{aligned}\|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)\|_2^2 \le \mathbb{E}\big[\|e_i^T\big(\hat{M} - \Pi_{T(\hat{M})}(M)\big)\|_2^2\big] \\ + \frac{16}{p^2}\big(v_{\beta^0}(p) \cdot \|M\|_\infty^2 + \eta^2\big) \cdot \left\{\sqrt{\log\left(\frac{2r}{\delta}\right)} \vee \log\left(\frac{2r}{\delta}\right)\right\}.\end{aligned}$$

The second term in (3.35) By (3.34), it follows that

$$\Pi_{T(\hat{M})}(M) - M = -(P_{\mathrm{csp}(\hat{M})} - I_{n_1})M(P_{\mathrm{rsp}(\hat{M})} - I_{n_2}).$$

Observe that $P_{\mathrm{csp}(\hat{M})^\perp} = I_{n_1} - P_{\mathrm{csp}(\hat{M})}$ and $P_{\mathrm{rsp}(\hat{M})^\perp} = I_{n_2} - P_{\mathrm{rsp}(\hat{M})}$. Thus, for any $i \in [n_1]$,

$$e_i^T\big(\Pi_{T(\hat{M})}(M) - M\big) = -e_i^T \cdot P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp}.$$

and therefore,

$$\begin{aligned}\|e_i^T\big(\Pi_{T(\hat{M})}(M) - M\big)\|_2^2 &\le e_i^T \cdot P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp} \cdot e_i \\ &\le \big|\big|\big|P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp}\big|\big|\big|_2 \cdot \|e_i\|_2^2. \end{aligned} \tag{3.36}$$

Moreover, we can observe that $M = P_{\mathrm{csp}(M)} M P_{\mathrm{rsp}(M)}$, and therefore,

$$P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp}$$

$$= P_{\mathrm{csp}(\hat{M})^\perp} \cdot \big(P_{\mathrm{csp}(M)} M P_{\mathrm{rsp}(M)}\big) \cdot P^2_{\mathrm{rsp}(\hat{M})^\perp} \cdot \big(P_{\mathrm{rsp}(\hat{M})^\perp} P_{\mathrm{rsp}(M)} M P_{\mathrm{csp}(M)}\big) \cdot P_{\mathrm{csp}(\hat{M})^\perp}$$
$$= \big(P_{\mathrm{csp}(\hat{M})^\perp} P_{\mathrm{csp}(M)}\big) \cdot M \cdot \big(P_{\mathrm{rsp}(M)} P_{\mathrm{rsp}(\hat{M})^\perp}\big) \cdot \big(P_{\mathrm{rsp}(\hat{M})^\perp} P_{\mathrm{rsp}(M)}\big) \cdot M \cdot \big(P_{\mathrm{csp}(M)} P_{\mathrm{csp}(\hat{M})^\perp}\big).$$

As a result,

$$\begin{aligned} &\big|\!\big|\!\big| P_{\mathrm{csp}(\hat{M})^\perp} M P_{\mathrm{rsp}(\hat{M})^\perp} \cdot P_{\mathrm{rsp}(\hat{M})^\perp} M^T P_{\mathrm{csp}(\hat{M})^\perp} \big|\!\big|\!\big|_2 \\ &\qquad \le |\!|\!| M |\!|\!|_2^2 \cdot \Big|\!\Big|\!\Big| \sin\Theta\big(\mathrm{csp}(\hat{M}), \mathrm{csp}(M)\big) \Big|\!\Big|\!\Big|_2^2 \cdot \Big|\!\Big|\!\Big| \sin\Theta\big(\mathrm{rsp}(\hat{M}), \mathrm{rsp}(M)\big) \Big|\!\Big|\!\Big|_2^2. \end{aligned} \tag{3.37}$$

Lastly, we note that by Wedin's $\sin\Theta$ theorem (Theorem 2.2.5),

$$\big|\!\big|\!\big| \sin\Theta\big(\mathrm{csp}(\hat{M}), \mathrm{csp}(M)\big) \big|\!\big|\!\big|_2 \vee \big|\!\big|\!\big| \sin\Theta\big(\mathrm{rsp}(\hat{M}), \mathrm{rsp}(M)\big) \big|\!\big|\!\big|_2 \le \frac{\big|\!\big|\!\big| \hat{M} - M \big|\!\big|\!\big|_2}{s_r(M)}. \tag{3.38}$$

All in all, we get the following inequality by combining (3.36), (3.37), and (3.38) together:

$$\big\| e_i^T \big(\Pi_{T(\hat{M})}(M) - M\big) \big\|_2^2 \le \frac{s_1(M)^2}{s_r(M)^4} \big|\!\big|\!\big| \hat{M} - M \big|\!\big|\!\big|_2^4.$$

To conclude the proof, it suffices to apply Theorem 3.4.3 to control $\big|\!\big|\!\big| \hat{M} - M \big|\!\big|\!\big|_2$. ■

Chapter 4

Errors-in-variables Regression for Prediction

4.1 Introduction

Regression is a statistical process for estimating the relationship between a dependent variable (often called the response, or the label) and a set of independent variables (often called 'covariates', 'predictors', or 'features'). Regression analysis is primarily used for two conceptually distinct purposes: prediction and causal explanation. For the predictive use, the predicted value of the dependent variable given a configuration of covariates is valuable, whereas for the explanatory use, it is the relationship between the variables that is of interest.

In standard formulations of regression, it is assumed that covariates are fully observed without noise. However, these assumptions may not be realistic for many applications, in which covariates may be only partially observed with corruption. Surveys often suffer from missing data as respondents might refuse or fail to respond to some questions. Sensor network data [133] and microarray gene expression data [93] also tend to be subject to missing data and measurement error. The regression models that account for measurement errors in the covariates are referred to as errors-in-variables regression [35].

In this chapter, we tackle the errors-in-variables regression for prediction purpose with the idea of imputing the data using low-rank matrix completion methods.

4.1.1 Problem Statement

Suppose we observe a response variable $y_i \in \mathbb{R}$ linked to a covariate vector $x_i \in \mathbb{R}^d$ via the linear model $y_i = \langle x_i, \beta^* \rangle + \epsilon_i$ where $\beta^* \in \mathbb{R}^d$ is the unknown regression vector, and $\epsilon_i \in \mathbb{R}$ denotes observation noise. This model can be compactly written as

$$y = X\beta^* + \epsilon, \tag{4.1}$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$. We want to estimate a function $f : x \mapsto y$ from data such that $f(x)$ is close to $\langle x, \beta^* \rangle$.

In many practical scenarios, we may not have direct access to $x_i \in \mathbb{R}^d$; rather, we only

observe a random vector $z_i \in \overline{\mathbb{R}}^d$ (recall that $\overline{\mathbb{R}} := \mathbb{R} \cup \{*\}$ where $*$ denotes 'unknown') linked to x_i via some conditional distribution, i.e.,

$$z_i \sim \mathbb{Q}(\ \cdot\ |x_i) \quad \text{for } i = 1, 2, \ldots, n. \tag{4.2}$$

As a result, we have access to data $\{(z_1, y_1), \ldots, (z_n, y_n)\}$ and want to use these data to estimate a regression function f such that $f(z) \approx \langle x, \beta^* \rangle$.

Specifically, we define two types of prediction objectives as follows.

In-sample Prediction First of all, we want the regression function to predict the response accurately for the samples used in training, i.e., $f(z_i) \approx \langle x_i, \beta^* \rangle$ for all $i \in [n]$. Thus, we consider the fixed design setting, where the covariates $x_1, \ldots, x_n$ are fixed (deterministic), and measure the prediction performance of f using the mean squared error (MSE), which is defined by

$$\mathrm{MSE}(f) = \frac{1}{n} \sum_{i=1}^{n} \big(f(z_i) - \langle x_i, \beta^* \rangle\big)^2.$$

Note that $\mathrm{MSE}(f)$ only addresses the prediction accuracy of f at the n sample points that are already observed.

Out-of-sample Prediction Oftentimes, the main goal of prediction is to use the learned function f to predict the response y^{new} given a new sample x^{new}, which is observed through $z^{\mathrm{new}} \sim \mathbb{Q}(\cdot|x^{\mathrm{new}})$ in our model. Therefore, we are interested in the prediction error for the response at an unseen feature, i.e., $f(z^{\mathrm{new}}) - \langle x^{\mathrm{new}}, \beta^* \rangle$. We define the squared forecasting error at z^{new} as follows:

$$\mathrm{SFE}(f; z^{\mathrm{new}}) = \big(f(z^{\mathrm{new}}) - \langle x^{\mathrm{new}}, \beta^* \rangle\big)^2.$$

■ 4.1.2 Contributions and Organization of the Chapter

The main contribution of this chapter is the analysis of the regression using the data imputed by low-rank matrix completion. We argue that if the main objective of regression analysis is in prediction rather than model identification, then data imputation is not harmful. As a byproduct of our analysis, we observe that the well-established technique of principal component regression (PCR) algorithm can be interpreted as an instance of 'regression-after-imputation' algorithm (Algorithm 2) considered in this chapter. This observation enables extending our analysis to yield finite-sample analysis for PCR. Overall, we provide a new perspective on errors-in-variables regression centered at prediction rather than model identification, with provable theoretical guarantees on prediction error and covariate estimation error as summarized in Table 4.1.

Table 4.1: Summary of contributions in this chapter in comparison with likelihood-based works in the errors-in-variables regression literature (Section 4.2.1). Here, n denotes the number of samples, and d, r, p denote the dimension, the rank, and the fraction of observed entries in the covariates, respectively.

	Objective	Assumptions	Knowledge of noise distribution	Prediction error	Covariate estimation error
Likelihood-based (Section 4.2.1)	Model identification	sparsity of β^* restricted eigenvalue cond.	Required	N/A	N/A
This thesis	**Prediction**	**low-rank covariates with balanced spectrum**	**Not required**	$\frac{1}{p}\frac{r}{n \wedge d}$ (Corollary 4.5.1)	$\lesssim \frac{r}{p^2}\frac{n \vee d}{n}$ (Corollary 4.5.3)

The rest of this chapter is organized as follows. In Section 4.2, we briefly survey the literature on errors-in-variables regression and principal component regression. In Section 4.3, we describe the algorithm for linear regression with the data imputed through low-rank matrix completion. Specifically, we describe the training phase (summarized in Algorithm 2) as well as the prediction phase where we use the learned regression function to predict the response at a new, noisy covariate vector. In Section 4.4, we present an upper bound on the in-sample prediction error, and discuss an informal upper bound on the out-of-sample prediction error. In Section 4.5, we consider a specific instance of Algorithm 2 such that the simple SVT algorithm (Algorithm 1, Section 3.3) is used as the matrix completion oracle. Specifically, we discuss its equivalence to PCR algorithm, prediction error bound, and covariate estimation error bound. Section 4.6 contains some numerical experiments that support our claims in the chapter. Lastly, we conclude the chapter with a summary in Section 4.7.

4.2 Related Work

4.2.1 Errors-in-variables Regression

There is a large body of work regarding errors-in-variables regression problem, e.g., [35, 120, 96, 18, 42], as well as references therein. However, much of the earlier work is either about asymptotic theory in classical setting (where the sample size n diverges with the dimension d fixed) or concerned with identifying a high-dimensional sparse linear model, assuming the existence of a ground truth model. As a result, their efforts are focused on proposing computationally tractable algorithms that correctly estimate the underlying parameter, under certain identifiability conditions such as sparsity of β^* and restricted eigenvalue condition on the covariance of the design matrix [96, 42].

Perhaps one of the first methods trained statisticians would consider to deal with missing data is a likelihood-based method involving the expectation-maximization (EM) algorithm

[95]. Städler and Bühlmann [137] developed an EM-based ℓ_1-regularized likelihood method for sparse inverse covariance estimation of the multivariate normal model in the presence of missing data, and used the result to derive an algorithm for sparse linear regression with missing data. However, it is difficult to guarantee the EM method will converge to a global optimum because the negative likelihood often becomes nonconvex with missing or noisy data.

Loh and Wainwright [96] consider an ℓ_1-regularized quadratic program, obtained by replacing the empirical covariance $\frac{1}{n}X^TX$ and the response vector $\frac{1}{n}X^Ty$ in the least squares formulation with the quantities that these quantities would be if there were no noise or missing values in the covariates. They show guarantees for the convergence (using projected gradient descents) and statistical error of the optimal solution under the restricted eigenvalue condition and additional mild technical conditions. Nevertheless, their approach requires explicit knowledge of the distribution of covariate noise, i.e., $\mathbb{Q}$ in (4.2), to compensate for the noise contribution in the covariance and the response. In fact, many other likelihood-based estimators in the literature require such a priori knowledge about the covariate noise [120, 42, 18].

Unlike these works, we aim at revealing a model that is valid for prediction use, rather than identifying the true model. Thus, our approach does not require prior knowledge about the covariate noise distribution. In addition, it is worth noting that it is not clear how the methods from the aforementioned works can be used to predict the response associated with unseen, noisy covariates because they only focus on estimating β^* and do not pay much attention to de-noising covariates.

■ 4.2.2 Principal Component Regression

Principal component regression (PCR) is a general term that refers to statistical techniques that use principal component analysis (PCA) in regression analysis [75]. The best known approach in this line simply performs PCA on the covariate data matrix and then regresses the response vector on a selected subset of principal components with the largest variance, which is commonly referred to as the principal component regression. There are other variants depending on the use of PCA, and one notable work by Bair et al. [14] proposes to use 'supervised' principal components, namely, principal components of a small subset of covariate variables that are most correlated to the response variable.

The effectiveness of PCR such as dimensionality reduction and shrinkage effect are well understood. However, the formal literature providing a rigorous analysis of PCR is sparse, despite its widespread use in practice. Especially, the ability of PCR to handle covariate data with noisy, missing entries is less understood, and its analysis remain as an open challenge. Although various methodological proposals have been made to use PCA with missing data [67, 141, 60], these works fall short of providing meaningful theoretical guarantees.

■ 4.3 Regression with Imputed Data

In this section, we describe how we estimate a prediction function $f : z \mapsto y$ using the imputed covariate data. In the training phase, we process corrupted covariate data Z with a low-rank matrix completion algorithm, and then apply standard regression analysis. In the prediction phase, we use the imputed covariates and the estimated regression function to make predictions, either for existing covariates in the training set or for a new covariate vector out of the samples. These two phases are described below.

Training Phase Given a partially observed, noisy set of covariates $z_1, \dots, z_n$, we estimate the true, de-noised covariates $x_1, \dots, x_n$ using a low-rank matrix completion algorithm. To be specific, we let $Z \in \overline{\mathbb{R}}^{n \times d}$ be the matrix whose i-th row is z_i, and estimate the underlying low-rank covariate matrix $\hat{X} \in \mathbb{R}^{n \times d}$ with any matrix completion algorithm of the user's choice. Then with the imputed covariate data $\hat{X}$, we apply standard regression analysis to estimate the regression vector β^*. For the sake of concreteness, here we estimate β^* by choosing the minimum ℓ_2-norm least squares solution, i.e.,

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \left\{ \|\beta\|_2 : \beta \text{ minimizes } \|y - \hat{X}\beta\|_2^2 \right\},$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ is the response data. This estimator can be equivalently written as $\hat{\beta} = \left(\hat{X}^T \hat{X}\right)^\dagger \hat{X}^T y$. This estimator is also known under the name 'ridgeless' regression estimator in the literature, motivated by the fact that $\hat{\beta} = \lim_{\lambda \to 0} \hat{\beta}_\lambda$ where $\hat{\beta}_\lambda$ denotes the ridge regression estimator $\hat{\beta}_\lambda = \arg\min_\beta \left\{ \frac{1}{n} \|y - \hat{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\} = \left(\hat{X}^T \hat{X} + \lambda I_d\right)^\dagger \hat{X}^T y$.

The procedure described above is summarized in Algorithm 2.

Algorithm 2: Linear regression with covariates estimated by matrix completion

Input: $Z \in \overline{\mathbb{R}}^{n \times d}$ and $y \in \mathbb{R}^n$
Output: $\hat{X} \in \mathbb{R}^{n \times d}$ and $\hat{\beta} \in \mathbb{R}^d$

1. Estimate the design matrix (covariates) X as $\hat{X} \leftarrow \varphi_{\mathrm{MC}}(Z)$ where φ_{MC} is any low-rank matrix completion oracle.
2. Estimate the regression vector β^* by solving a least squares problem as $\hat{\beta} \leftarrow \arg\min_{\beta \in \mathbb{R}^d} \left\{ \|\beta\|_2 : \beta \text{ minimizes } \|y - \hat{X}\beta\|_2^2 \right\}$.

Here we remark on two aspects of Algorithm 2. First of all, we have freedom to choose any low-rank matrix completion oracle in Step 1. For example, we can employ the simple SVT algorithm (Algorithm 1) studied in Chapter 3, or any optimization-based methods that solve (3.3) or (3.4). Second, we choose the minimum ℓ_2-norm solution among the minimizers of the

least squares problem in Step 2 because there may exist multiple minimizers when rank $\hat{X} < d$. We also remark that the minimum ℓ_2-norm solution satisfies $\hat{\beta} \in \mathrm{rsp}(\hat{X})$. It is because if $\hat{\beta}_1, \hat{\beta}_2 \in \arg\min \|y - \hat{X}\beta\|_2^2$, then $\hat{X}\hat{\beta}_1 = \hat{X}\hat{\beta}_2$, and therefore, $\hat{\beta}_1 - \hat{\beta}_2 \in \ker(\hat{X}) = \mathrm{rsp}(\hat{X})^\perp$.

Prediction Phase With the $\hat{X}$ and $\hat{\beta}$ identified in the training phase, we define a prediction function. For precise description, we introduce some notations. Let $x^{\mathrm{new}} \in \mathbb{R}^d$ denote a new covariate vector, and $z^{\mathrm{new}} \in \overline{\mathbb{R}}^d$ be an observed instance of x^{new}, cf. (4.2). Let $\omega = \{j \in [d] : z_j^{\mathrm{new}} \neq *\}$ denote the set of indices $j \in [d]$ such that the j-th coordinate of z^{new} is not missing. Then the prediction is made in two steps:

1. Estimate x^{new} from z^{new} based on the side information $\hat{X}$. Specifically, we let

$$\hat{x}^{\mathrm{new}} = \arg\min_{x \in \mathrm{rsp}(\hat{X})} \left\{ \|x\|_2 : x \text{ minimizes } \sum_{j \in \omega} (x_j - z_j^{\mathrm{new}})^2 \right\}. \tag{4.3}$$

2. Predict y^{new} to be $f(z^{\mathrm{new}}) = \langle \hat{x}^{\mathrm{new}}, \hat{\beta} \rangle$.

For the data in the training set, we do not need to estimate the covariates again, and can just let $f(z_i) = \langle \hat{x}_i, \hat{\beta} \rangle = (\hat{X}\hat{\beta})_i$ for all $i \in [n]$.

4.4 Prediction Error Analysis

4.4.1 Model Assumptions

We make a few assumptions on our data generation model described in (4.1) and (4.2). The following assumptions are arguably the simplest and natural assumptions to study theoretical behaviors of linear regression with missing and corrupted covariates.

Assumption 4.1. The noise ϵ in the model (4.1) is sub-Gaussian with parameter σ^2.

Assumption 4.2. The disturbance to the covariates in (4.2) is a combination of additive noise and missing data as follows. When $z \sim \mathbb{Q}(\cdot|x)$, $z_j = x_j + w_j$ with probability p and $z_j = *$ with probability $1 - p$ independently for each $j \in [d]$, where $p \in (0, 1]$, and $w \in \mathbb{R}^d$ be a η^2-sub-Gaussian random vector.

Observe that Assumption 4.2 resembles Assumptions 3.1 and 3.2 that are used for our analysis of the SVT algorithm in Chapter 3.

4.4.2 In-sample Prediction Error

In this section, we discuss an upper bound on the in-sample prediction error of linear regression based on imputed design matrix, which is stated in the next theorem.

Theorem 4.4.1. *Let $\hat{X}, \hat{\beta}$ be the outputs of Algorithm 2. Then the MSE of the resulting predictor $\hat{X}\hat{\beta}$ satisfies*

$$n \cdot \mathrm{MSE}(\hat{X}\hat{\beta}) \le \left\| \sin \Theta\big(\mathrm{csp}(\hat{X}), \mathrm{csp}(X)\big) \right\|_2^2 \left\| X\beta^* \right\|_2^2 + \left\| \Pi_{\mathrm{csp}(\hat{X})}(\epsilon) \right\|_2^2. \tag{4.4}$$

Moreover, if Assumption 4.1 holds, then for any $\delta > 0$, the following inequality is true with probability at least $1 - \delta$:

$$\left\| \Pi_{\mathrm{csp}(\hat{X})}(\epsilon) \right\|_2^2 \le \mathop{\mathbb{E}}_{\epsilon} \left[\left\| \Pi_{\mathrm{csp}(\hat{X})}(\epsilon) \right\|_2^2 \right] + 16\sigma^2 \Big(\log\left(\hat{r}/\delta\right) \vee \sqrt{\log\left(\hat{r}/\delta\right)} \Big)$$

where $\hat{r} = \mathrm{rank}(\hat{X})$ and

$$\mathop{\mathbb{E}}_{\epsilon} \left[\left\| \Pi_{\mathrm{csp}(\hat{X})}(\epsilon) \right\|_2^2 \right] \le 4\sigma^2 \hat{r}.$$

Before presenting the proof of Theorem 4.4.1, we make a few comments on the upper bound (4.4) in Theorem 4.4.1. This upper bound consists of two terms: the first term is due to the mis-specification of the column space of X; and the other term is subject to the measurement noise in the response vector. Note that the second term is inevitable even when we have direct access to X, and it actually matches the minimax optimal lower bound for linear regression up to a constant factor. Thus, the first term provides an upper bound for 'the price to pay' incurred by the mis-specification of the column space of X, or in other words, the excess risk due to imputing the missing values.

Now, the remaining question is "how large (or small) the first term in the upper bound will be?" This boils down to the question of finding an upper bound for the sine of the principal angle $\Theta\big(\mathrm{csp}(\hat{X}), \mathrm{csp}(X)\big)$. Suppose that we have an upper bound for the error in the design matrix estimation, $\hat{X} - X$, which is measured in an orthogonally invariant norm. Thanks to the celebrated $\sin\Theta$ theorem due to Davis and Kahan [43, 160], we can derive an upper bound on the sine of $\Theta\big(\mathrm{csp}(\hat{X}), \mathrm{csp}(X)\big)$ from the estimation error bound for $\hat{X} - X$, measured in the same norm, cf. Theorem 2.2.5. Therefore, if the matrix completion oracle in use estimates $\hat{X}$ accurately so that $\left\| \hat{X} - X \right\|_2$ is small, then we are golden. Note that the argument so far is valid for any matrix completion oracle as long as it comes with a spectral norm error guarantee. In Section 4.5.2, we consider a concrete example of the singular value thresholding algorithm for matrix completion, and derive an asymptotic upper bound $\mathrm{MSE}(\hat{X}\hat{\beta}) \lesssim r/n$.

Proof of Theorem 4.4.1. Recall that $\hat{\beta}$ is a solution of the least squares problem $\min_{\beta \in \mathbb{R}^d} \|y - \hat{X}\beta\|_2^2$, and thus, $\hat{X}\hat{\beta} = \Pi_{\mathrm{csp}(\hat{X})}(y) = \Pi_{\mathrm{csp}(\hat{X})}(X\beta^*) + \Pi_{\mathrm{csp}(\hat{X})}(\epsilon)$. By the Pythagorean theorem,

$$n \cdot \mathrm{MSE}(\hat{X}\hat{\beta}) = \left\| \hat{X}\hat{\beta} - X\beta^* \right\|_2^2 = \left\| \Pi_{\mathrm{csp}(\hat{X})}(X\beta^*) - X\beta^* \right\|_2^2 + \left\| \Pi_{\mathrm{csp}(\hat{X})}(\epsilon) \right\|_2^2. \tag{4.5}$$

Now we observe that $\Pi_{\mathrm{csp}(\hat{X})^\perp} = I_{\mathbb{R}^n} - \Pi_{\mathrm{csp}(\hat{X})}$ and that $X\beta^* = \Pi_{\mathrm{csp}(X)}(X\beta^*)$. Then it follows

from the definition of the operator norm and the characterization of the sine of principal angles in (2.1) that

$$\begin{aligned}\left\|\Pi_{\mathrm{csp}(\hat{X})}(X\beta^*) - X\beta^*\right\|_2 &= \left\|\Pi_{\mathrm{csp}(\hat{X})^\perp}(X\beta^*)\right\|_2 = \left\|\Pi_{\mathrm{csp}(\hat{X})^\perp}\Pi_{\mathrm{csp}(X)}(X\beta^*)\right\|_2 \\ &\le \left|\!\left|\!\left|P_{\mathrm{csp}(\hat{X})^\perp}P_{\mathrm{csp}(X)}\right|\!\right|\!\right|_2 \left\|X\beta^*\right\|_2 \\ &= \left|\!\left|\!\left|\sin\Theta\big(\mathrm{csp}(\hat{X}),\mathrm{csp}(X)\big)\right|\!\right|\!\right|_2 \left\|X\beta^*\right\|_2.\end{aligned} \tag{4.6}$$

Combining (4.5) and (4.6) proves the first statement.

Now it remains to prove the second statement. Recall that ϵ is σ^2-sub-Gaussian due to Assumption 4.1. We conclude the proof by applying Lemma 3.4.2 with $\mathcal{V} = \mathrm{csp}(\hat{X})$. ■

■ 4.4.3 Out-of-sample Prediction Error

In this section, we briefly discuss an informal upper bound on the out-of-sample prediction error, $\langle \hat{x}^{\mathrm{new}}, \hat{\beta}\rangle - \langle x^{\mathrm{new}}, \beta^*\rangle$. Although we do not present a formal theorem, the discussion in this section will give a clue why the out-of-sample prediction error is likely to be small when $\hat{X}$ well approximates X.

To begin with, we let $\mathcal{V} = \mathrm{rsp}(X), \hat{\mathcal{V}} = \mathrm{rsp}(\hat{X})$, and observe that $\hat{x}^{\mathrm{new}} \in \hat{\mathcal{V}}$ due to its construction in (4.3). Then, we get the following upper bound:

$$\begin{aligned}\left|\langle \hat{x}^{\mathrm{new}}, \hat{\beta}\rangle - \langle x^{\mathrm{new}}, \beta^*\rangle\right| &\le \left|\langle \hat{x}^{\mathrm{new}}, \hat{\beta} - \beta^*\rangle\right| + \left|\langle \hat{x}^{\mathrm{new}} - x^{\mathrm{new}}, \beta^*\rangle\right| \\ &\le \|\hat{x}^{\mathrm{new}}\|_2 \|P_{\hat{\mathcal{V}}}(\hat{\beta} - \beta^*)\|_2 + \|\hat{x}^{\mathrm{new}} - x^{\mathrm{new}}\|_2 \|\beta^*\|_2.\end{aligned} \tag{4.7}$$

In the rest of this section, we argue why $\|P_{\hat{\mathcal{V}}}(\hat{\beta} - \beta^*)\|_2$ and $\|\hat{x}^{\mathrm{new}} - x^{\mathrm{new}}\|_2$ are likely to be 'small.'

Error in Estimating β^* Recall that $\hat{\beta} = \hat{X}^\dagger y = \hat{X}^\dagger X\beta^* + \hat{X}^\dagger\epsilon$. Thus, it follows that

$$\begin{aligned}\left\|P_{\hat{\mathcal{V}}}(\hat{\beta} - \beta^*)\right\|_2 &\le \left\|P_{\hat{\mathcal{V}}}(\hat{X}^\dagger - X^\dagger)X\beta^*\right\|_2 + \left\|P_{\hat{\mathcal{V}}}\hat{X}^\dagger\epsilon\right\|_2 \\ &\le \left\|\hat{X}^\dagger(\hat{X} - X)\beta^*\right\|_2 + \left\|\hat{X}^\dagger\epsilon\right\|_2 \qquad \text{cf. [138, Theorem 3.2]} \\ &\le \frac{\left|\!\left|\!\left|\hat{X} - X\right|\!\right|\!\right|_2}{s_r(\hat{X})}\|\beta^*\|_2 + \frac{1}{s_r(\hat{X})}\|\epsilon\|_2.\end{aligned}$$

As discussed in Chapter 3, $\left|\!\left|\!\left|\hat{X} - X\right|\!\right|\!\right|_2 \lesssim \sqrt{\frac{n\vee d}{p}}$, whereas $s_r(\hat{X}) \asymp \sqrt{\frac{nd}{r}}$ provided that $\frac{s_1(\hat{X})}{s_r(\hat{X})} \asymp 1$. Then, we get $\left\|P_{\hat{\mathcal{V}}}(\hat{\beta} - \beta^*)\right\|_2 \lesssim \sqrt{\frac{r}{(n\wedge d)p}}\|\beta^*\|_2$.

Error in Estimating x^{new} If $p = 1$, then $\hat{x}^{\mathrm{new}} = \mathcal{P}_{\hat{\mathcal{V}}}(z^{\mathrm{new}})$. Observe that $z^{\mathrm{new}} - x^{\mathrm{new}}$ is a η^2-sub-Gaussian random vector and $x^{\mathrm{new}} \in \mathcal{V}$. Thus, we can see (e.g., by using Lemma 3.4.2

and Lemma 2.2.3) that

$$\begin{aligned}\|\hat{x}^{\text{new}} - x^{\text{new}}\|_2^2 &= \|\hat{x}^{\text{new}} - \mathcal{P}_{\hat{\mathcal{V}}}(x^{\text{new}})\|_2^2 + \|\mathcal{P}_{\hat{\mathcal{V}}}(x^{\text{new}}) - x^{\text{new}}\|_2^2 \\ &= \|\mathcal{P}_{\hat{\mathcal{V}}}(z^{\text{new}} - x^{\text{new}})\|_2^2 + \|\mathcal{P}_{\hat{\mathcal{V}}^\perp}(x^{\text{new}})\|_2^2 \\ &\lesssim \dim(\hat{\mathcal{V}})\eta^2 + \|x^{\text{new}}\|_2^2 \big\|\big| \sin^2\Theta(\hat{\mathcal{V}}, \mathcal{V})\big|\big\|_2^2 \\ &\le \eta^2 r + \|x^{\text{new}}\|_2^2 \frac{\|\|\hat{X} - X\|\|_2^2}{s_r(X)^2}.\end{aligned}$$

Therefore, $\|\hat{x}^{\text{new}} - x^{\text{new}}\|_2^2 \lesssim \frac{r}{p(n\wedge d)}\|x^{\text{new}}\|_2^2$, provided that $s_r(\hat{X}) \asymp \sqrt{\frac{nd}{r}}$. We believe it is possible to derive a similar conclusion when $p < 1$, as long as $p \gtrsim \frac{r}{d}$ (up to log factors) and $\mathcal{V}$ is incoherent with respect to the standard basis vectors.

Conclusion When $p \gtrsim \frac{r}{n\wedge d}$, we have $\|\hat{x}^{\text{new}} - x^{\text{new}}\|_2 \lesssim \|x^{\text{new}}\|_2$, and it follows from (4.7) that

$$\big|\langle \hat{x}^{\text{new}}, \hat{\beta}\rangle - \langle x^{\text{new}}, \beta^*\rangle\big| \lesssim \sqrt{\frac{r}{(n\wedge d)p}}\|x^{\text{new}}\|_2\|\beta^*\|_2.$$

Therefore, we expect the following upper bound on the normalized squared forecasting error at a new sample:

$$\frac{\text{SFE}(\hat{f}; z^{\text{new}})}{\|x^{\text{new}}\|_2^2\|\beta^*\|_2^2} \lesssim \frac{r}{(n\wedge d)p}$$

where $\hat{f} : z^{\text{new}} \mapsto \langle \hat{x}^{\text{new}}, \hat{\beta}\rangle$.

■ 4.5 Principal Component Regression

In this section, we study a specific example of matrix completion oracle, namely, the simple SVT algorithm discussed in Section 3.3. In Section 4.5.1, we observe that Algorithm 2 equipped with the simple SVT algorithm as the matrix completion oracle is equivalent to the traditional principal component regression algorithm. Then, in Section 4.5.2, we establish an upper bound for its prediction error as a corollary of Theorem 4.4.1, and the spectral norm error upper bound for the simple SVT (Theorem 3.4.3). Lastly, in Section 4.5.3, we use the $\ell_{2,\infty}$-norm bound for the SVT algorithm (Theorem 3.4.5) to argue that each of the in-sample covariate vectors can be reliably estimated in the ℓ_2 sense.

■ 4.5.1 Equivalence of PCR and Regression-after-SVT

Suppose that $p = 1$, i.e., every entry of Z is observed without any missing values. Then Step 1 of Algorithm 2 returns $\hat{X} = Z_r$, the best rank-r approximation of the noisy design matrix Z. In Step 2 of Algorithm 2, we obtain the ridgeless regression estimator $\hat{\beta} = (\hat{X}^T\hat{X})^\dagger\hat{X}^T y$.

This is equivalent to applying principal component regression with retaining only top-r principal components. Let $Z_r = U_r \Sigma_r V_r^T$ be a compact SVD of Z_r. In the literature, principal component regression is usually described as a three-step process as follows.

- Let $\Phi = ZV_r \in \mathbb{R}^{n \times r}$ denote the new design matrix, whose i-th row vector $\phi_i = V_r^T z_i \in \mathbb{R}^r$ represents the new i-th covariate vector in the transformed space.
- Let $\gamma = (\Phi^T \Phi)^\dagger \Phi^T y \in \mathbb{R}^r$ denote the regression vector in the transformed space.
- Assign $\beta = V_r \gamma$.

Note that this is exactly the same process to the steps in Algorithm 2 using the simple SVT algorithm, just described in the space of the transformed covariates.

Consequently, our Algorithm 2 can be viewed as a generalization of PCR, and the analysis in this chapter can be naturally carried over to the analysis of PCR.

■ 4.5.2 Prediction Error for PCR: Corollary of Theorem 4.4.1

Combining the prediction error upper bound from Theorem 4.4.1 and the spectral norm upper bound for the SVT algorithm (Theorem 3.4.3) yields the following corollary.

Corollary 4.5.1. *Let $\hat{X}, \hat{\beta}$ be the outputs of Algorithm 2 with the matrix completion oracle in use being the simple SVT algorithm (Algorithm 1) discussed in Section 3.3. If Assumptions 4.1 and 4.2 hold, then for any $\delta_1, \delta_2 > 0$, the following inequality is true for the resulting predictor $\hat{X}\hat{\beta}$ with probability at least $1 - \delta_1 - \delta_2$:*

$$\begin{aligned} n \cdot \mathrm{MSE}(\hat{X}\hat{\beta}) \le \frac{16\|X\beta^*\|_2^2}{p^2 \cdot s_r(X)^2} \cdot \left\{ \sqrt{v_* \log\left(\frac{n+d}{\delta_1}\right)} \vee 2L_* \log\left(\frac{n+d}{\delta_1}\right) \right\}^2 \\ + 4\sigma^2 r + 16\sigma^2 \left\{ \log\left(\frac{r}{\delta_2}\right) \vee \sqrt{\log\left(\frac{r}{\delta_2}\right)} \right\} \end{aligned} \tag{4.8}$$

where $r = \mathrm{rank}(X) = \mathrm{rank}(\hat{X})$ and

$$v_* = 6\sqrt{2\pi} \cdot p \cdot (n \vee d) \cdot \left[(1-p)\|X\|_\infty \cdot \left\{ \left[1 - \left(1 - \frac{e^2}{8\sqrt{2\pi}} \right) p \right] \|X\|_\infty + 2\eta \right\} + \eta^2 \right],$$

$$L_* = (1-p)\|X\|_\infty + \eta + \frac{e^2}{8\sqrt{2\pi}} \frac{p(1-p)\|X\|_\infty^2}{(1-p)\|X\|_\infty + \eta}.$$

Interpretation of Corollary 4.5.1 Now, we parse the upper bound in Corollary 4.5.1, as we did in Section 3.4.2. First of all, we observe that $v_* \asymp p \cdot (n \vee d) \cdot \{\|X\|_\infty \vee \eta\}^2$ and

$L_* \asymp \|X\|_\infty \vee \eta$. Therefore, we can see that

$$\sqrt{v_* \log\Big(\frac{n+d}{\delta_1}\Big)} \gtrsim L_* \log\Big(\frac{n+d}{\delta_1}\Big) \quad \text{if and only if} \quad p \gtrsim \frac{1}{n \vee d} \log\Big(\frac{n+d}{\delta_1}\Big).$$

Thus, when $p \gtrsim \frac{1}{n\vee d} \log\Big(\frac{n+d}{\delta_1}\Big)$ we get the following simplified inequality

$$n \cdot \mathrm{MSE}(\hat{X}\hat{\beta}) \lesssim \frac{\|X\beta^*\|_2^2}{s_r(X)^2} \cdot \{\|X\|_\infty \vee \eta\}^2 \cdot \frac{n \vee d}{p} \log\Big(\frac{n+d}{\delta_1}\Big) + \sigma^2 \Big\{ r + \log\Big(\frac{r}{\delta_2}\Big)\Big\}.$$

Further, if $s_1(X) \asymp s_r(X)$ and $X_{ij} \asymp \|X\|_\infty$ for most $(i,j) \in [n] \times [d]$, then $s_r(X) \asymp \|X\|_\infty \sqrt{nd/r}$. Eventually, we obtain the following high-probability upper bound for the relative MSE (up to log factor):

$$\frac{n \cdot \mathrm{MSE}(\hat{X}\hat{\beta})}{\|X\beta^*\|_2^2} = \frac{\|\hat{X}\hat{\beta} - X\beta^*\|_2^2}{\|X\beta^*\|_2^2} \lesssim \frac{1}{p}\Big\{1 \vee \frac{\eta}{\|X\|_\infty}\Big\}^2 \cdot \frac{r}{n \wedge d} + \frac{\sigma^2 r}{\|X\beta^*\|_2^2}. \tag{4.9}$$

Remark 4.5.2. Observe that when $\|X\beta^*\|_2^2 \asymp n$, the error rate of the relative prediction error upper bound in (4.9) is $O\big(\frac{1}{p}\frac{r}{n\wedge d}\big)$. Actually, the dependence on $n \wedge d$ is an artifact of the use of a suboptimal perturbation result, i.e., Wedin's version of $\sin\Theta$ theorem (Theorem 2.2.5) that yields the same upper bound for the row space and column space. If we use a more refined perturbation result, e.g., [30, Theorem 3], then it is possible to get the correct rate of $O\big(\frac{1}{p}\frac{r}{n}\big)$.

■ 4.5.3 Covariate Estimation Error: Corollary of Theorem 3.4.5

Sometimes the covariate vectors themselves can be the target of estimation. We obtain an upper bound on $\big|\big|\big|\hat{X} - X\big|\big|\big|_{2,\infty}^2 = \max_{i\in[n]} \|\hat{x}_i - x_i\|_2^2$ as a corollary of Theorem 3.4.5.

Corollary 4.5.3. *Let $\hat{X}$ be the estimate of X that is produced by applying the simple SVT algorithm (Algorithm 1) to Z. If Assumption 4.2 holds, then for any $\delta_1, \delta_2 > 0$, the following inequality is true with probability at least $1 - \delta_1 - \delta_2$:*

$$\begin{aligned}\big|\big|\big|\hat{X} - X\big|\big|\big|_{2,\infty}^2 \le{}& \frac{8r}{p^2} \cdot \big(v_{\beta^0}(p) \cdot \|X\|_\infty^2 + \eta^2\big) \\ &+ \frac{16}{p^2}\big(v_{\beta^0}(p) \cdot \|X\|_\infty^2 + \eta^2\big) \cdot \Big\{ \sqrt{\log\Big(\frac{2r}{\delta_1}\Big)} \vee \log\Big(\frac{2r}{\delta_1}\Big)\Big\} \\ &+ \frac{s_1(X)^2}{s_r(X)^4}\frac{256}{p^4} \cdot \Big\{ \sqrt{v_* \log\Big(\frac{n_1+n_2}{\delta_2}\Big)} \ \vee \ 2L_* \log\Big(\frac{n_1+n_2}{\delta_2}\Big)\Big\}^4\end{aligned}$$

where $v_{\beta^0}(\beta)$ denotes the sub-Gaussian norm of the centered Bernoulli random variable with parameter p defined in (2.9), and v_*, L_* have the same expressions as in Corollary 4.5.1.

Interpretation of Corollary 4.5.3 Again, we observe that $v_* \asymp p \cdot (n \vee d) \cdot \{\|X\|_\infty \vee \eta\}^2$ and $L_* \asymp \|X\|_\infty \vee \eta$, and therefore,

$$\sqrt{v_* \log\Big(\frac{n+d}{\delta_1}\Big)} \gtrsim L_* \log\Big(\frac{n+d}{\delta_1}\Big) \quad \text{if and only if} \quad p \gtrsim \frac{1}{n \vee d} \log\Big(\frac{n+d}{\delta_1}\Big).$$

If $s_1(X) \asymp s_r(X)$ and $X_{ij} \asymp \|X\|_\infty$ for most $(i,j) \in [n] \times [d]$, then $s_1(X), s_r(X) \asymp \|X\|_\infty \sqrt{nd/r}$. Then we get the following high-probability upper bound (up to log factor):

$$\big|\!\big|\!\big|\hat{X} - X\big|\!\big|\!\big|_{2,\infty}^2 \lesssim \frac{r}{p^2}\big\{\|X\|_\infty \vee \eta\big\}^2\Big(1 + \frac{n \vee d}{n \wedge d}\Big).$$

As already discussed in Remark 4.5.2, the aspect ratio $\frac{n \vee d}{n \wedge d}$ appears as an artifact of our analysis using sub-optimal perturbation bound (Wedin's $\sin\Theta$ theorem), and can be improved. Then, we obtain

$$\big|\!\big|\!\big|\hat{X} - X\big|\!\big|\!\big|_{2,\infty}^2 \lesssim \big\{\|X\|_\infty \vee \eta\big\}^2 \frac{r}{p^2} \frac{n \vee d}{n},$$

which implies that nontrivial recovery of every covariate vector (in the ℓ_2 sense) is possible as long as $p \gtrsim \sqrt{r/d}$ (assuming $n \geq d$).

We conjecture that this upper bound provides a tight analysis for the SVT algorithm, and thus, for the PCR. However, we mention that optimization-based algorithms might be able to achieve an improved rate of $\big|\!\big|\!\big|\hat{X} - X\big|\!\big|\!\big|_{2,\infty}^2 \lesssim \eta^2 \frac{r}{p} \frac{n \vee d}{n}$, as noted earlier in Remark 3.4.7.

■ 4.6 Numerical Experiments

In this section, we confirm our claims with simple numerical simulations.

Model We construct a low-rank design matrix $X = UV^T$ as the product of two random matrices $U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{d \times r}$ such that $U_{ij}, V_{ij} \sim N(0,1)$ are independent and identically distributed Gaussian random variables. We let $W \in \mathbb{R}^{n \times d}$ be a Gaussian random matrix with i.i.d. entries distributed as $N(0, \eta^2)$, and construct Ω by letting $(i,j) \in \Omega$ with probability p independently for all $(i,j) \in [n] \times [d]$. Recall that we only have access to X through Z such that $Z_{ij} = X_{ij} + W_{ij}$ for $(i,j) \in \Omega$, and $Z_{ij} = *$ otherwise. The response vector y is generated as $y = X\beta^* + \epsilon$ where $\beta^* \sim N(0, 10^2 I_d)$ and $\epsilon \sim N(0, \sigma^2 I_d)$.

Experiment 1 In our first experiment, we measure the normalized mean squared error of Algorithm 2, implemented with two types of matrix completion algorithms – simple SVT algorithm (Algorithm 1), and alternating minimization that solves the nonconvex formulation (3.4). Precisely, for each parameter configuration $(n, d, r, p, \eta, \sigma)$, we generate 100 random instances of $(X, W, \Omega, \beta^*, \epsilon)$, and estimate $(\hat{X}, \hat{\beta})$. Then we measure the normalized MSE,

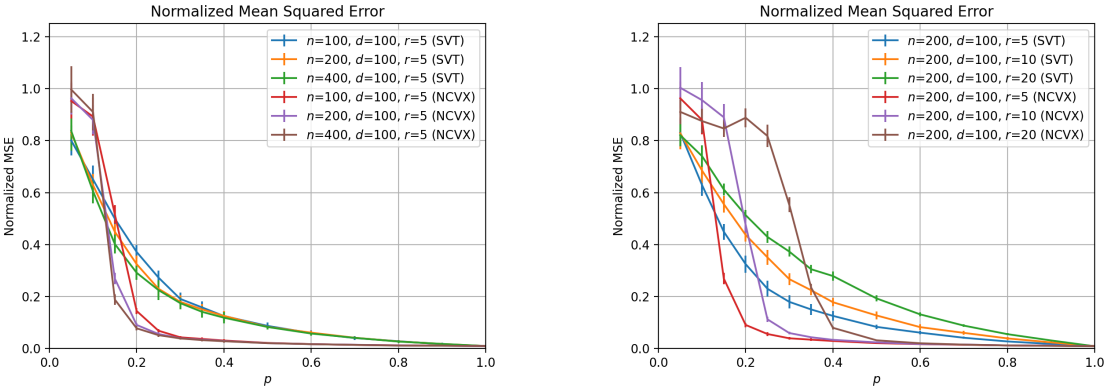
$\mathbb{E}_{\text{emp}}\|\hat{X}\hat{\beta} - X\beta^*\| / \mathbb{E}_{\text{emp}}\|X\beta^*\|$, where $\mathbb{E}_{\text{emp}}$ denotes the sample mean over the 100 instances.

Specifically, we consider the following parameter configurations:

- (Configuration 1) Fix $d = 100, r = 5, \eta = \sigma = 1$, and let $n \in \{100, 200, 400\}$.
- (Configuration 2) Fix $n = 200, d = 100, r = 5, \eta = \sigma = 1$, and let $r \in \{5, 10, 20\}$.

For each setup, we compute the normalized MSE for $p \in \{k/20 : k \in [20]\}$. The normalized MSE for the two configurations are illustrated in Figure 4.1 along with confidence intervals (2 standard errors).

Observe that even when we only have partial access to the design matrix (i.e., $p < 1$), the normalized MSE stays nearly as low as that of the full observation setup ($p = 1$), until p decreases below a certain threshold. For example, when $r = 5$, the normalized MSE of Algorithm 2 implemented with the nonconvex method remains almost unchanged for all $p \geq 0.2$ (see Figure 4.1a). This implies that the imputed dataset contains almost the same amount of information as the completely observed dataset that can be used for regression. Moreover, the minimum required fraction of measurements to capture the information increases proportionally to the rank of X (Figure 4.1b).



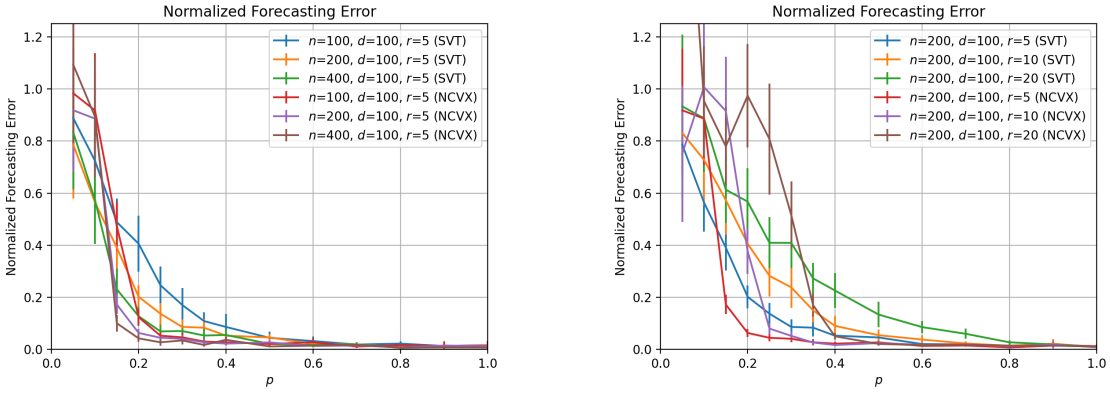
(a) In-sample prediction, configuration 1

(b) In-sample prediction, configuration 2

Figure 4.1: Normalized mean squared error of Algorithm 2, implemented with two matrix completion methods (the simple SVT and the nonconvex methods).

Experiment 2 In our second experiment, we consider the out-of-sample prediction error of Algorithm 2. We let $x^{\text{new}} = Vu$ where $V \in \mathbb{R}^{d \times r}$ is the factor matrix used to generate X, and $u \in \mathbb{R}^r$ is an arbitrary vector. For the concreteness of our exposition, here we present results with $u \sim N(0, I_r)$, however, we observe the same trend, regardless of the distributional assumption on u. z^{new} is generated from x^{new} by letting $z_j^{\text{new}} = x^{\text{new}} + w_j$ ($w_j \sim N(0, \eta^2)$) with probability p, and $z_j^{\text{new}} = *$ with probability $1-p$, independently for each $j \in [d]$. We conduct

experiments in the same setup as in our experiment 1. The normalized squared forecasting error (SFE), $\mathbb{E}_{\text{emp}}\|\hat{x}^{\text{new}}\hat{\beta} - x^{\text{new}}\beta^*\|_2^2 / \mathbb{E}_{\text{emp}}\|x^{\text{new}}\beta^*\|_2^2$, is used to measure the performance of algorithms. We can observe that the out-of-sample prediction error exhibits similar trends to that of the in-sample prediction error (MSE) by comparing Figure 4.1 with Figure 4.2. Note that the normalized out-of-sample prediction error has much larger variability than the normalized MSE because the MSE is the sum of squared errors from n independent samples, whereas the SFE consists of the error for a single instance.



(a) Out-of-sample prediction, configuration 1

(b) Out-of-sample prediction, configuration 2

Figure 4.2: Normalized out-of-sample prediction error of Algorithm 2, implemented with two matrix completion methods (the simple SVT and the nonconvex methods).

■ 4.7 Summary of the Chapter

In this chapter, we considered the errors-in-variables regression problem and studied prediction performance of ridgeless regression estimator that is computed from imputed covariate data. Unlike other works in the errors-in-variables regression literature, our main aim is at predicting the response rather than identifying the true underlying linear regression model. In Section 4.4, we argued upper bounds for prediction error of Algorithm 2 implemented with arbitrary matrix completion oracle, and in Section 4.5, we focused on a specific instance of the algorithm implemented with the simple SVT method. Noticing that Algorithm 2 implemented with the SVT method is equivalent to principal component regression when $p = 1$, our analysis in this chapter provides finite-sample analysis for the principal component regression. An interesting open question for further research is to investigate if the imputation idea can be also beneficial in kernel regression setting, or more broadly, in the nonlinear settings such as autoencoders.

Chapter 5

Sample-efficient Reinforcement Learning

5.1 Introduction

Reinforcement Learning (RL) has emerged as a promising technique for a variety of decision-making tasks, highlighted by impressive successes of deep reinforcement learning[1] in solving Atari games [103] and Go [131, 132]. However, generic RL methods suffer from the "curse-of-dimensionality," i.e., they require prohibitively many number of samples to achieve the learning goals as the number of model parameters increases. For example, the classical minimax theory for nonparametric estimation [140, 151] suggests that learning a Q-function (a function that encodes the value of an action at a state) within an accuracy of ϵ would require $\Omega\big(\frac{1}{\epsilon^{d_1+d_2+2}}\big)$ samples where d_1, d_2 denote the dimensions of the state and action spaces.

On the other hand, many RL tasks seem to require much less number of samples in practice. As a matter of fact, the conditions and the algorithms to achieve sample-efficient RL have been studied for at least two decades. Early works posit that the Q-function can be approximated well using low-dimensional features associated with states and actions, and use that low-dimensional representation of the Q-function to boost sample efficiency [149, 150, 101, 110, 98]. More recently, there is a growing body of results that attempt to show how sample efficiency is achievable in RL for particular model classes, e.g., Reactive POMDPs [84], Block MDPs [49], Linear MDPs [163], and Linear Mixture MDPs [104] to name a few. While these works have significantly advanced our understanding of the sample efficiency in RL, we may not know what features to use a priori, and the model assumptions may not be verifiable.

With these motivations, the primary goal in this chapter is to learn the optimal Q-function in a data-efficient manner by automatically detecting the underlying lower-dimensional structure (if any), *without* the need for any additional information such as the knowledge of features. In summary, we ask the following questions[2] in this chapter:

> "*Is there a universal representation of Q-function that facilitates learning the optimal Q-function in a data-efficient manner? If so, how can we exploit it?*"

[1] a deep neural network version of Q-learning, termed deep Q-networks by DeepMind.
[2] See Questions 5.1 and 5.2 for precise statements of the two questions

■ 5.1.1 Setup and Problem Statement

Markov Decision Process (MDP) We consider the standard setup of infinite-horizon discounted MDP, which is described by a quintuple, $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively; $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel where $\Delta(\mathcal{S})$ denotes the space of probability measures over $\mathcal{S}$; $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function; and $\gamma \in [0, 1)$ is the factor of discounting rewards in the future.

A policy $\pi : \mathcal{S} \to \mathcal{A}$ specifies a decision-making strategy, with which an agent chooses her actions based on the current state. A policy can be either deterministic or stochastic, however, we restrict ourselves to deterministic policies in this chapter. Given a policy π, we can define a function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that

$$Q^\pi(s, a) = \mathbb{E}_{P,\pi}\bigg[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \;\bigg|\; s_0 = s, a_0 = a\bigg],$$

which is the expectation of the infinite series of discounted rewards $\gamma^t R(s_t, a_t)$, on condition that $(s_0, a_0) = (s, a)$, $s_t \sim P(s_{t-1}, a_{t-1})$, and $a_t = \pi(s_t)$. Similarly, we can also define another function $V^\pi : \mathcal{S} \to \mathbb{R}$ by replacing the condition $a_0 = a$ with $a_0 = \pi(s_0)$ as follows:

$$V^\pi(s) = \mathbb{E}_{P,\pi}\bigg[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \;\bigg|\; s_0 = s, a_0 = \pi(s)\bigg].$$

In the reinforcement learning literature, Q^π and V^π are commonly referred to as the Q-function ('quality,' or action-value function) and the V-function ('value,' or state-value function) associated to the policy π, respectively.

A policy $\pi^* \in \mathcal{A}^\mathcal{S}$ is called an optimal policy if $V^{\pi^*}(s) = \sup_\pi V^\pi(s)$ for all $s \in \mathcal{S}$, and the function $V^* : \mathcal{S} \to \mathbb{R}$ such that $V^*(s) = \sup_{\pi \in \mathcal{A}^\mathcal{S}} V^\pi(s),\ \forall s \in \mathcal{S}$ is called the optimal value function. Remarkably, an optimal policy exists, and we can find at least one by decoding the optimal value function. To see the reason why, we define an operator $\mathcal{T} : \mathbb{R}^\mathcal{S} \to \mathbb{R}^\mathcal{S}$, which is commonly referred to as the Bellman operator, such that

$$\mathcal{T}V(s) = \sup_{a \in \mathcal{A}} \Big\{ R(s, a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a)}\big[V(s')\big] \Big\}.$$

Then we observe that V^* is a fixed point of $\mathcal{T}$ due to the optimality of V^*, i.e., the following Bellman equation holds:

$$V^*(s) = \mathcal{T}V^*(s), \quad \forall s \in \mathcal{S}.$$

It is known that $\mathcal{T}$ is a contraction with respect to the ℓ_∞-norm. Thus, there exists a unique

fixed point of $\mathcal{T}$, which is V^*. We can also define an optimal Q-function from V^*:

$$\begin{aligned} Q^*(s,a) &= R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a)}\big[V^*(s')\big] \\ &= R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a)}\Big[\max_{a' \in \mathcal{A}} Q^*(s',a')\Big]. \end{aligned} \tag{5.1}$$

An optimal policy π^* can be derived from Q^* by greedily choosing a that maximizes $Q^*(s,a)$ for given s, i.e., by letting $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$.

Q-Learning Although π^* can be extracted from Q^*, we do not have access to the Q^*-function. Actually, Q^* itself must be estimated from data generated by the MDP. Q-learning algorithm – proposed in the Ph.D. thesis of Watkins [159] – is a model-free reinforcement learning algorithm that attempts to learn Q^* using the Bellman equation (5.1) for Q^*. Q-learning is model-free in the sense that it does not require prior knowledge of model parameters P and R, but instead, it uses simulation or experimental information to compute estimates of the expected rewards.

Q-learning proceeds in a similar manner to Sutton's method of temporal differences [145, 143]. More precisely, Q-learning starts with an initial guess for Q^* (given in the form a table when $|\mathcal{S}|, |\mathcal{A}| < \infty$), and then repeat the loop of three inner steps – (i) choose an action a_t at state s_t, (ii) measure the reward $R(s_t, a_t)$, (iii) update the guess for Q^* using (5.1) – which is referred to as value iteration update – and move to $s_{t+1} \sim P(s_t, a_t)$ – until termination. It is known that Q-learning converges to Q^* with probability 1, provided that all state-action pairs are sampled sufficiently often [158, 148, 72, 54].

Despite the nice convergence property, Q-learning requires $\Omega\big(|\mathcal{S} \times \mathcal{A}|\big)$ number of samples to explore all possible actions at adll states. Also, we note that a real-valued function defined on $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ has $|\mathcal{S} \times \mathcal{A}|$ degrees of freedom, and thus, this sample complexity lower bound is unavoidable unless we impose additional structural assumptions on Q^* (or the underlying MDP). In this chapter, we ask the following two questions in pursuit of finding a principled way to reduce the sample complexity requirement.

Question 5.1. Is there a class of Q^*-functions, which can be learned from only $o(|\mathcal{S} \times \mathcal{A}|)$ number of samples?

Question 5.2. If the answer to Question 5.1 is positive, then is there an algorithmic procedure for sample-efficient Q-learning? What would be its sample complexity to produce $\hat{Q}$ such that $\|\hat{Q} - Q^*\|_\infty \le \epsilon$?

5.1.2 Contributions and Organization of the Chapter

The main contributions in this chapter are the methodological proposal of Q-learning algorithm that exploits low-rank structure in Q^* (Algorithm 3) and its theoretical analysis. We argue that if the Q^* function admits a low-rank structure (to be discussed in Section 5.3), that can be utilized by means of matrix completion to enhance the sample efficiency in value iteration updates. That is, it suffices to explore only a small fraction of state-action pairs instead of probing all pairs in order to update the current guess about Q^*.

As a main theoretical contribution of the chapter, we answer to both Questions 5.1 and 5.2 in the affirmative by showing that the proposed Q-learning procedure converges to Q^* with the desired improvement in sample complexity. Without any structural assumptions, the number of samples (explorations of state-action pairs) required to learn Q^* within an accuracy of ϵ scales as $|\mathcal{S}||\mathcal{A}|/\epsilon^2$. We show that if the matrix completion oracle in use satisfies a certain ℓ_∞-contraction property, then it is possible to learn Q^* from only $O\big(\text{rank}(Q^*)\cdot(|\mathcal{S}|\vee|\mathcal{A}|)/\epsilon^2\big)$ number of samples. This conclusion is summarized in Table 5.1 with comparison to other results that do not utilize such structural assumptions. We also complement our analysis with numerical experiments on simple stochastic control tasks.

The rest of this chapter is organized as follows. In Section 5.3, we consider the spectral representation of the Q-function by viewing it as a possibly infinite-dimensional matrix $Q\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. This allows us to define the notion of rank for Q^* and to consider a class of (approximately) low-rank Q^* functions. In Section 5.4, we propose an algorithmic framework to estimate 'low-rank' Q^*-functions with aid of a matrix completion oracle. The key idea is that it suffices to explore $\text{rank}(Q^*)\cdot\big(|\mathcal{S}|+|\mathcal{A}|\big)$ number of state-action pairs to execute one round of value iteration, which can be much smaller than $|\mathcal{S}||\mathcal{A}|$. In Section 5.5, we argue that if the matrix completion oracle in use satisfies a certain ℓ_∞-contraction property (Section 5.5.1), then the proposed Q-learning procedure converges to Q^* with the desired improvement in sample complexity (Section 5.5.2). It is proven in Section 5.6 that there exists at least one matrix completion oracle that has such a property. In Section 5.7, we provide some numerical results. We find that the Q^*-function has a low-rank structure for several well-known control tasks, e.g., the inverted pendulum. Moreover, we observe that our proposed method that exploits the low-rank structure leads to a significant improvement in sample complexity over the methods that do not.

5.2 Related Work

Reinforcement Learning For problems with finite $\mathcal{S}$ and finite $\mathcal{A}$, there has been a great effort in learning an ϵ-optimal policy rather than just learning an ϵ-optimal Q-function. Re-

Table 5.1: Summary of our sample complexity results, a few selected works from the literature, and lower bounds. Our results are paraphrased from Theorem 5.5.2.

	Finite $\mathcal{S}$ & Finite $\mathcal{A}$	Cont. $\mathcal{S}$ & Finite $\mathcal{A}$	Cont. $\mathcal{S}$ & Cont. $\mathcal{A}$
This thesis (w/ low-rank assmp.)	$\tilde{O}\Big(\frac{\lvert\mathcal{S}\rvert \vee \lvert\mathcal{A}\rvert}{\epsilon^2}\Big)$	$\tilde{O}\Big(\frac{1}{\epsilon^{d_1+2}}\Big)$	$\tilde{O}\Big(\frac{1}{\epsilon^{d_1 \vee d_2+2}}\Big)$
Selected works from the literature (w/o low-rank assmp.)	$\tilde{O}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon^2}\Big)$ [129, 130]	$\tilde{O}\Big(\frac{1}{\epsilon^{d_1+2}}\Big)$ [55]	N/A
Lower bound (w/o low-rank assmp.)	$\tilde{\Omega}\Big(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon^2}\Big)$ [13]	$\tilde{\Omega}\Big(\frac{1}{\epsilon^{d_1+2}}\Big)$ [126]	$\Omega\Big(\frac{1}{\epsilon^{d_1+d_2+2}}\Big)$ [151]

cently, Sidford et al. [129] showed a sample complexity upper bound of $\tilde{O}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\big)$ that is optimal with respect to both accuracy ϵ and the dependence on the discount factor γ. A matching lower bound was proved earlier by Azar, Munos, and Kappen [13]. RL problems with continuous state and action spaces have received significantly less attention in the literature. While there are practical RL algorithms to deal with continuous domains [154, 94, 68], theoretical understanding (including sample complexity analysis) on this class of problems is limited [9].

Sample-efficient RL There have been numerous works that study the conditions and algorithms to achieve sample-efficient RL. Early works focus on approximating Q function using low-dimensional features associated with states and actions, and use the obtained low-dimensional representation to achieve sample efficiency [149, 150, 101, 110, 98]. Following the recent resurgence of interest in RL, there is a growing body of theoretical works that attempt to show how sample efficiency is possible in RL for particular model classes; notable examples include Reactive POMDPs [84], Block MDPs [49], Linear MDPs [163], and Linear Mixture MDPs [104] to list a few. There are also a few works that propose more general frameworks based on structural conditions that allows for sample-efficient RL, e.g., Bellman rank [74], Witness rank [142], Bilinear classes [50], etc.

Low-rank Q^* Perhaps, the most closely related work to this chapter is the recent empirical paper by Yang et al. [164] that investigates the performance of deep Q-network in the setting where the Q-function has approximately low rank. Their method is similar to our approach that will be described in this chapter, however, they only consider the finite state/action space setup, and moreover, they do not provide any theoretical analysis. In a sense, the results in this chapter provide a formal framework to understand the empirical success reported in [164], resolving the theoretical open problems raised in that work.

5.3 Spectral Representation and Low-rank Q^*-function

5.3.1 Model Assumptions on MDP

We make three assumptions on the regularity of the underlying MDP and its optimal Q-function. Note that we can discard Assumptions 5.2 and 5.3 when $\mathcal{S}, \mathcal{A}$ are finite.

Assumption 5.1. There exists $R_{\max} < \infty$ such that $|R(s,a)| \le R_{\max}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Assumption 5.2. The state space $\mathcal{S} \subseteq \mathbb{R}^{d_1}$ and the action space $\mathcal{A} \subseteq \mathbb{R}^{d_2}$ are compact.

Assumption 5.3. The optimal Q-function is L-Lipshitz with respect to the 1-product in $\mathcal{S} \times \mathcal{A}$, i.e.,

$$\big|Q^*(s_1,a_1) - Q^*(s_2,a_2)\big| \le L \cdot d_{\mathcal{S}\times\mathcal{A}}\big((s_1,a_1);(s_2,a_2)\big)$$

where $d_{\mathcal{S}\times\mathcal{A}}\big((s_1,a_1);(s_2,a_2)\big) = \|s_1 - s_2\|_2 + \|a_1 - a_2\|_2$.

Note that Assumption 5.1 implies that for any policy π, $|V^\pi(s)| \le V_{\max}$ for all s, where $V_{\max} := R_{\max}/(1-\gamma)$. By definition, this yields $|Q^*(s,a)| \le V_{\max}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Also, we remark that some form of smoothness assumption as in Assumption 5.3 is required to learn a MDP with continuous state/action space via discretization, and typically adopted in the literature [9, 126, 55].

5.3.2 Spectral Representation of Q^*

Let $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be the optimal Q-function. We consider the integral operator $K = K_{Q^*} : L^2(\mathcal{S}) \to L^2(\mathcal{A})$ induced by Q^* such that

$$(Kh)(a) = \int_{s\in\mathcal{S}} Q^*(s,a) \cdot h(s)\, ds, \quad \forall a \in \mathcal{A}.$$

Here, $L^2(\mathcal{S}) = \{f : \mathcal{S} \to \mathbb{R} \text{ such that } \int_{s\in\mathcal{S}} f(s)^2 ds < \infty\}$ is the Hilbert space of (equivalent classes of) square-integrable functions on $\mathcal{S}$ endowed with the inner product $\langle f_1, f_2\rangle = \int_{s\in\mathcal{S}} f_1(s) \cdot f_2(s)\, ds$, and $L^2(\mathcal{A})$ is defined similarly. Through this lens, we obtain the spectral representation for Q^*. Note that when $\mathcal{S}, \mathcal{A}$ are finite, this reduces to the singular value decomposition of the matrix $[Q^*(s,a)] \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Theorem 5.3.1. *If Assumptions 5.1, 5.2, and 5.3 hold, then then there exist a nonincreasing sequence $(\sigma_i \ge \mathbb{R}_+ : i \in \mathbb{N})$ with $\sum_{i=1}^\infty \sigma_i^2 < \infty$ and orthonormal sets of functions $\{f_i \in L^2(\mathcal{S}) : i \in \mathbb{N}\}$ and $\{g_i \in L^2(\mathcal{A}) : i \in \mathbb{N}\}$ such that*

$$Q^*(s,a) = \sum_{i=1}^\infty \sigma_i f_i(s) g_i(a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{5.2}$$

The proof of Theorem 5.3.1 is omitted due to its technicalities and similarity to the classical proof of the spectral theorems; we refer the interested readers to [127, Appendix B] for a proof.

Theorem 5.3.1 motivates us to define the rank of Q^* as

$$\operatorname{rank}(Q^*) = \inf\big\{i \in \mathbb{N} : \exists \text{ a decomposition of } Q^* \text{ of the form (5.2) s.t. } \sigma_j = 0,\ \forall j > i\big\}.$$

We can also define the notion of δ-approximate rank by letting $\operatorname{rank}_\delta(Q^*) = \inf\{i \in \mathbb{N} : \exists \text{ a decomposition of } Q^* \text{ of the form (5.2) s.t. } \sum_{j=i+1}^{\infty} \sigma_i^2 < \delta\}$. In this chapter, we focus on Q-learning for Q^* with rank $r \ll |\mathcal{S}| \wedge |\mathcal{A}|$ for the conciseness of our exposition. We refer the interested readers to [127, Section 5.3] for results on the case when the rank of Q^* may not be small, but its δ-approximate rank is small for some $\delta > 0$.

Before moving to the next section, we provide an example of a classical MDP that exhibits low-rank structure in Q^* to motivate the readers. This example suggests that the rank of the optimal Q-function is approximately the same with the order of the dynamical system, and do not scale up with the size of state or action spaces.

Example 5.3.2. The linear quadratic regulator (LQR) problem considers designing a linear controller π for a linear dynamical system given by $s_{t+1} = As_t + Ba_t,\ \ a_t = \pi s_t$, by minimizing a quadratic cost (negative reward) function $R(s_t, a_t) = s_t^T E s_t + a_t^T F a_t$ where E, F are some positive definite matrices. Here, $s_t \in \mathbb{R}^{d_1}$ is the state of the system at time t, $a_t \in \mathbb{R}^{d_2}$ is the control input at t, and $A \in \mathbb{R}^{d_1 \times d_1}, B \in \mathbb{R}^{d_1 \times d_2}, \pi \in \mathbb{R}^{d_2 \times d_1}$ are matrices describing the system. According to linear-quadratic control theory [19], the value function for policy π can be expressed as $V^\pi(s_t) = s_t^T K_\pi s_t$ for some matrix K_π; thus, the Q-function is written as

$$\begin{aligned} Q^\pi(s,a) &= R(s,a) + \gamma V^\pi\big((As + Ba)\big) \\ &= s^T(E + \gamma A^T K_\pi A)s + 2\gamma s^T A^T K_\pi B a + a^T(F + \gamma B^T K_\pi B)a. \end{aligned}$$

Letting $A^T K_\pi B = \sum_{i=1}^r \tau_i u_i v_i^T$ be the SVD of $A^T K_\pi B$, we observe that

$$Q^\pi(s,a) = 2\gamma \sum_{i=1}^r \tau_i (u_i^T s) \cdot (v_i^T a) + (s^T M_{\mathcal{S}} s) \cdot 1_{\mathcal{A}}(a) + 1_{\mathcal{S}}(s) \cdot (a^T M_{\mathcal{A}} a)$$

where $M_{\mathcal{S}} = E + \gamma A^T K_\pi A$, $M_{\mathcal{A}} = F + \gamma B^T K_\pi B$ and $1_{\mathcal{S}}$ ($1_{\mathcal{A}}$, resp.) denotes the constant-1 function on $\mathcal{S}$ ($\mathcal{A}$, resp.). Observe that $1_{\mathcal{S}}$, $s^T M_{\mathcal{S}} s$, and $\{u_i^T s\}_{i=1}^r$ form an orthogonal set in $L^2(\mathcal{S})$ as long as $\mathcal{S}$ is symmetric (i.e., $\mathcal{S} = -\mathcal{S}$). Similarly, $1_{\mathcal{A}}$, $a^T M_{\mathcal{A}} a$, and $\{v_i^T a\}_{i=1}^r$ form an orthogonal set in $L^2(\mathcal{A})$ when $\mathcal{A}$ is symmetric. Thus, the rank of Q^π is at most $\min\{d_1, d_2\} + 2$, and so is the rank of Q^*. Therefore, $\operatorname{rank}(Q^*)$ is significantly smaller than $|\mathcal{S}| \sim 2^{d_1}$ or $|\mathcal{A}| \sim 2^{d_2}$ (after quantization).

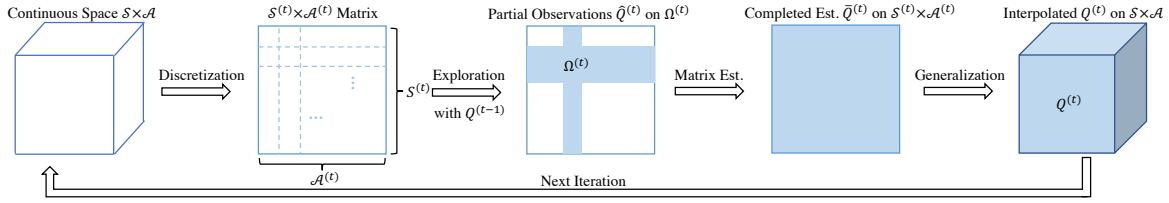


Figure 5.1: An illustration of the 4 steps in Algorithm 3 (Q-learning with matrix completion).

■ 5.4 Sample-efficient Q-learning with Matrix Completion

■ 5.4.1 Sampling with Generative Model

Here, we specify the sampling model assumed throughout the remainder of this chapter. The sampling model described in the next assumption allows the agent to freely access to any state-action pair (s, a), query the reward $R(s, a)$ at the state-action pair, and draw a sample instance of next state $s' \sim P(s, a)$.

Assumption 5.4. The agent has access to a generative model (or a simulator), which provides sampling access to any state in the environment of the agent's choice.

Assuming access to a generative model can be a natural assumption when we have a simulator of the environment [78, 77]. Note that the model in Assumption 5.4 is a stronger sampling model than the so-called online simulation model, in which the agent only has access to the trajectory determined by the environment itself. However, it is a weaker assumption than having complete knowledge of the environment.

■ 5.4.2 Algorithm: Q-learning with Low-rank Matrix Completion

In this section, we describe a Q-learning algorithm that uses a matrix completion oracle as a subroutine to improve sample efficiency. At each iteration, our proposed Q-learning algorithm explores only a small fraction of the entire state-action pairs, and extrapolates the revealed information to update the current guess of Q^* at all state-action pairs with aid of the matrix completion algorithm. More precisely, each iteration consists of four steps: discretization, exploration, matrix estimation and generalization. Here we provide a narrative overview of each step; see Algorithm 3 for a compact description and Figure 5.1 for a visual summary.

- **Step 1. Discretization.** At iteration t, discretize $\mathcal{S}, \mathcal{A}$ by constructing $\beta^{(t)}$-nets, $\mathcal{S}^{(t)} \subseteq \mathcal{S}$ and $\mathcal{A}^{(t)} \subseteq \mathcal{A}$, for an appropriately chosen resolution $\beta^{(t)} \in (0, 1)$. Note that $|\mathcal{S}^{(t)}| = O\big((1/\beta^{(t)})^{d_1}\big)$, $|\mathcal{A}^{(t)}| = O\big((1/\beta^{(t)})^{d_2}\big)$ with Assumption 5.2.
- **Step 2. Exploration.** Choose a subset $\Omega^{(t)} \subseteq \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$ and refine the current guess of Q^* at all $(s, a) \in \Omega^{(t)}$. For each $(s, a) \in \Omega^{(t)}$, the agent generates $N^{(t)}$ independent

Algorithm 3: Low-rank Q-learning using matrix completion

Input: $\mathcal{S}$, $\mathcal{A}$, γ, $Q^{(0)}$, T, $\{\beta^{(t)}\}_{t=1,\dots,T}$, $\{N^{(t)}\}_{t=1,\dots,T}$
simulator (generative model), matrix completion oracle φ_{MC}

Output: $Q^{(T)}$, the Q-value oracle after T iterations

For $t \in [T]$:

1. Discretize $\mathcal{S}$ and $\mathcal{A}$ so that $\mathcal{S}^{(t)}$ is a $\beta^{(t)}$-net of $\mathcal{S}$ and $\mathcal{A}^{(t)}$ is a $\beta^{(t)}$-net of $\mathcal{A}$.
2. Explore a few selected state-action pairs. Specifically,
   (a) Select a subset $\Omega^{(t)} \subseteq \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$.
   (b) For each $(s,a) \in \Omega^{(t)}$, generate $N^{(t)}$ independent samples of the next states $\{s_i'\}_{i\in[N^{(t)}]}$ with the simulator, and let
   $$\hat{Q}^{(t)}(s,a) \leftarrow R(s,a) + \gamma \frac{1}{N^{(t)}} \sum_{i=1}^{N^{(t)}} V^{(t-1)}(s_i').$$
3. Extrapolate $\hat{Q}^{(t)}$ defined on $\Omega^{(t)}$ to the entire $\mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$ with matrix completion:
   $$\bar{Q}^{(t)} \leftarrow \varphi_{\mathrm{MC}}\big(\hat{Q}^{(t)}; \Omega^{(t)}\big).$$
4. Generalize the estimate $\bar{Q}^{(t)}$ defined on $\mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$ to the original space $\mathcal{S} \times \mathcal{A}$. Subsequently, update $V^{(t)}(s) \leftarrow \sup_{a\in\mathcal{A}} Q^{(t)}(s,a)$ for all $s \in \mathcal{S}$.

samples of the next states $\{s_i'\}_{i\in[N^{(t)}]}$ with the simulator, and refine her guess of $Q^*(s,a)$ based on $Q^{(t-1)}$ and $\{s_i'\}_{i\in[N^{(t)}]}$:

$$\hat{Q}^{(t)}(s,a) \leftarrow R(s,a) + \gamma \cdot \frac{1}{N^{(t)}} \sum_{i=1}^{N^{(t)}} \sup_{a'\in\mathcal{A}} Q^{(t-1)}(s,a'). \tag{5.3}$$

- **Step 3. Matrix Completion.** Given $\hat{Q}^{(t)}(s,a) : \Omega^{(t)} \to \mathbb{R}$ updated in Step 2, the agent extrapolates $\hat{Q}^{(t)}$ to obtain an improved estimate of Q^* for the entire $\mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$. This can be viewed as a matrix estimation problem. With the matrix completion oracle φ_{MC} provided, the agent gets $\bar{Q}^{(t)} \leftarrow \varphi_{\mathrm{MC}}(\hat{Q}^{(t)})$.
- **Step 4. Generalization.** The estimates $\bar{Q}^{(t)}(s,a)$, $(s,a) \in \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$ are generalized to $\mathcal{S} \times \mathcal{A}$ in this step. This task can be done with any supervised learning algorithm. We simply utilize the 1-nearest neighbor: for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q^{(t)}(s,a) = \bar{Q}^{(t)}(s',a')$ where $(s',a') \in \arg\min_{(s_t,a_t)\in\mathcal{S}^{(t)}\times\mathcal{A}^{(t)}} d_{\mathcal{S}\times\mathcal{A}}\big((s_t,a_t),(s,a)\big)$.

We remark that Steps 2 and 3 are the essential steps that lead to the enhanced sample efficiency of Algorithm 3. Steps 1 and 4 are auxiliary steps that discretize and interpolate the state/action spaces.

■ 5.5 Convergence and Sample Complexity Analysis

In this section, we argue that our proposed algorithm (Algorithm 3) correctly converges to Q^* with high probability, requiring only $\widetilde{O}\big(\mathrm{rank}(Q^*) \cdot (|\mathcal{S}| \vee |\mathcal{A}|\big)$ number of queries to the simulator. To prove this claim, we need the matrix completion oracle in use to satisfy a certain ℓ_∞-contraction property. We start our discussion by defining the property.

■ 5.5.1 ℓ_∞-contraction Property of Matrix Completion

Recall that we described Algorithm 3 with a generic matrix completion oracle φ_{MC}, avoiding specifying what particular algorithm is used. That was intentional because the proposed Q-learning algorithm successfully converges to Q^* as long as the MC oracle in use satisfies a certain ℓ_∞-contraction property, which is stated in the next.

Before defining the property, recall from (3.7) that $\mathcal{R}_\Omega(M)_{ij} = M_{ij}$ if $(i,j) \in \Omega$ and $\mathcal{R}_\Omega(M)_{ij} = 0$ otherwise. We define $\overline{\mathcal{R}}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \overline{\mathbb{R}}^{n_1 \times n_2}$ similarly so that $\overline{\mathcal{R}}_\Omega(M)_{ij} = M_{ij}$ if $(i,j) \in \Omega$ and $\overline{\mathcal{R}}_\Omega(M)_{ij} = *$ otherwise.

Definition 5.5.1. A matrix completion oracle φ_{MC} has the (c_{MC}, ϵ_{MC})-contraction property at $M \in \mathbb{R}^{n_1 \times n_2}$ with respect to $\Omega \subseteq [n_1] \times [n_2]$ if

$$\big\|\varphi_{\mathrm{MC}}\big(\overline{\mathcal{R}}(M+E)\big) - M\big\|_\infty \le c_{MC}\big\|\mathcal{R}_\Omega(E)\big\|_\infty, \qquad \forall E \text{ s.t. } \|E\|_\infty \le \epsilon_{MC}.$$

Now, we make the following assumption, which captures the operational meaning of the 'success' of the MC oracle and will serve as a pivotal premise for the success of the entire Q-learning algorithm. For notational convenience, we let $Q(\mathcal{S}', \mathcal{A}')$ denote the $|\mathcal{S}'| \times |\mathcal{A}'|$ matrix $\big[Q(s,a) : \ (s,a) \in \mathcal{S}' \times \mathcal{A}'\big]$, whose entries are indexed by $(s,a) \in \mathcal{S}' \times \mathcal{A}'$.

Assumption 5.5. There exists a constant $C_\Omega \ge 1$ such that for every iteration $t \in [T]$ in Algorithm 3, it is possible to choose $\Omega^{(t)}$ that satisfies

1. $|\Omega^{(t)}| \le C_\Omega\big(|\mathcal{S}^{(t)}| + |\mathcal{A}^{(t)}|\big)$, and
2. the matrix completion oracle φ_{MC} in use has (c_{MC}, ϵ_{MC})-contraction property at $Q^*\big(\mathcal{S}^{(t)}, \mathcal{A}^{(t)}\big)$ with respect to $\Omega^{(t)}$ for some $\epsilon_{MC} > 0$.

Here we remark that Assumption 5.5 ensures that it is possible to choose $\Omega^{(t)}$ at Step 2

of every iteration in Algorithm 3 so that $\Omega^{(t)} \leq C_\Omega\big(|\mathcal{S}| + |\mathcal{A}|\big)$ and

$$\max_{(s,a)\in\mathcal{S}^{(t)}\times\mathcal{A}^{(t)}} \big|\bar{Q}^{(t)}(s,a) - Q^*(s,a)\big| \leq c_{MC} \cdot \max_{(s,a)\in\Omega^{(t)}} \big|\hat{Q}^{(t)}(s,a) - Q^*(s,a)\big|.$$

In Section 5.5.2, we prove the convergence of Algorithm 3, under Assumption 5.5. Thereafter, in Section 5.6, we will show that there exists at least one matrix completion oracle that satisfies Assumption 5.5.

■ 5.5.2 Convergence and Sample Complexity of Algorithm 3

To that end, let the algorithm start with initialization $Q^{(0)}(s,a) = 0,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$ and hence $V^{(0)}(s) = 0,\ \forall s \in \mathcal{S}$. That is, $|Q^{(0)}(s,a) - Q^*(s,a)| \leq V_{\max},\ \forall (s,a)\in\mathcal{S}\times\mathcal{A}$. For the sake of notational brevity, we let $d_1 = d_2 = d$ in the sequel. We remark that our theorems apply equally to the general settings by simply replacing d with $d_1 \vee d_2$ whenever it appears in the proof.

Theorem 5.5.2. *Let $Q^{(0)}(s,a) = 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. If Assumption 5.5 is satisfied with $\epsilon_{MC} \geq 2\gamma V_{\max}$, then for any $\delta \in (0,1)$, there exists an algorithmic choice of $\{\beta^{(t)}, N^{(t)}, \Omega^{(t)}\}_{t\in[T]}$ for Algorithm 3 that leads to*

$$\mathbb{P}\bigg(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big|Q^{(t)}(s,a) - Q^*(s,a)\big| \leq (2\gamma c_{MC})^t V_{\max}, \quad \forall 1 \leq t \leq T\bigg) \geq 1-\delta. \tag{5.4}$$

Furthermore, if $\gamma < \frac{1}{2c_{MC}}$, then there exists $T = \Theta\big(\log\frac{1}{\epsilon}\big)$ for which

$$\mathbb{P}\bigg(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big|Q^{(T)}(s,a) - Q^*(s,a)\Big| \leq \epsilon\bigg) \geq 1-\delta, \tag{5.5}$$

and the sample complexity measured in the number of calls to the simulator is $\widetilde{O}\big(\frac{1}{\epsilon^{d+2}}\cdot\log\frac{1}{\delta}\big)$.

Proof Sketch of Theorem 5.5.2 While the complete proof of Theorem 5.5.2 is deferred until Section 5.9.1, here we sketch the main ideas in the proof. Observe that Algorithm 3 updates the guess $Q^{(t)}$ following the sequence depicted in Figure 5.2. To prove the linear convergence in (5.4), we decouple the analysis of the three steps (Steps 2, 3, and 4 of Algorithm 3) and observe the following inequalities.

- In Step 2, we generate $N^{(t)}$ samples using a simulator, and make a Bellman update based on the current belief $Q^{(t-1)}$, which may be biased. Thus, $Q^{(t)}(s,a) - Q^*(s,a)$ is the sum of $N^{(t)}$ (possibly biased) sub-Gaussian random variables, and we obtain the inequality

(Lemma 5.9.1):

$$\left\|\mathcal{R}_{\Omega^{(t)}}\left(\hat{Q}^{(t)}-Q^*\right)\right\|_\infty \le \gamma\Bigg(\underbrace{\left\|Q^{(t-1)}-Q^*\right\|_\infty}_{\text{systematic bias}} + \underbrace{\sqrt{\frac{2V_{\max}^2}{N^{(t)}}\log\left(\frac{2|\Omega^{(t)}|T}{\delta}\right)}}_{\text{statistical fluctuation due to sampling}}\Bigg).$$

- In Step 3, if φ_{MC} is assumed to have (c_{MC}, ϵ_{MC})-contraction property at $Q^*\left(\mathcal{S}^{(t)}, \mathcal{A}^{(t)}\right)$ with respect to $\Omega^{(t)}$ by Assumption 5.5, and therefore,

$$\left\|\bar{Q}^{(t)}-Q^*(\mathcal{S}^{(t)},\mathcal{A}^{(t)})\right\|_\infty \le c_{MC}\left\|\mathcal{R}_{\Omega^{(t)}}\left(\hat{Q}^{(t)}-Q^*\right)\right\|_\infty.$$

- In Step 4, L-Lipschitzness of Q^* (Assumption 5.3) implies that

$$\left\|Q^{(t)}-Q^*\right\|_\infty \le \left\|\bar{Q}^{(t)}-Q^*(\mathcal{S}^{(t)},\mathcal{A}^{(t)})\right\|_\infty + 2L\beta^{(t)}.$$

Now it remains to combine the three inequalities and choose appropriate parameters to balance them. If we choose parameters

$$\beta^{(t)} = \frac{V_{\max}}{8L}(2\gamma c_{MC})^t, \quad |\Omega^{(t)}| = C_\Omega\left(|\mathcal{S}^{(t)}| + |\mathcal{A}^{(t)}|\right), \quad N^{(t)} = \frac{8}{(2\gamma c_{MC})^{2(t-1)}}\log\left(\frac{2|\Omega^{(t)}|T}{\delta}\right),$$

for each $t \in [T]$, then we can prove the claims in Theorem 5.5.2.

We remark that it is possible to show $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{(t)}(s,a)-Q^*(s,a)\right| \le \alpha^t V_{\max}$ for any $\alpha > \gamma c_{MC}$ with a more sophisticated choice of parameters. Accordingly, the conclusion for sample complexity, (5.5), requires $\gamma < \frac{\alpha}{c_{MC}}$. Thus, the constant c_{MC} in Assumption 5.5 determines the range of MDP parameter γ, for which such reduction in sample complexity – $\widetilde{O}\left(\frac{1}{\epsilon^{d+2}}\cdot\log\frac{1}{\delta}\right)$ instead of $\widetilde{O}\left(\frac{1}{\epsilon^{2d+2}}\cdot\log\frac{1}{\delta}\right)$ – can be achieved.

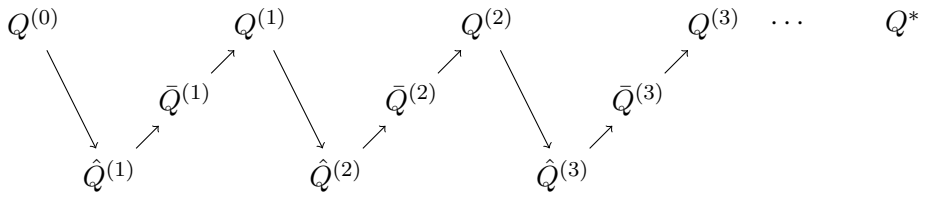


Figure 5.2: A schematic illustration of the sequence of updating $Q^{(t)}$ in Algorithm 3.

■ 5.6 A Matrix Completion Oracle Fulfilling Assumption 5.5

Although we showed Algorithm 3 linearly converges to Q^* in the ℓ_∞-norm[3] with the promised reduction in sample complexity, it is contingent on Assumption 5.5. In this section, we show that there exists at least one matrix completion oracle φ_{MC} that has the (c_{MC}, ϵ_{MC})-contraction property, and thus, Assumption 5.5 can be fulfilled.

Anchor Sets To begin with, we define the notion of anchor sets of states and actions for Q^*. Intuitively, this notion intends to identify a small subset of states and actions, $\mathcal{S}^\sharp \subseteq \mathcal{S}, \mathcal{A}^\sharp \subseteq \mathcal{A}$, such that the function value $Q^*(s,a)$ at any state-action pair (s,a) can be estimated from the matrix $Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp)$. In other words, the states in $\mathcal{S}^\sharp$ are sufficiently diverse to represent the behavior of Q^* over all states $s \in \mathcal{S}$. Likewise, a similar interpretation holds for $\mathcal{A}^\sharp$.

Definition 5.6.1. Let $\mathcal{S}^\sharp = \{s_i^\sharp\}_{i=1}^{R_s} \subseteq \mathcal{S}$ be a set of R_s states in $\mathcal{S}$ and $\mathcal{A}^\sharp = \{a_i^\sharp\}_{i=1}^{R_a} \subseteq \mathcal{A}$ be a set of R_a actions in $\mathcal{A}$. The pair $(\mathcal{S}^\sharp, \mathcal{A}^\sharp)$ is an anchor set of states and actions for Q^* if $\text{rank } Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp) = \text{rank } Q^*$.

Note that if $(\mathcal{S}^\sharp, \mathcal{A}^\sharp)$ is an anchor set of states and actions for Q^*, then $\text{rank } Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp) = \text{rank } Q^*(\mathcal{S}', \mathcal{A}')$ for all $\mathcal{S}', \mathcal{A}'$ such that $\mathcal{S}^\sharp \subseteq \mathcal{S}' \subseteq \mathcal{S}$ and $\mathcal{A}^\sharp \subseteq \mathcal{A}' \subseteq \mathcal{A}$. It is because $\text{rank } Q^*(\mathcal{S}', \mathcal{A}') \leq \text{rank } Q^*$ for all $\mathcal{S}' \subseteq \mathcal{S},\ \mathcal{A}' \subseteq \mathcal{A}$.

Indeed, $\mathcal{S}^\sharp$ and $\mathcal{A}^\sharp$ will be used to construct our exploration sets and we want them to have small size. Finding only a few diverse states and actions is arguably easy in practice. For stochastic control tasks experimented in Section 5.7, we simply pick a few states and actions that are far from each other in their respective metric. We remark that assuming the existence of some "anchor" elements (i.e., elements having some special, relevant properties) is common in feature-based reinforcement learning [163] or matrix factorization such as topic modeling [10].

Schur-complement-based Matrix Completion Next, we make the following observation, motivated by the definition of the Schur complement.

Lemma 5.6.2. *Let* $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathbb{R}^{n_1 \times n_2}$. *If* $\text{rank } A = \text{rank } M$, *then* $D = CA^\dagger B$.

Proof. Since $\text{rrank} \begin{bmatrix} A & B \end{bmatrix} \geq \text{rrank } A = \text{rank } A = \text{rank } M = \text{rrank } M$, there exists a matrix P such that $\begin{bmatrix} C & D \end{bmatrix} = P \begin{bmatrix} A & B \end{bmatrix}$. Also, observe that $\text{crank} \begin{bmatrix} A & B \end{bmatrix} \leq \text{crank } M = \text{rank } M = \text{rank } A = \text{crank } A$. That is, the column space of B is a subspace of the column space of A. It follows that $AA^\dagger A = A$ and $AA^\dagger B = B$ because the left multiplication of $AA^\dagger$ is the

[3] L^∞ when $\mathcal{S} \times \mathcal{A}$ is not countable.

projection on the column space of A. We obtain

$$\begin{bmatrix} C & D \end{bmatrix} = P \begin{bmatrix} A & B \end{bmatrix} = P \begin{bmatrix} AA^\dagger A & AA^\dagger B \end{bmatrix} = \begin{bmatrix} PA & PAA^\dagger B \end{bmatrix}.$$

Therefore, $PA = C$ and $D = PAA^\dagger B = CA^\dagger B$. ■

That is, if $r = \text{rank } A = \text{rank } M \ll n_1 \wedge n_2$, then we can recover the entire matrix M from the three submatrices A, B, C. Note that the total number of elements in A, B, C can be as small as $r \cdot (n_1 + n_2 - r) \ll n_1 n_2$. Based on this observation, we consider the following map:

$$\varphi_{\text{Schur}} : Z = \begin{bmatrix} A & B \\ C & * \end{bmatrix} \quad \mapsto \quad \hat{M} = \begin{bmatrix} A & B \\ C & CA^\dagger B \end{bmatrix}. \tag{5.6}$$

This gives a simple method for matrix completion, and we show in the next theorem that φ_{Schur} has the (c_{MC}, ϵ_{MC})-contraction property for some $c_{MC}, \epsilon_{MC} > 0$.

Theorem 5.6.3. *Let* $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathbb{R}^{n_1 \times n_2}$ *with* $A \in \mathbb{R}^{r_1 \times r_2}$ *for some* $r_1 < n_1$ *and* $r_2 < n_2$. *Let* $\widetilde{A}, \widetilde{B}, \widetilde{C}$ *be matrices of the same dimension with* A, B, C, *respectively, and let* $\epsilon_A = \|\widetilde{A} - A\|_\infty$, $\epsilon_B = \|\widetilde{B} - B\|_\infty$, *and* $\epsilon_C = \|\widetilde{C} - C\|_\infty$. *If* rank A = rank $M = r$ *and* $\epsilon_A \le \frac{s_r(A)}{2\sqrt{r_1 r_2}}$, *then*

$$\left\|\widetilde{C}\widetilde{A}^\dagger\widetilde{B} - D\right\|_\infty \le \frac{\sqrt{r_1 r_2}}{s_r(A)} \left\{ \sqrt{2}\Big(\|B\|_\infty \epsilon_C + \|C\|_\infty \epsilon_B + \epsilon_B \epsilon_C \Big) + (1+\sqrt{5}) \frac{\sqrt{r_1 r_2}}{s_r(A)} \|B\|_\infty \|C\|_\infty \epsilon_A \right\}.$$

Let's see how Theorem 5.6.3 leads to the (c_{MC}, ϵ_{MC})-contraction property. To this end, we let $M, E \in \mathbb{R}^{n_1 \times n_2}$, $\Omega = \{(i,j) \in [n_1] \times [n_2] : i \in [r_1] \text{ or } j \in [n_2]\}$ for some $r_1 < n_1$, $r_2 < n_2$, and $Z = \overline{\mathcal{R}}_\Omega(M + E)$. Note that if $\|\mathcal{R}_\Omega(M)\|_\infty \vee \|\mathcal{R}_\Omega(Z)\|_\infty \le L$ and rank A = rank $M = r$ where $A = M([r_1], [r_2])$, then Theorem 5.6.3 states that

$$\left\|\varphi_{\text{Schur}}(Z) - M\right\|_\infty \le \frac{\sqrt{r_1 r_2}}{s_r(A)} \left\{ 4\sqrt{2}L + (1+\sqrt{5}) \frac{\sqrt{r_1 r_2}}{s_r(A)} L^2 \right\} \cdot \left\|\mathcal{R}_\Omega(E)\right\|_\infty,$$

for all E such that $\left\|\mathcal{R}_{[r_1]\times[r_2]}(E)\right\|_\infty \le \frac{s_r(A)}{2\sqrt{r_1 r_2}}$. Therefore, φ_{Schur} has the (c_{MC}, ϵ_{MC})-contraction property at M with respect to Ω with

$$c_{MC} = \frac{\sqrt{r_1 r_2}}{s_r(A)} \left\{ 4\sqrt{2}L + (1+\sqrt{5}) \frac{\sqrt{r_1 r_2}}{s_r(A)} L^2 \right\}, \quad \text{and} \quad \epsilon_{MC} = \frac{s_r(A)}{2\sqrt{r_1 r_2}}.$$

Note that we set $\epsilon_{MC} = \frac{s_r(A)}{2\sqrt{r_1 r_2}}$ because of the requirement $\left\|\mathcal{R}_{[r_1]\times[r_2]}(E)\right\|_\infty \le \frac{s_r\left(\mathcal{R}_{[r_1]\times[r_2]}(M)\right)}{2\sqrt{r_1 r_2}}$, and that choice is not necessary (i.e., $\epsilon_{MC} \to \infty$) if this requirement can be certified otherwise.

Satisfaction of Assumption 5.5 Finally, we combine the Schur-complement-based method discussed above with the notion of anchor sets to argue how Assumption 5.5 is satisfied. Suppose that a pair of anchor sets $(\mathcal{S}^\sharp, \mathcal{A}^\sharp)$ for Q^* is available to the agent.

At each iteration t, we augment the discretized state/action sets with the anchor sets, namely, we let $\bar{\mathcal{S}}^{(t)} \leftarrow \mathcal{S}^{(t)} \cup \mathcal{S}^\sharp$ and $\bar{\mathcal{A}}^{(t)} \leftarrow \mathcal{A}^{(t)} \cup \mathcal{A}^\sharp$. Then we choose the exploration set $\Omega^{(t)}$ to be $\Omega^{(t)} = \{(s,a) \in \bar{\mathcal{S}}^{(t)} \times \bar{\mathcal{A}}^{(t)} : s \in \mathcal{S}^\sharp \text{ or } a \in \mathcal{A}^\sharp\}$. For each $(s,a) \in \Omega^{(t)}$, we update $\hat{Q}^{(t)}(s,a)$ with the simulator as previously described in Step 2-(b) of Algorithm 3. Lastly, we estimate $\bar{Q}^{(t)}(\bar{\mathcal{S}}^{(t)}, \bar{\mathcal{A}}^{(t)})$ using φ_{Schur}. This entire process is summarized in Algorithm 4.

From the discussions above, we can verify that Assumption 5.5 is satisfied with

$$c_{MC} = \frac{\sqrt{r_1 r_2}}{s_r\big(Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp)\big)} \left\{ 4\sqrt{2} V_{\max} + (1+\sqrt{5}) \frac{\sqrt{r_1 r_2}}{s_r\big(Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp)\big)} V_{\max}^2 \right\}, \quad \epsilon_{MC} = \frac{s_r\big(Q^*(\mathcal{S}^\sharp, \mathcal{A}^\sharp)\big)}{2\sqrt{r_1 r_2}}.$$

Algorithm 4: An implementation of Steps 2 and 3 from Algorithm 3 (Schur-MC)

Input: $\mathcal{S}^{(t)}, \mathcal{A}^{(t)}$, $(\mathcal{S}^\sharp, \mathcal{A}^\sharp)$, and simulator (generative model)

Output: $\bar{Q}^{(t)}(\mathcal{S}^{(t)}, \mathcal{A}^{(t)}) \in \mathbb{R}^{|\mathcal{S}^{(t)}| \times |\mathcal{A}^{(t)}|}$

1. At the beginning of Step 2 of Algorithm 3, augment the discretized state/action sets $\mathcal{S}^{(t)}, \mathcal{A}^{(t)}$ with the anchor set by letting $\bar{\mathcal{S}}^{(t)} \leftarrow \mathcal{S}^{(t)} \cup \mathcal{S}^\sharp$ and $\bar{\mathcal{A}}^{(t)} \leftarrow \mathcal{A}^{(t)} \cup \mathcal{A}^\sharp$.
2. At Step 2-(a), choose $\Omega^{(t)} = \{(s,a) \in \bar{\mathcal{S}}^{(t)} \times \bar{\mathcal{A}}^{(t)} : s \in \mathcal{S}^\sharp \text{ or } a \in \mathcal{A}^\sharp\}$.
3. At Step 2-(b), update $\hat{Q}^{(t)}(s,a)$ with the simulator for all $(s,a) \in \Omega^{(t)}$.
4. At Step 3, estimate the $|\bar{\mathcal{S}}^{(t)}| \times |\bar{\mathcal{A}}^{(t)}|$ matrix $\bar{Q}^{(t)}(\bar{\mathcal{S}}^{(t)}, \bar{\mathcal{A}}^{(t)})$ by letting

$$\bar{Q}^{(t)}(\bar{\mathcal{S}}^{(t)}, \bar{\mathcal{A}}^{(t)}) \leftarrow \varphi_{\text{Schur}}\left(\begin{bmatrix} \hat{Q}^{(t)}(\mathcal{S}^\sharp, \mathcal{A}^\sharp) & \hat{Q}^{(t)}(\mathcal{S}^\sharp, \bar{\mathcal{A}}^{(t)}) \\ \hat{Q}^{(t)}(\bar{\mathcal{S}}^{(t)}, \mathcal{A}^\sharp) & * \end{bmatrix} \right). \tag{5.7}$$

Then obtain $\bar{Q}^{(t)}(\mathcal{S}^{(t)}, \mathcal{A}^{(t)})$ by restricting $\bar{Q}^{(t)}(\bar{\mathcal{S}}^{(t)}, \bar{\mathcal{A}}^{(t)})$ to $(s,a) \in \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$.

■ 5.7 Numerical Experiments

To complement our theoretical analysis in the chapter, we validate the effectiveness of Algorithm 3 with simple numerical experiments on the task of balancing an inverted pendulum. Additional experiments on a few other tasks – mountain car, double integrator, cart-pole, acrobot – can be found in [127].

5.7.1 Experimental Setup

Inverted Pendulum The aim of this control task is to balance a pendulum at the unstable equilibrium position, i.e., the upright position ($\theta = 0$). The dynamics of the pendulum system can be described by a 2-dimensional state, $(\theta, \dot{\theta})$ – the angle and the angular speed. Specifically, its equations of motion can be formulated as follows [144]:

$$\begin{aligned}\theta &:= \theta + \dot{\theta}\,\tau,\\ \dot{\theta} &:= \dot{\theta} + \Big(\sin\theta - \dot{\theta} + u\Big)\,\tau + \eta,\end{aligned}$$

where τ is the time interval between decisions, u denotes the input torque (control, or action) on the pendulum, and $\eta \sim \mathcal{N}(\mu, \sigma^2)$ refers to the random noise term modeling external disturbance. We use the following reward function to stabilize the pendulum:

$$R(\theta, u) = -0.1u^2 + \exp\left(\cos\theta - 1\right).$$

Experimental Setup In our simulations, we limit the input torque u in $[-1, 1]$ and set $\tau = 0.3$, $\mu = 0$, and $\sigma = 0.1$. First of all, we compute the 'ground truth' Q^* by running the standard value iteration on a very fine grid on the state-action space. Specifically, we discretize each dimension of the 2-dimensional state space into 50 values, and the action space into 1000 values, which results in a grid of size 2500×1000.

We observe that this 'ground truth' Q^*-matrix has a very small approximate rank in the inverted pendulum task (as well as in the other 4 tasks in [127]). Thus, we set $r_1 = r_2 = 10$ for our experiments, and select r_1 states and r_2 actions that are randomly sampled from each of the uniformly divided cells in the state/action spaces.

While our theory (Theorem 5.5.2) requires γ to be small, we find the method works well in our experiments even when γ is large. Here we report our empirical results for $\gamma = 0.9$.

Lastly, when we compare the performance of different matrix completion oracles, we use different sampling schemes (i.e., choice of $\Omega^{(t)}$). While traditional matrix completion methods typically work with samples drawn uniformly at random, our matrix completion method φ_{Schur} – the Schur-complement-based method in (5.6) – explores a few entire rows and columns, as discussed in Section 5.6. Therefore, we let the traditional methods explore a set of randomly chosen entries, but restricting it to have the same size as the set φ_{Schur} explores.

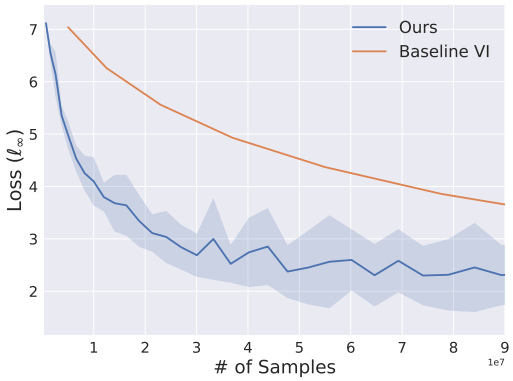
Measures of Performance We measure the performance of Q-learning with two measures:

- ℓ_∞ error: $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\hat{Q}(s,a) - Q^*(s,a)\right|$, and
- Mean error (mean absolute error): $\frac{1}{|\mathcal{S}||\mathcal{A}|}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\hat{Q}(s,a) - Q^*(s,a)\right|$.

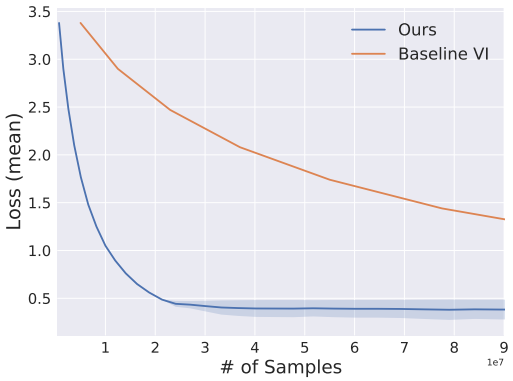
We use the fine grid and the 'ground truth' obtained by running the standard value iteration on it (discussed above) as proxies for $\mathcal{S} \times \mathcal{A}$ and Q^*, respectively.

■ 5.7.2 Simulation Results

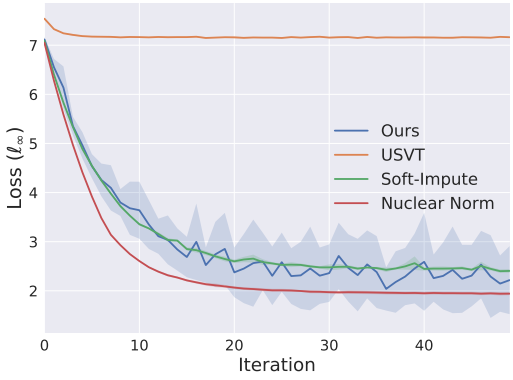
Improved Sample Complexity with MC In the first experiment, we evaluate the performance of our proposed Q-learning algorithm with aid of matrix completion (Algorithm 3) versus the same algorithm implemented without the matrix completion step (Step 3). In other words, the baseline algorithm explores and updates all $(s, a) \in \mathcal{S}^{(t)} \times \mathcal{A}^{(t)}$ at each iteration, which is equivalent to performing the standard simulated value iteration on the entirety of the discretized set of state-action pairs. The resulting errors for the two algorithms are plotted against the number of samples explored (=the number of calls to the simulator) in Figure 5.3a (ℓ_∞-error) and Figure 5.3b (mean error). It is evident from the plots that the sample complexity of Q-learning is significantly reduced with the use of matrix completion. Similar patterns are observed in the other 4 tasks that are not included here; see [127].



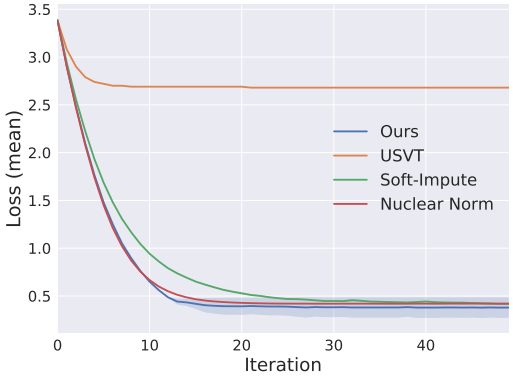
(a) Sample complexity (ℓ_∞ error).



(b) Sample complexity (mean error).

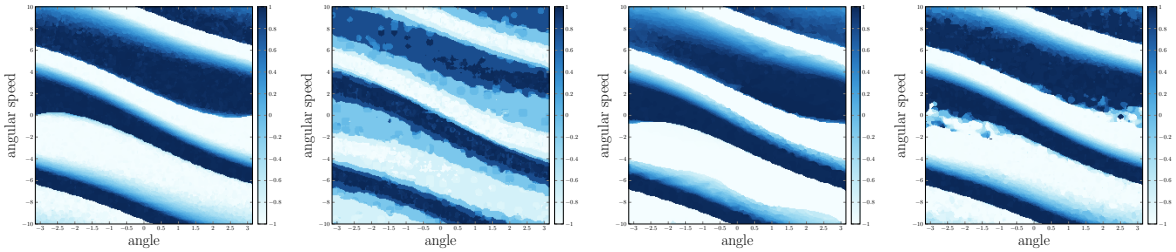


(c) Comparison of MC methods (ℓ_∞ error).



(d) Comparison of MC methods (Mean error).

Figure 5.3: Empirical results for the Inverted Pendulum task (averaged over 5 runs).



(a) Optimal policy. (b) Soft-impute. (c) Nuc-norm min. (d) Schur-MC.

Figure 5.4: Visualization of the derived policy for different matrix completion methods. The policy $\hat{\pi}$ is derived from $Q^{(T)}$ by taking $\hat{\pi}(s) = \arg\max_{a \in \mathcal{A}} Q^{(T)}(s, a)$.

Performance of the Schur-MC Next, we compare the performance of Algorithm 3 implemented with Schur-MC (Algorithm 4) to Algorithm 3 implemented with three different matrix completion methods, namely, nuclearn-norm minimization [32], soft-impute [100], and universal singular value thresholding [37]. These methods represent the common approaches to matrix completion – convex relaxation, iterative SVT, and one-shot SVT, respectively – that have been extensively studied in the literature.

The resulting errors for each of the methods are plotted against the number of iterations in Figure 5.3c (ℓ_∞-error) and Figure 5.3d (mean error). In addition, the policy $\hat{\pi}$ induced[4] from $Q^{(T)}$ for each of the methods except the USVT, are illustrated in Figure 5.4 for comparison with the optimal policy (Figure 5.4a). We omitted the policy for the USVT as it did not lead to a meaningful policy, as the high errors observed in Figures 5.3c, 5.3d already suggest.

First of all, we remark that the traditional matrix completion methods perform well in practice, although their theoretical analysis still remains insufficient to yield provable guarantees for their use in Q learning, or more broadly in reinforcement learning. In particular, the nuclear-norm minimization demonstrates the best overall performance among the methods tested, in terms of estimation accuracy and stability.

We observe that the Schur-MC exhibits competitive performance both in ℓ_∞ error and in mean error, nearly as good as the nuclear-norm minimization. Also, its resulting policy is very close to the policy induced by Q^*, except some jitters at the points of discontinuity in Q^*. Lastly, we note that the Schur-MC is a simple method, and therefore, it runs much faster, when compared to other optimization-based matrix completion methods. For example, the runtime for a single run of matrix completion on the Inverted Pendulum discretized to a 2500×1000 matrix with $10 \cdot (2500 + 1000 - 10) = 34900$ samples was $1.9 \pm .6$ seconds (averaged over 5 runs) for the Schur-MC, whereas it took 76.3 ± 8.2 seconds for the nuclear norm minimization, and 41.5 ± 1.7 seconds for the soft-impute.

[4] $\hat{\pi}(s) = \arg\max_{a \in \mathcal{A}} Q^{(T)}(s, a)$ for all s.

■ 5.8 Summary of the Chapter

In this chapter, we investigated how the low-rank structure in the Q^*-function can be exploited to provably reduce the sample complexity in model-free Q-learning. Unlike the existing works in the reinforcement learning literature that impose structural assumptions on the model dynamics, etc., we proposed to utilize the structure in the resulting Q-function. In that sense, our approach is more closely aligned with the spirit of nonparametric function estimation. Specifically, we described a Q-learning procedure that uses a matrix completion oracle to efficiently explore the set of state-action pairs. Thereafter, we proved that if the matrix completion oracle in use has a certain ℓ_∞-contraction property (Definition 5.5.1), then the proposed Q-learning procedure converges to the correct Q^*-function within ϵ-accuracy after exploring only $\mathrm{rank}(Q^*) \cdot \big(|\mathcal{S}| + |\mathcal{A}|\big)$ number of state-action pairs. In addition, we empirically verified our claims in this chapter with numerical simulations on some stochastic control tasks. An interesting problem for future research could be the verification of the ℓ_∞-contraction property for standard matrix completion methods such as the nuclear-norm minimization. Our experiments suggest that traditional optimization-based methods are likely to satisfy the property, but we do not know the answer yet and leave it as an open question.

■ 5.9 Proofs

■ 5.9.1 Proof of Theorem 5.5.2

Error Bound for the Lookahead Subroutine: Step 2-(b) in Algorithm 3 First of all, we present an upper bound for the error incurred by the lookahead (Bellman update with a simulator) based on the current oracle $V^{(t-1)}$, cf. Eq. (5.3) and Step 2-(b) of Algorithm 3.

Lemma 5.9.1. *Suppose that we have access to a value oracle $V : \mathcal{S} \to \mathbb{R}$ such that*

$$\sup_{s \in \mathcal{S}} \big| V(s) - V^*(s) \big| \le B.$$

Given $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $s'_1, \ldots, s'_N$ be the next states of (s, a) independently drawn from a simulator and let $\hat{Q}(s, a) = R(s, a) + \gamma \cdot \frac{1}{N} \sum_{i=1}^N V(s'_i)$. Then for any $\delta > 0$,

$$\big| \hat{Q}(s, a) - Q^*(s, a) \big| \le \gamma \left(B + \sqrt{\frac{2V_{\max}^2}{N} \log\left(\frac{2}{\delta}\right)} \right)$$

with probability at least $1 - \delta$.

Proof. Note that $Q^*(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P_{s,a}}[V^*(s')]$ by definition of Q^* and V^*; see (5.1).

It follows that

$$\begin{aligned}|\hat{Q}(s,a) - Q^*(s,a)| &= \gamma\left|\frac{1}{N}\sum_{i=1}^{N} V(s_i') - \mathbb{E}_{s'\sim P_{s,a}}[V^*(s')]\right| \\ &\leq \gamma\left|\frac{1}{N}\sum_{i=1}^{N} V(s_i') - \frac{1}{N}\sum_{i=1}^{N} V^*(s_i')\right| + \gamma\left|\frac{1}{N}\sum_{i=1}^{N} V^*(s_i') - \mathbb{E}_{s'\sim P_{sa}}\left[V^*(s')\right]\right| \\ &= \frac{\gamma}{N}\sum_{i=1}^{N}\left|V(s_i') - V^*(s_i')\right| + \gamma\left|\frac{1}{N}\sum_{i=1}^{N} V^*(s_i') - \mathbb{E}_{s'\sim P_{sa}}\left[V^*(s')\right]\right|. \quad (5.8)\end{aligned}$$

By assumption, the first term in Eq. (5.8) is bounded by γB. Meanwhile, we use Hoeffding's inequality (Theorem 2.4.18) to control the second term; for any $\tau > 0$,

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} V^*(s_i') - \mathbb{E}_{s'\sim P_{sa}}\left[V^*(s')\right] > \tau\right) \leq \exp\left(-\frac{N\tau^2}{2V_{\max}^2}\right).$$

Note that $\delta \leq 2\exp\left(-\frac{N\tau^2}{2V_{\max}^2}\right)$ if and only if $\tau \geq \sqrt{\frac{2V_{\max}^2}{N}\log\left(\frac{2}{\delta}\right)}$ to complete the proof. ■

Proof of Theorem 5.5.2

Proof of Theorem 5.5.2. We prove the first statement by mathematical induction. Observe that $Q^{(0)}(s,a) = 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, and thus, $|Q^{(0)}(s,a) - Q^*(s,a)| \leq V_{\max}$ for all (s,a). It remains to show that for $t = 1,\ldots,T$,

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{(t)}(s,a) - Q^*(s,a)\right| \leq 2\gamma c_{MC}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{(t-1)}(s,a) - Q^*(s,a)\right|. \quad (5.9)$$

To prove the inequality in (5.9), we trace back Steps 2 -4 in Algorithm 3.

- **(Step 4. interpolation)** For each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $\hat{s}^{(t)} \in \arg\min_{s'\in\mathcal{S}^{(t)}}\|s'-s\|_2$ and $\hat{a}^{(t)} \in \arg\min_{a'\in\mathcal{A}^{(t)}}\|a'-a\|_2$. Since $\mathcal{S}^{(t)}$ is a $\beta^{(t)}$-net of $\mathcal{S}$, $\|\hat{s}^{(t)} - s\| \leq \beta^{(t)}$. Likewise, $\|\hat{a}^{(t)} - a\| \leq \beta^{(t)}$. As $Q^{(t)}(s,a) = \bar{Q}^{(t)}(\hat{s}^{(t)},\hat{a}^{(t)})$ and Q^* is L-Lipschitz,

$$\begin{aligned}\left|Q^{(t)}(s,a) - Q^*(s,a)\right| &= \left|\bar{Q}^{(t)}(\hat{s}^{(t)},\hat{a}^{(t)}) - Q^*(s,a)\right| \\ &= \left|\bar{Q}^{(t)}(\hat{s}^{(t)},\hat{a}^{(t)}) - Q^*(\hat{s}^{(t)},\hat{a}^{(t)})\right| + \left|Q^*(\hat{s}^{(t)},\hat{a}^{(t)}) - Q^*(s,a)\right| \\ &\leq \left|\bar{Q}^{(t)}(\hat{s}^{(t)},\hat{a}^{(t)}) - Q^*(\hat{s}^{(t)},\hat{a}^{(t)})\right| + 2L\beta^{(t)}.\end{aligned}$$

Therefore, we get the following upper bound:

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{(t)}(s,a) - Q^*(s,a)\right| \leq \max_{(s,a)\in\mathcal{S}^{(t)}\times\mathcal{A}^{(t)}}\left|\bar{Q}^{(t)}(s,a) - Q^*(s,a)\right| + 2L\beta^{(t)}. \quad (5.10)$$

- **(Step 3. matrix completion)** By Assumption 5.5, we have the following upper bound:

$$\max_{(s,a)\in\mathcal{S}^{(t)}\times\mathcal{A}^{(t)}} \left|\bar{Q}^{(t)}(s,a) - Q^*(s,a)\right| \le c_{MC} \max_{(s,a)\in\Omega^{(t)}} \left|\hat{Q}^{(t)}(s,a) - Q^*(s,a)\right|. \tag{5.11}$$

- **(Step 2. exploration)** Lastly, applying Lemma 5.9.1 and taking union bound over $(s,a)\in\Omega^{(t)}$, we obtain

$$\begin{aligned}&\max_{(s,a)\in\Omega^{(t)}} \left|\hat{Q}^{(t)}(s,a) - Q^*(s,a)\right| \\ &\quad\le \gamma\left(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t-1)}(s,a) - Q^*(s,a)\right| + \sqrt{\frac{2V_{\max}^2}{N^{(t)}}\log\left(\frac{2|\Omega^{(t)}|T}{\delta}\right)}\right)\end{aligned} \tag{5.12}$$

with probability at least $1-\frac{\delta}{T}$. Here we used the fact that $Q^{(t-1)}$ and Q^* are continuous, and $\mathcal{A}$ is compact[5], and thus,

$$\sup_{s\in\mathcal{S}} \left|V^{(t-1)}(s) - V^*(s)\right| \le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t-1)}(s,a) - Q^*(s,a)\right|.$$

Combining the inequalities in (5.10), (5.11), (5.12) yields

$$\begin{aligned}&\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t)}(s,a) - Q^*(s,a)\right| \\ &\quad\le \gamma c_{MC}\left(\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t-1)}(s,a) - Q^*(s,a)\right| + \sqrt{\frac{2V_{\max}^2}{N^{(t)}}\log\left(\frac{2|\Omega^{(t)}|T}{\delta}\right)}\right) + 2L\beta^{(t)}.\end{aligned} \tag{5.13}$$

with probability at least $1-\frac{\delta}{T}$. For each $1\le t\le T$, we choose

$$\beta^{(t)} = \frac{V_{\max}}{8L}(2\gamma c_{MC})^t \quad\text{and}\quad N^{(t)} = \frac{8}{(2\gamma c_{MC})^{2(t-1)}}\log\left(\frac{2|\Omega^{(t)}|T}{\delta}\right). \tag{5.14}$$

With this choice of $\beta^{(t)}$ and $N^{(t)}$, we can see from (5.13) that if $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t-1)}(s,a) - Q^*(s,a)\right| \le (2\gamma c_{MC})^{t-1}V_{\max}$, then $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t)}(s,a) - Q^*(s,a)\right| \le (2\gamma c_{MC})^t V_{\max}$ with probability at least $1-\frac{\delta}{T}$ for each $t\in[T]$.

[5]For each $s\in\mathcal{S}$, there exist $a^{(t-1)}(s), a^*(s)\in\mathcal{A}$ such that $V^{(t-1)}(s) = Q^{(t-1)}(s,a^{(t-1)}(s))$ and $V^*(s) = Q^*(s,a^*(s))$. If $V^{(t-1)}(s)\ge V^*(s)$, then $V^{(t-1)}(s) - V^*(s) = Q^{(t-1)}(s,a^{(t-1)}(s)) - Q^*(s,a^*(s)) \le Q^{(t-1)}(s,a^{(t-1)}(s)) - Q^*(s,a^{(t-1)}(s))$. If $V^{(t-1)}(s) < V^*(s)$, then $V^*(s) - V^{(t-1)}(s) = Q^*(s,a^*(s)) - Q^{(t-1)}(s,a^{(t-1)}(s)) \le Q^*(s,a^*(s)) - Q^{(t-1)}(s,a^*(s))$. Therefore, $|V^{(t-1)}(s) - V^*(s)| \le \max_{a\in\{a^{(t-1)}(s),a^*(s)\}}\{Q^{(t-1)}(s,a) - Q^*(s,a)\}$.

Taking union bound over $t \in [T]$, we can conclude that with probability at least $1-\delta$,

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q^{(t)}(s,a) - Q^*(s,a)\right| \le (2\gamma c_{MC})^t V_{\max}, \quad \forall t = 1,\dots,T.$$

Sample complexity. If $\gamma < \frac{1}{2c_{MC}}$, then $2\gamma c_{MC} < 1$. Let $T_\epsilon = \left\lceil \frac{\log\left(\frac{V_{\max}}{\epsilon}\right)}{\log\left(\frac{1}{2\gamma c_{MC}}\right)} \right\rceil$ and observe that $(2\gamma c_{MC})\cdot\epsilon \le (2\gamma c_{MC})^{T_\epsilon} V_{\max} \le \epsilon$. Recall that for each $t \in [T]$, we call the simulator $N^{(t)}$ times for each state-action pair $(s,a) \in \Omega^{(t)}$. Therefore, the total sample complexity of Algorithm 3 with $T = T_\epsilon$ is $\sum_{t=1}^{T_\epsilon} \left|\Omega^{(t)}\right| N^{(t)}$.

By standard argument on covering number (e.g. Section 2.2.3), we can see that $|\mathcal{S}^{(t)}|, |\mathcal{A}^{(t)}| \le C'\left(\frac{1}{\beta^{(t)}}\right)^d = C'\left(\frac{8L}{V_{\max}}\right)^d \left(2\gamma c_{MC}\right)^{-dt}$ for some constant $C' > 0$. This is an increasing function of t. Recall that we may assume $|\Omega^{(t)}| = C_\Omega\left(|\mathcal{S}^{(t)}| + |\mathcal{A}^{(t)}|\right)$ by Assumption 5.5. This is also increasing with respect to t, and thus, so is $N^{(t)}$ as described in (5.14). Therefore,

$$\sum_{t=1}^{T_\epsilon} \left|\Omega^{(t)}\right| N^{(t)} \le T_\epsilon \left|\Omega^{(T_\epsilon)}\right| N^{(T_\epsilon)}.$$

Recall that $(2\gamma c_{MC})^{T_\epsilon} V_{\max} \ge (2\gamma c_{MC})\cdot\epsilon$. Thus, $\beta^{(T_\epsilon)} = \frac{V_{\max}}{8L}(2\gamma c_{MC})^{T_\epsilon} \ge \frac{2\gamma c_{MC}}{8L}\epsilon$, and $|\mathcal{S}^{(T_\epsilon)}|, |\mathcal{A}^{(T_\epsilon)}| \le C'\left(\frac{8L}{2\gamma c_{MC}}\right)^d \frac{1}{\epsilon^d}$. As a result, we can observe that

$$\begin{aligned}
T_\epsilon \left|\Omega^{(T_\epsilon)}\right| N^{(T_\epsilon)} &\le T_\epsilon \cdot C_\Omega\left(|\mathcal{S}^{(T_\epsilon)}| + |\mathcal{A}^{(T_\epsilon)}|\right) \cdot \frac{8}{(2\gamma c_{MC})^{2(T_\epsilon-1)}} \log\left(\frac{2C_\Omega\left(|\mathcal{S}^{(T_\epsilon)}| + |\mathcal{A}^{(T_\epsilon)}|\right)T_\epsilon}{\delta}\right) \\
&\le T_\epsilon \cdot 2C_\Omega C' \left(\frac{8L}{2\gamma c_{MC}}\right)^d \frac{1}{\epsilon^d} \cdot 8\left(\frac{V_{\max}}{\epsilon}\right)^2 \log\left(\frac{4C_\Omega C' T_\epsilon}{\delta}\left(\frac{8L}{2\gamma c_{MC}}\right)^d \frac{1}{\epsilon^d}\right) \\
&= 16 C_\Omega C' V_{\max}^2 \left(\frac{8L}{2\gamma c_{MC}}\right)^d \cdot \frac{T_\epsilon}{\epsilon^{d+2}} \cdot \log\left(4C_\Omega C'\left(\frac{8L}{2\gamma c_{MC}}\right)^d \cdot \frac{T_\epsilon}{\epsilon^d} \cdot \frac{1}{\delta}\right). \qquad (5.15)
\end{aligned}$$

To conclude the proof, it suffices to observe that $T_\epsilon = \left\lceil \frac{\log\left(\frac{V_{\max}}{\epsilon}\right)}{\log\left(\frac{1}{2\gamma c_{MC}}\right)} \right\rceil = O\left(\log\frac{1}{\epsilon}\right)$, and therefore, the total sample complexity scales as $O\left(\frac{1}{\epsilon^{d+2}} \log\frac{1}{\epsilon} \cdot \left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$ by (5.15). ■

■ 5.9.2 Proof of Theorem 5.6.3

Proof. First of all, observe that $D = CA^\dagger B$ by Lemma 5.6.2. Then we can decompose $\widetilde{C}\widetilde{A}^\dagger\widetilde{B} - D = \widetilde{C}\left(\widetilde{A}^\dagger - A^\dagger\right)\widetilde{B} + \left(\widetilde{C}A^\dagger\widetilde{B} - D\right)$, and therefore, for any $(i,j) \in [n_1]\times[n_2]$,

$$\begin{aligned}
\left(\widetilde{C}\widetilde{A}^\dagger\widetilde{B} - D\right)_{ij} &= e_i^T\left(\widetilde{C}\widetilde{A}^\dagger\widetilde{B} - D\right)e_j = \mathrm{Tr}\left[e_i^T\widetilde{C}\left(\widetilde{A}^\dagger - A^\dagger\right)\widetilde{B}e_j\right] + \mathrm{Tr}\left[e_i^T\left(\widetilde{C}A^\dagger\widetilde{B} - D\right)e_j\right] \\
&= \mathrm{Tr}\left[\left(\widetilde{A}^\dagger - A^\dagger\right)\widetilde{B}e_j e_i^T\widetilde{C}\right] + \mathrm{Tr}\left[A^\dagger\left(\widetilde{B}e_j e_i^T\widetilde{C} - Be_j e_i^T C\right)\right].
\end{aligned}$$

Since $\big|\mathrm{Tr}(XY)\big| \le |||X|||_2 \|Y\|_* \le \sqrt{\mathrm{rank}\, Y}\,|||X|||_2 \|Y\|_F$, we obtain

$$\begin{aligned}\big\|\widetilde{C}\widetilde{A}^\dagger\widetilde{B} - D\big\|_\infty &\le \max_{i,j}\Big|\mathrm{Tr}\Big[\big(\widetilde{A}^\dagger - A^\dagger\big)\widetilde{B}e_j e_i^T\widetilde{C}\Big]\Big| + \max_{i,j}\Big|\mathrm{Tr}\Big[A^\dagger\big(\widetilde{B}e_j e_i^T\widetilde{C} - Be_j e_i^T C\big)\Big]\Big| \\ &\le |||\widetilde{A}^\dagger - A^\dagger|||_2 \cdot \max_{i,j}\big\|\widetilde{B}e_j e_i^T\widetilde{C}\big\|_F + \sqrt{2}|||A^\dagger|||_2 \cdot \max_{i,j}\big\|\widetilde{B}e_j e_i^T\widetilde{C} - Be_j e_i^T C\big\|_F.\end{aligned}$$

To conclude the proof, it suffices to observe that

$$\begin{aligned}|||A^\dagger|||_2 &\le \frac{1}{s_r(A)}, \\ |||A^\dagger|||_2 &\le \frac{1}{s_r(A) - \epsilon_A\sqrt{r_1 r_2}}, && \because \text{Weyl's ineq. (Thm. 2.2.4)} \\ |||\widetilde{A}^\dagger - A^\dagger|||_2 &\le \frac{1+\sqrt{5}}{2}|||A^\dagger|||_2|||A^\dagger|||_2|||\widetilde{A} - A|||_2 && \because \text{[138, Thm. 3.4]} \\ &\le \frac{1+\sqrt{5}}{2}\frac{\epsilon_A\sqrt{r_1 r_2}}{s_r(A)\cdot\big(s_r(A) - \epsilon_A\sqrt{r_1 r_2}\big)} \\ &\le (1+\sqrt{5})\frac{\epsilon_A\sqrt{r_1 r_2}}{s_r(A)^2}, && \because \epsilon_A \le \frac{s_r(A)}{2\sqrt{r_1 r_2}} \\ \big\|Be_j e_i^T C\big\|_F &\le \sqrt{r_1 r_2}\|B\|_\infty\|C\|_\infty, \\ \big\|\widetilde{B}e_j e_i^T\widetilde{C} - Be_j e_i^T C\big\|_F &\le \sqrt{r_1 r_2}\big(\|B\|_\infty\epsilon_C + \|C\|_\infty\epsilon_B + \epsilon_B\epsilon_C\big).\end{aligned}$$

■

Part III

Hardness of Global Approximation of a Large-scale SDP

Chapter 6

Semidefinite Programming, Scalability, and Approximating the PSD Cone

6.1 Introduction to Part III

6.1.1 Semidefinite Programming and Scalability Issue

Semidefinite programming (SDP) is a branch of convex optimization that considers problems of the form

$$\begin{aligned} \text{minimize} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m, \\ & X \in \boldsymbol{S}_+^n, \end{aligned} \tag{6.1}$$

where C, A_i's are $n \times n$ symmetric matrices, $\langle A, B \rangle = \mathrm{Tr}(AB)$, and $\boldsymbol{S}_+^n$ denotes the cone of $n \times n$ positive semidefinite (PSD) matrices. For every SDP of the form (6.1) (called the primal problem), there exists another associated SDP, called the dual problem, which can be stated as

$$\begin{aligned} \text{maximize} \quad & \langle b, y \rangle \\ \text{subject to} \quad & C - \sum_{i=1}^m y_i A_i \in \boldsymbol{S}_+^n, \end{aligned} \tag{6.2}$$

where $y = (y_1, \dots, y_m) \in \mathbb{R}^m$ is the dual decision variable.

SDP has attracted great interest in many fields as a powerful tool to provide theoretical guarantees as well as practical algorithms. For example, the current best approximation algorithms for hard combinatorial problems such as the maximum cut [62] and the maximal stable set [97] rely on semidefinite programming. SDPs also arise as computationally tractable relaxations for sparse recovery problems, e.g., rank minimization [116] and sparse PCA [41]. Moreover, the hierarchies of SDPs based on the notion of sum-of-squares (SOS) polynomials [86, 112] provide one of the most promising approaches for global optimization without convexity assumptions. For instance, the SOS approaches can be used to automate the search of a Lyapunov function that certifies various notions related to the safety, stability, and ro-

bustness in control and robotics [111]; and more recently, they have been extensively studied in the context of algorithmic robust statistics [85, 47, 36] because they can yield algorithms that achieve a near-optimal tradeoff between statistical and computational efficiencies.

In addition to the long list of applications, SDPs are also well known for their strong theoretical and computational properties. For instance, under mild assumptions[1] – e.g., if both primal and dual problems are strictly feasible – strong duality holds between the primal problem (6.1) and the dual problem (6.2), i.e., they achieve exactly the same optimal costs. Moreover, many algorithms for linear programs have been adapted to SDPs, and these algorithms can solve SDPs to arbitrary accuracy in polynomial time [155].

Despite these numerous virtues of SDPs, when it comes to solving SDPs for practical purposes, they suffer from one serious problem, which is scalability. Although current SDP solvers utilizing interior-point methods can solve an SDP up to arbitrary accuracy, they require prohibitively large computational cost and memory requirement when n is large. Indeed, such scalability issues remain as one of the major challenges in the field, and devising methods to address these impediments to make SDPs more scalable is an active area of research.

■ 6.1.2 Approximating the PSD Cone for Scalable SDP

Among the various attempts that have been made to improve the scalability of the SDPs (see the recent survey [99] and references therein), we study in this thesis the approaches that replace the PSD cone, $\boldsymbol{S}_+^n$, in the SDP formulation with another cone $\mathcal{K}$ that is computationally simpler to describe. More precisely, these approaches substitute the PSD constraint $X \in \boldsymbol{S}_+^n$ in the SDP (6.1) with a computationally easier cone membership constraint $X \in \mathcal{K}$ for some convex cone, $\mathcal{K} \subseteq \boldsymbol{S}_+^n$, thereby obtaining the following approximate optimization formulation:

$$\begin{aligned} \text{minimize} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m, \\ & X \in \mathcal{K}. \end{aligned} \tag{6.3}$$

Note that the optimal solution of the approximate program (6.3) is guaranteed to be feasible to the original SDP (6.1) because $\mathcal{K} \subseteq \boldsymbol{S}_+^n$. This property is particularly advantageous in safety-critical applications.

Notable examples of these approaches include the diagonally dominant sum-of-squares (DSOS) and scaled diagonally dominant sum-of-squares (SDSOS) optimization that are recently introduced by Ahmadi and Majumdar [6]. DSOS optimization considers the cone of diagonally dominant matrices, namely, $DD^n := \{A \in \boldsymbol{S}^n : a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \forall i \in [n]\}$ in

[1] usually referred to as constraint qualifications

place of $\boldsymbol{S}_+^n$. Note that DD^n is a polyhedral cone and that $DD^n \subseteq \boldsymbol{S}_+^n$, (e.g., by Gershgorin's circle theorem). As a result, DSOS optimization provides a linear-programming-based alternative to SDP by letting $\mathcal{K} = DD^n$ in (6.3), which trades off solution quality for reduced computation time. Similarly, SDSOS optimization considers the second-order cone, $SDD^n := \{A \in \boldsymbol{S}^n : \exists \rho \in \mathbb{R}_+^n \text{ such that } \mathrm{diag}(\rho) \cdot A \cdot \mathrm{diag}(\rho) \in DD^n\}$, and leads to a second-order-cone-programming-based approximation of SDP.

As these approaches trade off scalability with conservatism, there arises a tradeoff between the computational gain and the quality of optimal solution. Informally, the simpler the facial structure of $\mathcal{K}$ is, the faster the problem (6.3) could be solved, however, the resulting approximate optimal value could be farther away from the true optimal value of the original SDP (6.1); see Figure 6.1. As a result, there arises a natural question: "how much error is incurred in the optimal value when $\boldsymbol{S}_+^n$ is replaced by $\mathcal{K}$, or equivalently, how large is the optimality gap between the original problem (6.1) and the approximate problem (6.3)?" In spite of the promising empirical evidences that DSOS and SDSOS approaches can solve SDPs much faster than the standard methods without sacrificing the quality of the optimal solutions [6, Section 4], there is little theoretical understanding about the tradeoff.

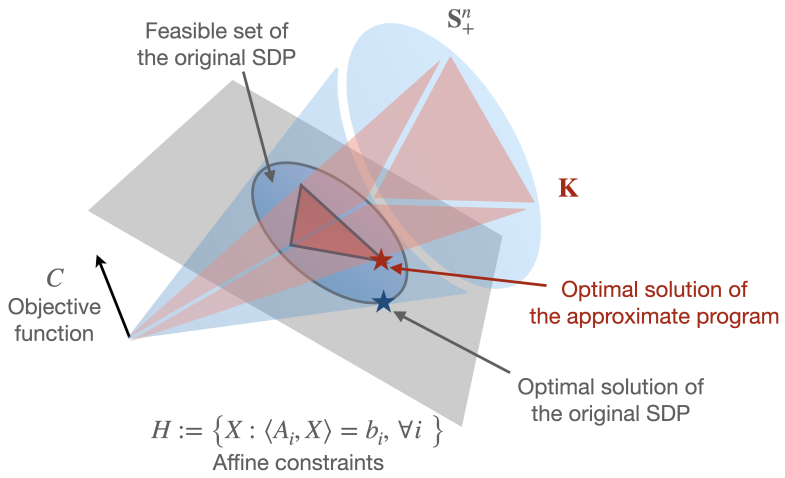


Figure 6.1: An illustration of approximating the SDP (6.1) by the program (6.3).

Motivated by this lack of theoretical understanding, we study the problem of approximating the PSD cone, $\boldsymbol{S}_+^n$, with a cone $\mathcal{K}$ that can be described by a small number of $k \times k$ PSD constraints. Note that DSOS corresponds to the special case with $k = 1$, and SDSOS corresponds to $k = 2$. Specifically, we ask the following question in this part of the thesis:

> *"How closely can we approximate $\boldsymbol{S}_+^n$ with a cone $\mathcal{K}$ that can be described using at most N number of $k \times k$ PSD constraints?"*

Intuitively, if the cone $\mathcal{K}$ closely approximates $\boldsymbol{S}_+^n$, then the optimality gap incurred by approximating (6.1) with (6.3) will be small. The notion of *approximation* and the measure of *complexity* (N) will be specified in Chapter 7 to make this question precise.

6.1.3 Contributions and Organization of Part III

Contributions As the main contributions in this part, we show the hardness of approximating the PSD cone using only a few, small-sized PSD constraints. To this end, in Chapter 7, we first specify the notions of approximation for cones (Section 7.2) that we use, and show that it is impossible to approximate $\boldsymbol{S}_+^n$ within a constant factor using only a polynomially many number of $k \times k$ PSD constraints, when $k \ll \frac{n}{\log^2 n}$ (see the theorems in Section 7.4 for the precise statements). We prove this claim by showing that *any* set that well approximates $\boldsymbol{S}_+^n$ must have superpolynomially large $\boldsymbol{S}_+^k$-extension complexity (to be defined in Section 6.2.1). We remark that this extension complexity lower bound implies the hardness of approximating $\boldsymbol{S}_+^n$, regardless of the construction of approximating sets. This is a strong implication as it rules out the possibility of approximating $\boldsymbol{S}_+^n$ for all construction schemes at once.

Here we clarify that our results only refute the *global, non-adaptive* approximability of the PSD cone $\boldsymbol{S}_+^n$ using a small number of $k \times k$ PSD constraints. Our hardness theorems do not contradict the empirical success of DSOS/SDSOS approaches because the optimal value of the approximate program (6.3) can still be close to that of the original SDP (6.1) as long as $\mathcal{K}$ *locally* approximates $\boldsymbol{S}_+^n$, e.g., in the direction of C.

Organization The rest of this part is organized as follows. In the remainder of this chapter (Chapter 6), we review additional technical background that will be used in our analysis in Chapter 7. Specifically, we review the concept of semidefinite extension complexity and some results about Fourier analysis on the Boolean hypercube. Thereafter, Chapter 7 discusses the main results of this part; see Section 7.1.2 for the organization of Chapter 7.

6.2 Additional Technical Background for Chapter 7

In this section, we review additional background materials in preparation for Chapter 7; these will be used in our proof of the main theorems stated in Section 7.4. Specifically, in Section 6.2.1, we briefly review the concepts of semidefinite extension complexity and slack operators, and Section 6.2.2 contains some results about Fourier analysis on the Boolean hypercube. Expert readers may want to skip this chapter and continue reading from Chapter 7.

■ 6.2.1 Lifts, Extension Complexity and Slack Operator

Here we briefly review the $\mathcal{K}$-extension complexity of a convex body and its connection to the $\mathcal{K}$-rank of a slack operator. We refer interested readers to [66] and [58] for more details.

Let $\mathcal{K}$ be a closed convex cone. Given a positive integer r, let $\mathcal{K}^r = \mathcal{K} \times \cdots \times \mathcal{K}$ (r times) denote the Cartesian product of r copies of $\mathcal{K}$. We say that a set $S \subset \mathbb{R}^d$ admits a $\mathcal{K}^r$-lift if S can be expressed as

$$S = \pi\big(\mathcal{K}^r \cap L\big)$$

where π is a linear map and L is an affine subspace. The convex set $\mathcal{K}^r \cap L$ is called a $\mathcal{K}^r$-lift of S. The $\mathcal{K}$-extension complexity of S, denoted by $\mathrm{xc}_{\mathcal{K}}(S)$, is defined as the smallest r such that S admits a $\mathcal{K}^r$-lift. We remark that if $\mathrm{xc}_{\mathcal{K}}(S) = N$, then one can optimize a linear function on S by solving a conic program involving N decision variables in $\mathcal{K}$.

Let P, Q be two convex bodies such that $P \subseteq Q \subseteq \mathbb{R}^d$ and the origin is contained in the interior of P. Let Q° be the polar of Q; see (2.4). We let $\mathrm{ext}(P)$ denote the set of extreme points of P and define the slack operator $s_{P,Q}$ for the pair (P, Q) as follows.

Definition 6.2.1 (slack operator)**.** For a pair of convex bodies $P \subseteq Q$ with 0 in the interior of P, the map $s_{P,Q} : \mathrm{ext}(P) \times \mathrm{ext}(Q^\circ) \to \mathbb{R}$ such that $s_{P,Q}(x, y) = 1 - \langle x, y\rangle$ is called its associated slack operator. The slack operator $s_{P,Q}$ admits a $\mathcal{K}$-factorization if there exists a pair of maps $A : \mathrm{ext}(P) \to \mathcal{K}$ and $B : \mathrm{ext}(Q^\circ) \to \mathcal{K}^*$ such that $s_{P,Q}(x, y) = \langle A(x), B(y)\rangle$ for all $x \in \mathrm{ext}(P)$ and $y \in \mathrm{ext}(Q^\circ)$.

Note that $s_{P,Q}(x, y) \geq 0$ for all $(x, y) \in \mathrm{ext}(P) \times \mathrm{ext}(Q^\circ)$ because $P \subseteq Q$ and therefore $\langle x, y\rangle \leq 1$ for all $(x, y) \in P \times Q^\circ$ by definition of the polar.

The existence of a $\mathcal{K}$-lift of a convex body S is closely related to that of a $\mathcal{K}$-factorization of $s_{P,Q}$ associated to convex bodies P, Q such that $P \subseteq S \subseteq Q$. This connection is originally established by Yannakakis [165] for the special case with $\mathcal{K} = \mathbb{R}_+$, motivated by computational considerations about linear programming (LP). This special case of the $\mathbb{R}_+$-extension complexity is widely known as the LP extension complexity (or the extension complexity of polytopes), which counts the minimum number of linear inequalities required to describe S. If $\mathrm{xc}_{\mathbb{R}_+}(S) = N$, then one can optimize a linear function on S by solving a linear program with N inequality constraints.

Note that a polytope is generated by a finite number of extreme points, or by a finite number of facets from the dual perspective. Thus, the slack operator associated to a pair of polytopes is a finite-sized nonnegative matrix (so called the slack matrix). Yannakakis' theorem states that the LP extension complexity of a polytope is equal to the nonnegative rank of an associated slack matrix.

The Yannakakis' theorem is later generalized in [66]. Here we state a generalized version

of Yannakakis theorem in the next lemma (cf. [58, Proposition 3.12]), which immediately follows from the proof of [66, Theorem 3].

Lemma 6.2.2. *Let P, Q be a pair of convex bodies such that $P \subseteq Q$. If there is a convex body S such that S admits a proper $\mathcal{K}$-lift and $P \subseteq S \subseteq Q$, then $s_{P,Q}$ has a $\mathcal{K}$-factorization. Conversely, if $s_{P,Q}$ admits a $\mathcal{K}$-factorization, then there exists a convex set S such that S has a $\mathcal{K}$-lift and $P \subseteq S \subseteq Q$.*

We are interested in the case where $\mathcal{K}$ is a Cartesian product of small PSD cones, $\boldsymbol{S}_+^k$ where $k \geq 1$ is a fixed constant. Given a nonnegative operator s, we define $\operatorname{rank}_{\boldsymbol{S}_+^k}(s)$ to be the smallest positive integer r such that s admits a $(\boldsymbol{S}_+^k)^r$-factorization. As a consequence of Lemma 6.2.2, we obtain

$$\inf_{S: P \subseteq S \subseteq Q} \operatorname{xc}_{\boldsymbol{S}_+^k}(S) = \operatorname{rank}_{\boldsymbol{S}_+^k}(s_{P,Q}). \tag{6.4}$$

That is, the $\boldsymbol{S}_+^k$-rank of the slack operator $s_{P,Q}$ provides a lower bound on the $\boldsymbol{S}_+^k$-extension complexity that is valid for *all* convex set S such that $P \subseteq S \subseteq Q$.

■ 6.2.2 Fourier Analysis on the Hypercube

Later in the proof of Theorems 7.4.1 and 7.4.2, we consider a certain slack operator restricted to the n-dimensional hypercube and use its degree-2 Fourier component to prove extension complexity lower bounds. Specifically, we will need to control the norm of the degree-2 Fourier component. Here we review some necessary notions and refer the interested readers to a more comprehensive reference, e.g., [108].

Let $H_n = \{-1, 1\}^n$ denote the vertex set of the n-dimensional hypercube. Every function $f : H_n \to \mathbb{R}$ has a unique Fourier expansion

$$f = f_0 + f_1 + \cdots + f_n$$

where each f_k is a homogeneous multilinear polynomial of degree k. We call f_k the k-th harmonic component of f and let $\operatorname{proj}_k : f \mapsto f_k$ denote the projection onto the degree-k harmonic subspace (the subspace of homogeneous polynomials of degree k).

Given $\rho \in [0, 1]$, the noise operator T_ρ smooths $f : H_n \to \mathbb{R}$, by diminishing its high-frequency modes. To be precise, T_ρ acts on f multiplying the k-th Fourier coefficient of f by a factor of ρ^k, i.e.,

$$T_\rho f = \sum_{k=0}^{n} \rho^k f_k.$$

For $\rho < 1$, $T_\rho f$ is 'smoother' than f as the high-frequency terms of f are attenuated more heavily. In one extreme, $T_\rho f$ is constant equal to $\mathbb{E} f$ when $\rho = 0$; in the other extreme where

$\rho = 1$, there is no smoothing at all and $T_\rho f = f$.

Next, we recall that the p-norm ($p \geq 1$) of $f : H_n \to \mathbb{R}$ is defined as

$$\|f\|_p = \left(\mathbb{E}_{x \sim \mu(H_n)}\left[|f(x)|^p\right]\right)^{\frac{1}{p}}.$$

where $\mu(H_n)$ denotes the uniform probability measure over H_n. Note that $\|f\|_p \leq \|f\|_q$ for $p \leq q$. When $p < q$, there is no general way to control $\|f\|_q$ with $\|f\|_p$, and the ratio $\|f\|_q / \|f\|_p$ can be arbitrarily large; the ratio becomes larger as f fluctuates more wildly.

The hypercontractive inequality for T_ρ due to Bonami and Beckner [23, 16] provides an upper bound on $\|T_\rho f\|_q$ in terms of $\|f\|_p$ with $p < q$, thereby giving an estimate for how much smoother $T_\rho f$ is, when compared to f. It can be stated as follows.

Lemma 6.2.3 (Hypercontractivity)**.** *Given $f : H_n \to \mathbb{R}$, for any $0 < \rho \leq 1$ and $p \geq 1$, we have $\|T_\rho f\|_q \leq \|f\|_p$ where $q = 1 + \frac{1}{\rho^2}(p-1)$.*

Useful Lemmas about Degree-2 Harmonic Component We use Lemma 6.2.3 to control the norm of the degree-2 harmonic component of a bounded nonnegative function as stated below in Lemma 6.2.4, following [117, Lemma 2.3] and [56, Lemma 3].

Lemma 6.2.4. *Let $f : H_n \to \mathbb{R}$ satisfy (i) $0 \leq f(x) \leq \Lambda$ for all $x \in H_n$; and (ii) $\mathbb{E}_{x \sim \mu(H_n)}[f(x)] \leq 1$. Then*

$$\|\mathrm{proj}_2 f\|_2 \leq \begin{cases} \Lambda & \text{if } \Lambda < e, \\ e \log(\Lambda) & \text{if } \Lambda \geq e. \end{cases}$$

Proof. Let $f = f_0 + f_1 + f_2 + \cdots + f_n$ be the Fourier expansion of f. Then for $0 \leq \rho \leq 1$,

$$\|\mathrm{proj}_2 f\|_2^2 = \|f_2\|_2^2 = \frac{1}{\rho^4}\left(\rho^2 \|f_2\|_2\right)^2 \leq \frac{1}{\rho^4} \sum_{k=0}^{n} \rho^{2k} \|f_k\|_2^2 = \frac{1}{\rho^4} \|T_\rho f\|_2^2.$$

With $\rho = \sqrt{p-1}$ for $1 \leq p \leq 2$, we have $\|T_\rho f\|_2 \leq \|f\|_p$ by hypercontractivity. Then it follows that

$$\|\mathrm{proj}_2 f\|_2 \leq \frac{1}{\rho^2} \|T_\rho f\|_2 \leq \frac{1}{p-1} \|f\|_p \leq \frac{1}{p-1} \Lambda^{p-1}$$

because $\|f\|_p = \mathbb{E}[f^p]^{\frac{1}{p}} \leq \Lambda^{\frac{p-1}{p}} \mathbb{E}[f]^{\frac{1}{p}} \leq \Lambda^{\frac{p-1}{p}} \leq \Lambda^{p-1}$. If $\Lambda < e$, we choose $p = 2$ to get $\|\mathrm{proj}_2 f\|_2 \leq \Lambda$. Otherwise, we choose $p = 1 + \frac{1}{\log \Lambda}$ to obtain $\|\mathrm{proj}_2 f\|_2 \leq e \log(\Lambda)$. ∎

Observe that for any function $f : H_n \to \mathbb{R}$, its degree-2 projection, $\mathrm{proj}_2(f)$, is a multilinear quadratic form on H_n. That is, there exists some matrix A with zero diagonal[2] such that

[2] More precisely, $A_{ij} = \frac{1}{2}\mathbb{E}_{X \sim \mu(H_n)}[X_i X_j f(X)]$ for $i, j \in [n]$ such that $i \neq j$.

$\text{proj}_2(f)(x) = x^T Ax$ for all $x \in H_n$. Therefore, the random variable $\text{proj}_2(f)(X)$ derived from the uniform random vector $X \sim \mu(H_n)$ is sub-exponential by Lemma 2.4.20. We formally state this observation in the following lemma to use later in the proof of Theorem 7.4.1.

Lemma 6.2.5. *Let X be a random vector uniformly distributed over H_n. For any function $f : H_n \to \mathbb{R}$, the derived random variable $\text{proj}_2(f)(X)$ is sub-exponential with parameters $(c_1 M_f^2,\ c_2 M_f)$ where $M_f = \|\text{proj}_2 f\|_2/\sqrt{2}$, and $c_1, c_2 > 0$ are the same absolute constants that appear in Lemma 2.4.20. That is,*

$$\mathbb{E}_{X\sim\mu(H_n)} \exp\left(\lambda\ \text{proj}_2(f)(X)\right) \le \exp\left(\frac{\lambda^2}{2} c_1 M_f^2\right), \qquad \textit{for all } \lambda \textit{ s.t. } |\lambda| \le \frac{1}{c_2 M_f}.$$

Proof. Let A be a symmetric $n\times n$ matrix such that $A_{ii} = 0,\ \forall i$ and $A_{ij} = \frac{1}{2}\mathbb{E}_{Y\sim\mu(H_n)}[Y_i Y_j f(Y)]$ for $i \neq j$. Then we observe that for all $X \in H_n$,

$$\text{proj}_2(f)(X) = \sum_{\substack{i=1\\ j>i}}^{n} X_i X_j \mathbb{E}_{Y\sim\mu(H_n)}[Y_i Y_j f(Y)] = X^T A X.$$

Note that X_i is sub-Gaussian with sub-Gaussian parameter 1 for all i because $\mathbb{E}[e^{\lambda X_i}] = \frac{1}{2}(e^{\lambda} + e^{-\lambda}) \le e^{\frac{\lambda^2}{2}}$. To conclude the proof, we apply Lemma 2.4.20 and observe that $\|A\|_F^2 = \sum_{\substack{i=1\\ j\neq i}}^{n} \left(\frac{1}{2}\mathbb{E}_{X\sim\mu(H_n)}[X_i X_j f(X)]\right)^2 = \frac{1}{2}\|\text{proj}_2 f\|_2^2$ and $\|A\|_{op} \le \|A\|_F$. ∎

Chapter 7

On Approximating the PSD Cone by Smaller-sized PSD Constraints

7.1 Introduction

In this chapter, we ask the following question:

> *"How closely can we approximate the cone of $n \times n$ positive semidefinite matrices, with a cone that can be described using at most N number of $k \times k$ PSD constraints?"*

Our motivation for studying this question is from the need for theoretical understanding of the optimality gap (=the difference in the optimal value) arising from the relaxation of SDPs as discussed in Section 6.1.2. Specifically, we consider a convex cone[1] $\mathcal{K} \supseteq \boldsymbol{S}^n_+$ and the following relaxation of SDP:

$$\begin{array}{lll} \text{maximize} & \langle C, X \rangle & \\ \text{subject to} & \langle A_i, X \rangle = b_i, \quad i \in [m], & \\ & X \in \boldsymbol{S}^n_+ & \end{array} \quad \xrightarrow{\text{relax}} \quad \begin{array}{lll} \text{maximize} & \langle C, X \rangle & \\ \text{subject to} & \langle A_i, X \rangle = b_i, \quad i \in [m], & \\ & X \in \mathcal{K} & \end{array} \tag{7.1}$$

where $C, A_i \in \boldsymbol{S}^n$ are problem data. Intuitively, we expect if the cone $\mathcal{K}$ closely approximates $\boldsymbol{S}^n_+$, then the optimality gap resulting from the relaxation (7.1) will be small.

7.1.1 Overview and Contributions of the Chapter

In this chapter, we consider global, non-adaptive approximability of $\boldsymbol{S}^n_+$ by $\mathcal{K}$ that do not make use of the problem data $C, (A_i, b_i)_{i=1}^m$. The main contributions of this chapter are the lower bounds on N, the number of $k \times k$ PSD constraints required to describe the cone $\mathcal{K}$ that well approximates the PSD cone $\boldsymbol{S}^n_+$. In particular, our $\boldsymbol{S}^k_+$-extension complexity lower bounds (Section 7.4) imply that it is hard to globally approximate the PSD cone $\boldsymbol{S}^n_+$ using only a few, small-sized PSD constraints for any construction of the approximating cone $\mathcal{K}$.

[1] In Section 6.1.2, we discussed approximating $\boldsymbol{S}^n_+$ in an SDP with a cone $\mathcal{K} \subseteq \boldsymbol{S}^n_+$. These are essentially the same because we may interpret (7.1) as approximating the dual SDP by replacing $\boldsymbol{S}^n_+$ with $\mathcal{K}^* \supseteq \boldsymbol{S}^n_+$.

To this end, we begin by specifying the notion of approximation to formally state the question asked at the beginning of the chapter. Then, we present two types of lower bounds on N: one is specific to a certain construction scheme for approximating cones $\mathcal{K}$ (namely, k-PSD approximations), whereas the other is a construction-independent complexity lower bound that applies to *any* constructions of $\mathcal{K}$. In the rest of this section, we provide a more detailed overview of the results in this chapter.

Notions of Approximation First, we specify the notions of approximation for cones we use. Let $H = \{X \in \boldsymbol{S}^n : \text{Tr } X = 1\}$ and for any cone $\mathcal{K} \supseteq \boldsymbol{S}^n_+$, let $B_H(\mathcal{K}) = (\mathcal{K} \cap H) - \frac{1}{n} I_n$ where I_n is the $n \times n$ identity matrix. That is, $B_H(\mathcal{K})$ is the unit-trace affine section of $\mathcal{K}$ translated by $-\frac{1}{n} I_n$; note that $0 \in B_H\left(\boldsymbol{S}^n_+\right)$. For $\epsilon > 0$, we say $\mathcal{K}$ is an *ϵ-approximation* of $\boldsymbol{S}^n_+$ if $B_H\left(\boldsymbol{S}^n_+\right) \subseteq B_H(\mathcal{K}) \subseteq (1+\epsilon) B_H\left(\boldsymbol{S}^n_+\right)$.

This notion of approximation is natural and closely related to quantifying the difference in the optimal value (optimality gap) induced by relaxing $\boldsymbol{S}^n_+$ to $\mathcal{K}$ as in (7.1). Suppose that we are given a SDP of the form (6.1) with $m = 1$, $A_1 = I_n$, and $b_1 = 1$, and we relax the problem by replacing $\boldsymbol{S}^n_+$ with a cone $\mathcal{K} \supseteq \boldsymbol{S}^n_+$. If $\mathcal{K}$ is an ϵ-approximation of $\boldsymbol{S}^n_+$, then the relative optimality gap is at most ϵ for all $C \in \boldsymbol{S}^n$.

We also define two auxiliary notions of approximation for the convenience of our analysis. Observe that the notion of ϵ-approximation requires $B_H(\mathcal{K})$ to approximate $B_H\left(\boldsymbol{S}^n_+\right)$ well in all directions in the ambient space. We introduce more lenient notions of approximation by requiring the relative optimality gap to be small only on average for randomized C with standard Gaussian distribution. Specifically, $\mathcal{K}$ is called an ϵ-approximation of $\boldsymbol{S}^k_+$ in the *average sense* if $B_H\left(\boldsymbol{S}^n_+\right) \subseteq B_H(\mathcal{K})$ and $w_G\big(B_H(\mathcal{K})\big) \leq (1+\epsilon) \cdot w_G\big(B_H\left(\boldsymbol{S}^n_+\right)\big)$ where $w_G(S) = \mathbb{E}_g\left[\sup_{x \in S} \langle g, x \rangle\right]$ denotes the Gaussian width of S. Likewise, $\mathcal{K}$ is called an ϵ-approximation of $\boldsymbol{S}^k_+$ in the *dual-average sense* if $B_H\left(\boldsymbol{S}^n_+\right) \subseteq B_H(\mathcal{K})$ and $w_G\big(B_H(\mathcal{K})^\circ\big) \geq \frac{1}{1+\epsilon} \cdot w_G\big(B_H\left(\boldsymbol{S}^n_+\right)^\circ\big)$. More details about these notions can be found in Section 7.2.

k-PSD Approximations of $\boldsymbol{S}^n_+$ In Section 7.3, we consider approximating $\boldsymbol{S}^n_+$ by enforcing PSD constraints on certain k-dimensional subspaces in $\mathbb{R}^n$. We begin by formally defining the k-PSD approximation of $\boldsymbol{S}^n_+$.

Definition 7.1.1 (k-PSD approximation)**.** Let $\mathcal{V} = \{V_1, \ldots, V_N\}$ be a set of k-dimensional subspaces of $\mathbb{R}^n$. The k-PSD approximation of $\boldsymbol{S}^n_+$ induced by $\mathcal{V}$ is the convex cone

$$\boldsymbol{S}^{n,k}_+(\mathcal{V}) := \left\{X \in \boldsymbol{S}^n : v^T X v \geq 0,\ \forall v \in V_i,\ \forall i = 1, \ldots, N\right\}.$$

Note that $\boldsymbol{S}^n_+ \subseteq \boldsymbol{S}^{n,k}_+(\mathcal{V})$ and that $\boldsymbol{S}^{n,k}_+(\mathcal{V})$ can be characterized using at most $N = |\mathcal{V}|$ number of $k \times k$ PSD constraints. A prominent example is the so-called sparse k-PSD approximation, denoted by $\boldsymbol{S}^{n,k}_+$, which is a k-PSD approximation of $\boldsymbol{S}^n_+$ induced by the collection of $N = \binom{n}{k}$

subspaces of k-sparse vectors in $\mathbb{R}^n$.

Our first main results (Theorem 7.3.2 and Corollary 7.3.4) state that if $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$, then $N \geq \exp\left(n \cdot \max\{1/(1+\epsilon) - \sqrt{k/n},\ 0\}^2\right)$ is necessary, regardless of the choice of the subspaces $V_1, \ldots, V_N$; see Remark 7.3.5 in Section 7.3.1. For instance, Corollary 7.3.4 implies that for any $\epsilon > 0$, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$ unless $k = \Omega_n(n)$.

We remark that the conclusion of Theorem 7.3.2 (and Corollary 7.3.4) is possibly too conservative, especially when the subspaces have overlaps. It is because the proof of Theorem 7.3.2 only takes the number of subspaces into consideration, and is oblivious to the configuration of the subspaces in $\mathcal{V}$. In Section 7.3.2, we elaborate on this point with an example of the sparse k-PSD approximation. Although Corollary 7.3.4 already suggests that k must scale at least linearly as n in order for $\boldsymbol{S}_+^{n,k}$ to approximate $\boldsymbol{S}_+^n$, it becomes uninformative once k/n exceeds a certain threshold (approximately 0.137); see Section 7.3.2 and Figure 7.4b.

In Section 7.3.3, a tailored analysis for the sparse k-PSD approximation is provided. To be specific, we consider a carefully designed matrix in $\boldsymbol{S}_+^{n,k} \setminus \boldsymbol{S}_+^n$ to show $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \geq \frac{n-k}{k-1}$ (Theorem 7.3.7) where $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) := \inf\{\epsilon > 0 : \boldsymbol{S}_+^{n,k}$ is an ϵ-approximation of $\boldsymbol{S}_+^n\}$. Furthermore, we prove a sharper lower bound for $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ that is strictly positive for all $1 \leq k < n$, using the duality between $\boldsymbol{S}_+^{n,k}$ and the cone of factor width at most k (Theorem 7.3.8). See Figure 7.1a for comparison between these tailored results and the weak bound obtained from Corollary 7.3.4.

Approximate Extended Formulations of $\boldsymbol{S}_+^n$ Recall that a k-PSD approximation of $\boldsymbol{S}_+^n$ is the intersection of sets in $\boldsymbol{S}^n$, each of which is described with a $k \times k$ PSD constraint. Instead of directly intersecting sets in $\boldsymbol{S}^n$, we may introduce additional variables in pursuit of a more compact description. To be precise, we can lift $\boldsymbol{S}^n$ to a higher-dimensional space by embedding, intersect the lifted space with $k \times k$ PSD constraints, and then project the intersection back to describe a set in $\boldsymbol{S}^n$. The resulting description is called an extended formulation of the set, and the preimage of the projection is called the lifted representation (or PSD lift) of the set. The $\boldsymbol{S}_+^k$-extension complexity of a set S, denoted by $\text{xc}_{\boldsymbol{S}_+^k}(S)$, counts the minimum number of $k \times k$ PSD constraints required to describe S using extended formulation (i.e., with an arbitrary number of additional variables allowed in the description).

In Section 7.4, we argue that any set that well approximates $B_H\left(\boldsymbol{S}_+^n\right)$ must have $\boldsymbol{S}_+^k$-extension complexity at least superpolynomially large in n if k is much smaller than n. That is, it is impossible to approximate $B_H\left(\boldsymbol{S}_+^n\right)$ using only polynomially many $k \times k$ PSD constraints, for any construction of the approximating set. To be precise, if S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then $\text{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{1+\epsilon}\right)^{1/2},\ \frac{1}{1+\epsilon}\ \frac{n}{k}\right\}\right)$ (Theorem 7.4.1); and if S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then $\text{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3},\ \frac{1}{1+\epsilon}\left(\frac{n}{k}\right)^{1/2}\right\}\right)$

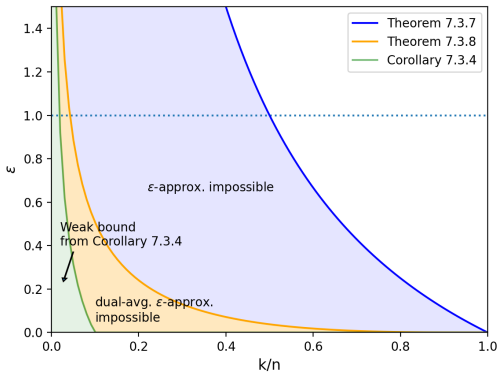
Table 7.1: Overview of our hardness results about approximating $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^k$, presented in terms of the number N of the $k \times k$ PSD constraints needed to construct an ϵ-approximation of $\boldsymbol{S}_+^n$. Here, $C_1, C_2 > 0$ are some universal constants and $\gtrsim$ indicates that the inequality holds asymptotically in the limit $n \to \infty$.

Notion of Approx.	k-PSD Approximations of $\boldsymbol{S}_+^n$	Approximate Extended Formulations of $\boldsymbol{S}_+^n$
ϵ-approx. (Definition 7.2.1)	$N \gtrsim \exp\left(n \cdot \max\left\{\frac{1}{1+\epsilon} - \sqrt{\frac{k}{n}},\ 0\right\}^2\right)$ (Theorem 7.3.2 & Corollary 7.3.4)	$N \gtrsim \exp\left(C_1 \cdot \min\left\{\sqrt{\frac{n}{1+\epsilon}},\ \frac{1}{1+\epsilon}\,\frac{n}{k}\right\}\right)$ (Theorem 7.4.1)
avg. ϵ-approx. (Definition 7.2.2)	Same lower bound as on the right	$N \gtrsim \exp\left(C_2 \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3},\ \frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}\right\}\right)$ (Theorem 7.4.2)

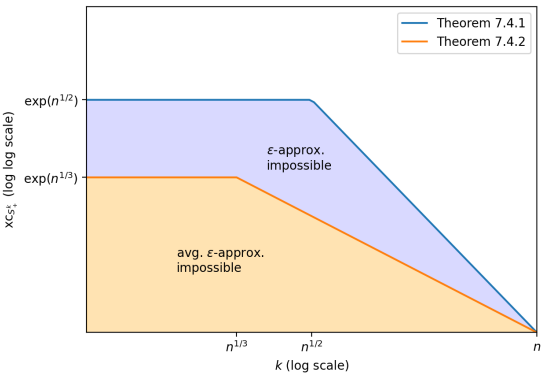
(Theorem 7.4.2). These results are visually illustrated in Figure 7.1b. We remark that these results extend [56, Theorems 1 & 2] beyond the special case $k = 1$.

Nevertheless, we do not know whether our extension complexity lower bounds are tight. It might be possible to achieve stronger extension complexity lower bounds (i.e., move the curves upward) by means of a more sophisticated analysis; for example, when $k = 1$, the lower bound from Theorem 7.4.1 reads as $\exp(C\sqrt{n})$, whereas the current best construction has size $\exp(cn)$. We are curious if it could be possible to achieve a matching exponential complexity lower bound of order $\exp(n)$ for $k = 1$, and leave it as an interesting open problem.

Summary of Results Table 7.1 summarizes the results in this paper. The lower bounds in the table imply the hardness of approximating $\boldsymbol{S}_+^n$ by using only a small number of $k \times k$ PSD constraints when $k = o(n)$.



(a) Hardness results for sparse k-PSD approximations. A weak bound from generic result (Cor. 7.3.4) and tailored bounds (Thms. 7.3.7 & 7.3.8).



(b) Impossibility of approximating $B_H(\boldsymbol{S}_+^n)$ with a polynomial number of $k \times k$ PSD constraints implied by Thm. 7.4.1 and Thm. 7.4.2.

Figure 7.1: Summary of the results in this chapter about the hardness of approximating $\boldsymbol{S}_+^n$.

7.1.2 Organization of the Chapter

In Section 7.2, we begin by defining the notions of approximation that will be used through this chapter. In Section 7.3, we consider a specific construction of approximating cones, namely, the k-PSD approximations of $\boldsymbol{S}^n_+$, which is defined as the set of $n \times n$ symmetric matrices that are PSD when restricted to a given fixed set of k-dimension al subspaces in $\mathbb{R}^n$. Specifically, Section 7.3.1 discusses a generic lower bound on the number of subspaces required to approximate $\boldsymbol{S}^n_+$, and Section 7.3.2 provides a more refined analysis tailored to the so-called sparse k-PSD approximation of $\boldsymbol{S}^n_+$. In Section 7.4, we consider the approximate extended formulations of $\boldsymbol{S}^n_+$, and study a construction-independent complexity lower bound of approximating the PSD cone. More precisely, we present two main theorems that imply the hardness of approximating $\boldsymbol{S}^n_+$. In Section 7.5, we make a few comments on the results in this chapter as well as some related work in the literature. Lastly, Section 7.6 contains deferred proofs of the theorems in the chapter.

7.1.3 Additional Notation

Given $X \in \boldsymbol{S}^n$ and $I \subset [n]$, let $X_I \in \boldsymbol{S}^{|I|}$ denote the principal submatrix of X with row/column indices in I. Throughout this chapter, the letter H is reserved to indicate the subspace of unit trace, $H = \{X \in \boldsymbol{S}^n : \operatorname{Tr} X = 1\}$. For a cone $\mathcal{K} \subseteq \boldsymbol{S}^n$, its base (translated by $-\frac{1}{n} I_n$) is the compact set defined to be $B_H(\mathcal{K}) := \left(\mathcal{K} \cap H\right) - \frac{1}{n} I_n = \{X - \frac{1}{n} I_n \in \boldsymbol{S}^n : X \in \mathcal{K} \cap H\}$, and we define $B^*_H(\mathcal{K}) := B_H(\mathcal{K}^*)$ for notational convenience.

7.2 Three Notions of Approximation

Recall that we want to approximate the positive semidefinite cone $\boldsymbol{S}^n_+$ with a convex cone $\mathcal{K} \supseteq \boldsymbol{S}^n_+$ so that the feasible set $B_H(\mathcal{K}) = \left(\mathcal{K} \cap H\right) - \frac{1}{n} I_n$ (cf. Remark 2.3.2) well approximates $B_H\left(\boldsymbol{S}^n_+\right)$. In Section 7.2.1, we introduce three notions of approximation for sets. In Section 7.2.2, we extend these notions to cones to assess the quality of $\mathcal{K}$ as an approximation of $\boldsymbol{S}^n_+$.

Specifically, we first define a natural notion of ϵ-approximation for sets that contain the origin (Definition 7.2.1). Then, we additionally describe two auxiliary notions of approximation for the convenience of our analysis, namely, the average ϵ-approximation (Definition 7.2.2) and the dual-average ϵ-approximation (Definition 7.2.6). These two auxiliary notions can be obtained by relaxing a quantifier in the definition of ϵ-approximation. These relaxed notions are closely related, but incomparable to each other. They will be respectively used in Section 7.3 and Section 7.4 to prove the hardness of approximating $\boldsymbol{S}^n_+$ with a small number of $k \times k$ PSD constraints.

■ 7.2.1 Notions of Approximation for Sets

To begin with, we define the notion of ϵ-approximation for sets containing the origin.

Definition 7.2.1 (ϵ-approximation)**.** Let P be a set containing 0. For $\epsilon > 0$, a set S is an *ϵ-approximation* of P if $P \subseteq S \subseteq (1+\epsilon)P$. Given two sets P, S that contain 0, we let

$$\epsilon^*(P, S) := \inf\{\epsilon > 0 : S \text{ is an } \epsilon\text{-approximation of } P\}.$$

This is a natural notion to quantify how tightly a set P containing 0 can be approximated by another set $S \supseteq P$. Recall the definition of the support function $h_S(x) := \sup_{z \in S}\langle x, z\rangle$, cf. (2.7). We observe that if S is an ϵ-approximation of P, then

$$h_P(x) \le h_S(x) \le (1+\epsilon)h_P(x) \quad \text{for all } x. \tag{7.2}$$

That is, if S is an ϵ-approximation of P, then for every direction in the ambient space, the distance from the supporting hyperplane of S to the origin is at most $(1+\epsilon)$ times the distance from the supporting hyperplane of P to the origin. Moreover, when P and S are convex, the converse is also true.

Next, we define a more lenient notion of approximation by relaxing the quantifier 'for all x' in (7.2) by taking average over random direction x. To this end, recall the notion of Gaussian width from Definition 2.3.4 that $w_G(S) = \mathbb{E}_G[h_S(G)]$ for any nonempty bounded set $S \subset \mathbb{R}^d$, where G is a standard Gaussian.

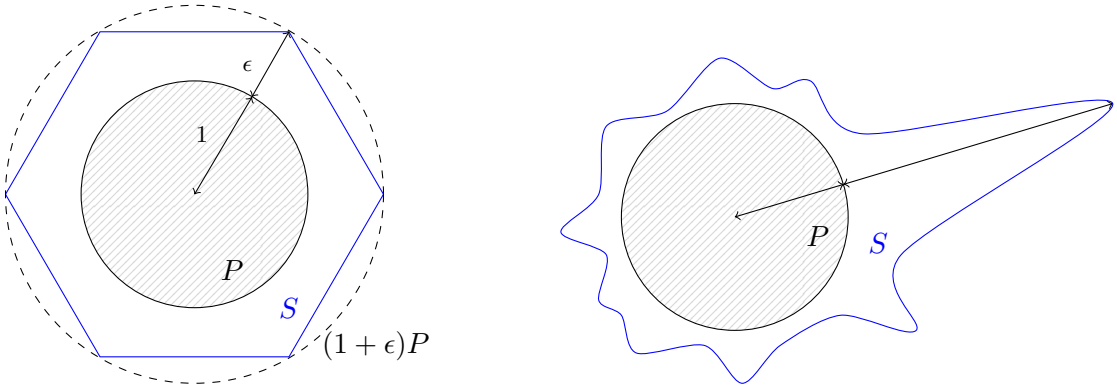
Definition 7.2.2 (average ϵ-approximation)**.** Let P be a set containing 0. For $\epsilon > 0$, a set S is an *average ϵ-approximation* of P, or *ϵ-approximation of P in the average sense*, if $P \subseteq S$ and $w_G(S) \le (1+\epsilon)w_G(P)$. Given two sets P, S that contain 0, we let

$$\epsilon^*_{\text{avg}}(P, S) := \inf\{\epsilon > 0 : S \text{ is an average } \epsilon\text{-approximation of } P\}.$$

By definition, S is an average ϵ-approximation of P if and only if $\mathbb{E}_G[h_S(G) - h_P(G)] \le \epsilon \cdot \mathbb{E}_G\big[h_P(G)\big]$ where G is a standard Gaussian random matrix in $\boldsymbol{S}^n$.

Note that average ϵ-approximation is a weaker notion than ϵ-approximation because $\epsilon^*(P, S) \ge \epsilon^*_{\text{avg}}(P, S)$. That is, for a fixed $\epsilon > 0$, if S is an ϵ-approximation of P, then S is also an average ϵ-approximation of P. As a matter of fact, average ϵ-approximation is a strictly weaker notion because there exists a pair of sets (P, S) such that $\epsilon^*(P, S) > \epsilon^*_{\text{avg}}(P, S)$, i.e., there exists some $\epsilon > 0$ for which S is not an ϵ-approximation of P whereas S is an average ϵ-approximation of P. We illustrate this point with the following two examples.

Example 7.2.3. Let $P = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le 1\}$ and $S = \{(x, y) \in \mathbb{R}^2 : x^2/4 + y^2 \le 1\}$. Then $\epsilon^*(P, S) = 1$. On the other hand, $\epsilon^*_{\text{avg}}(P, S) = \frac{4}{\pi}E(3/4) - 1 \approx 0.54196$ where $E(m)$



(a) S is an ϵ-approximation of P.

(b) S 'poorly' approximates P only in some directions.

Figure 7.2: An illustrating cartoon for ϵ-approximation versus average ϵ-approximation.

is the complete elliptic integral of the second kind with parameter $m = k^2$. The value of $\epsilon^*_{\text{avg}}(P, S)$ can be computed by observing that $w_G(P) = \mathbb{E}_{G\in N(0,I_2)}\|g\|_2$ and $w_G(S) = \mathbb{E}_{g\in N(0,I_2)}\|g\|_2 \cdot \frac{1}{2\pi}\int_0^{2\pi}\sqrt{4\cos^2\theta + \sin^2\theta}d\theta = \frac{4}{\pi}E(3/4)\mathbb{E}_{g\in N(0,I_2)}\|g\|_2$.

Example 7.2.4. Let $P = \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}\}$ and $S = \{x \in \mathbb{R}^n : \|x\|_1 \le \sqrt{n}\}\}$ where $\|\cdot\|_p$ denotes the ℓ_p-norm. Then $\epsilon^*(P, S) = \sqrt{n} - 1$. On the other hand, $\epsilon^*_{\text{avg}}(P, S) = f(n)$ where $f(n)$ is a function of n such that $f(n) \approx \sqrt{2\log n}$ for sufficiently large n. It is because $w_G(P) = \mathbb{E}_{g\sim N(0,I_n)}\|g\|_2 \approx \sqrt{n}$ and $w_G(S) = \sqrt{n}\cdot\mathbb{E}_{g\sim N(0,I_n)}\max_{i\in[n]}|g_i| \approx \sqrt{2n\log n}$.

In the two examples above, we observe that there exists $\epsilon > 0$ such that S is an average ϵ-approximation of P, while S is not an ϵ-approximation of P. This happens because $h_S(G) - h_P(G)$ is small on average, but the difference can be potentially large for some G. In other words, S approximates P well on average, but poorly for certain 'bad' directions in the ambient space, as illustrated in Figure 7.2. Nevertheless, the set of 'bad' directions might have only a small measure as in Example 7.2.4, and the notion of ϵ-approximation as in Definition 7.2.1 can be overly conservative. That is why we additionally consider the notion of average ϵ-approximation, which is more lenient with the shape of the approximating set S.

One drawback of evaluating the quality of approximation with the notion of average ϵ-approximation is that it only measures the difference averaged over an ensemble of random objectives. Thus, we cannot control the gap $h_S(x) - h_P(x)$ for any specific x, however, we can still establish a probabilistic upper bound on $h_S(G) - h_P(G)$ when G is randomly drawn from the standard Gaussian distribution.

Lemma 7.2.5. *Let S be an average ϵ-approximation of P for some $\epsilon > 0$. Then for all $\tau > 0$,*

$$\Pr_{G\sim \text{std Gaussian}}\Big[h_S(G) - h_P(G) > \tau\Big] \le \epsilon\frac{w_G(P)}{\tau}.$$

Lemma 7.2.5 operationally means that if S is an average ϵ-approximation of P for small ϵ, then $h_S(x) - h_P(x)$ can be large only for x in a set that has small measure. In particular, the probability upper bound converges to 0 as $\epsilon \to 0$. That is, $h_S(x) - h_P(x)$ converges to 0 for all x (but those in a set of measure-zero) as $\epsilon \to 0$.

Proof of Lemma 7.2.5. Note that $h_S(G) - h_P(G) \geq 0$ for all g because $P \subseteq S$. The conclusion follows from Markov's inequality and the observation that $w_G(S) - w_G(P) \leq \epsilon \cdot w_G(P)$. ■

Lastly, we revisit Definition 7.2.1 to introduce an alternative relaxation of ϵ-approximation, namely, the 'dual' version of average ϵ-approximation. Recall from (2.6) that the gauge function of S is defined as $p_S(x) := \inf\{\lambda \in \mathbb{R} : \lambda > 0 \text{ and } x \in \lambda S\}$. Observe that $P \subseteq S \subseteq (1+\epsilon)P$ if and only if $\frac{1}{1+\epsilon} p_P(x) \leq p_S(x) \leq p_P(x)$ for all x. When P and S are closed convex sets, $p_P(x) = h_{P^\circ}(x)$ and $p_S(x) = h_{S^\circ}(x)$ by Lemma 2.3.3. Therefore, S is an ϵ-approximation of P if and only if $\frac{1}{1+\epsilon} h_{P^\circ}(x) \leq p_{S^\circ}(x) \leq p_{P^\circ}(x)$ for all x. As before, we ease the condition "$\frac{1}{1+\epsilon} h_{P^\circ}(x) \leq p_{S^\circ}(x)$ for all x" by averaging over x to reach at the following definition.

Definition 7.2.6 (dual-average ϵ-approximation)**.** Let P be a set containing 0. For $\epsilon > 0$, a set S is a *dual-average ϵ-approximation* of P, or *ϵ-approximation of P in the dual-average sense*, if $P \subseteq S$ and $w_G(S^\circ) \geq \frac{1}{1+\epsilon} w_G(P^\circ)$. Given two sets P, S that contain 0, we define

$$\epsilon^*_{\text{dual-avg}}(P, S) := \inf\{\epsilon > 0 : S \text{ is a dual-average } \epsilon\text{-approximation of } P\}.$$

Note that dual-average ϵ-approximation is also a weaker notion than ϵ-approximation. That is, for a fixed $\epsilon > 0$, if S is an ϵ-approximation of P, then S is also a dual-average ϵ-approximation of P. In Section 7.3, we use the notion of dual-average ϵ-approximation as a technical tool to prove the hardness of k-PSD approximations of $\boldsymbol{S}^n_+$.

The notion of dual-average ϵ-approximation is closely related to the notion of average ϵ-approximation; indeed, they are dual to each other. However, they are not equivalent notions of approximation, i.e., there exist convex sets P, S such that S is a good approximation of P in the average sense, but not in the dual average sense; the opposite is also possible. See Remark 7.2.7 and Example 7.2.8.

Remark 7.2.7. For $\epsilon > 0$, S is an average ϵ-approximation of P if and only if P° is a dual-average ϵ-approximation of S°. In other words, $\epsilon^*_{\text{avg}}(P, S) = \epsilon^*_{\text{dual-avg}}(S^\circ, P^\circ)$. In this sense, the notion of dual-average ϵ-approximation is the dual of the notion of average ϵ-approximation.

Example 7.2.8 (Ball, Needle, and Pancake)**.** Consider a d-dimensional unit ℓ_2-ball and a 'needle' obtained by taking the convex hull of the union of the ball and two points that are located on the opposite side of the origin at distance d. The polar of this 'needle' is a 'pancake'

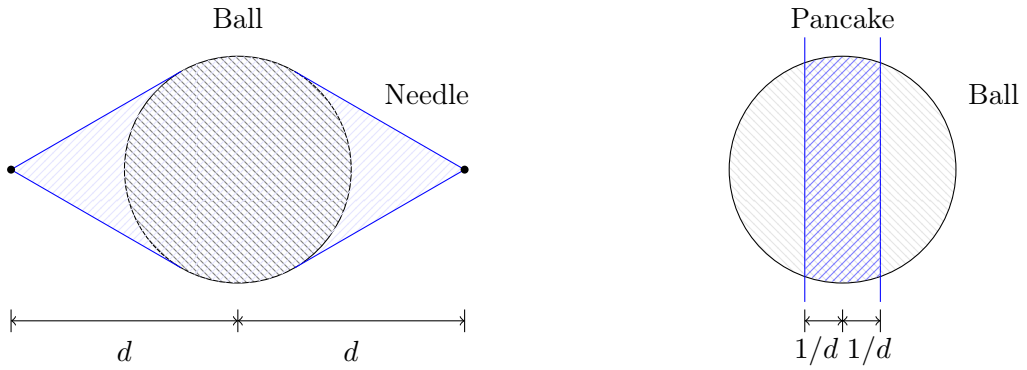


Figure 7.3: An illustration of the sets described in Example 7.2.8.

obtained by intersecting the unit ball with a slab of thickness $2/d$ along its equator. These sets are illustrated in Figure 7.3. We observe that the Gaussian width of the ball, the needle, and the pancake are approximately $\sqrt{d-1/2}, d\sqrt{2/\pi}$, and $\sqrt{d-3/2}$, respectively. Thus, the ball is a good approximation of the pancake in the average sense, but not in the dual-average sense. Likewise, the needle is a good approximation of the ball in the dual-average sense, but not in the average sense. See Section 7.7.1 for more details on this example.

■ 7.2.2 Notions of Approximation for Cones

Recall that our primary motivation for introducing the notions of approximation is to quantify the optimality gap that arises from a conic programming relaxation of an SDP. Consider the relaxation as described in (7.1) that is obtained by replacing the PSD cone $\boldsymbol{S}^n_+$ in an SDP with a larger cone $\mathcal{K} \supseteq \boldsymbol{S}^n_+$. Letting $P = \{X \in \boldsymbol{S}^n_+ : \langle A_i, X\rangle = b_i,\ i = 1,\dots,m\}$ and $S = \{X \in \mathcal{K} : \langle A_i, X\rangle = b_i,\ i = 1,\dots,m\}$ denote the feasible sets of the original and the relaxed problems, we can see that $S \supseteq P$ and there arises an increase in the optimal value, $\Gamma_{P,S}(C) := h_S(C) - h_P(C)$, as a result of the relaxation.

We extend the notions of approximation for sets, defined in Section 7.2.1, to the corresponding notions for cones by fixing a certain affine constraint. Recall that for a cone $\mathcal{K}$, we let $B_H(\mathcal{K}) := \big(\mathcal{K} \cap H\big) - \frac{1}{n} I_n = \big\{X - \frac{1}{n} I_n \in \boldsymbol{S}^n : X \in \mathcal{K} \cap H\big\}$ where $H = \{X \in \boldsymbol{S}^n : \operatorname{Tr} X = 1\}$ and I_n denotes the $n \times n$ identity matrix. Note that $B_H(\mathcal{K})$ is the feasible set of the relaxed problem (7.1), translated by $-\frac{1}{n} I_n$, when the affine constraint is the unit-trace constraint. We define the notions of approximation for cones as follows.

Definition 7.2.9 (ϵ-approximation for cones in $\boldsymbol{S}^n$)**.** A cone $\mathcal{K} \subseteq \boldsymbol{S}^n$ is an ϵ-approximation (average ϵ-approximation / dual-average ϵ-approximation, resp.) of $\boldsymbol{S}^n_+$ if $B_H(\mathcal{K})$ is an ϵ-approximation (average ϵ-approximation / dual-average ϵ-approximation, resp.) of $B_H\big(\boldsymbol{S}^n_+\big)$.

Also, we let

$$\epsilon^*(\boldsymbol{S}^n_+, \mathcal{K}) := \epsilon^*\big(B_H\left(\boldsymbol{S}^n_+\right), B_H\left(\mathcal{K}\right)\big)$$

and define $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \mathcal{K})$ and $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}^n_+, \mathcal{K})$ in a similar manner.

Remark 7.2.10. Remark here that $w_G\big(B_H\left(\boldsymbol{S}^n_+\right)\big) \leq \sqrt{2n}$ and that $\lim_{n\to\infty} \frac{w_G(B_H(\boldsymbol{S}^n_+))}{\sqrt{2n}} = 1$ because $\boldsymbol{S}^n_+ \cap H = \text{conv}\{vv^T : v \in \mathbb{S}^{n-1}\}$, cf. Lemma 2.4.10 and Remark 2.4.11.

■ 7.3 k-PSD Approximations of S^n_+

One possible option to relax the PSD constraint $X \in \boldsymbol{S}^n_+$ is to enforce the PSD constraints only on the smaller $k \times k$ principal submatrices of X, which leads to the following relaxation:

$$\begin{aligned} \text{maximize} \quad & \langle C, X\rangle \\ \text{subject to} \quad & \langle A_i, X\rangle = b_i, \quad i = 1, \ldots, m, \\ & k \times k \text{ principal submatrices of } X \in \boldsymbol{S}^k_+. \end{aligned} \tag{7.3}$$

Note that the PSD cone $\boldsymbol{S}^n_+$ is replaced with a relaxed cone that is defined using $(k \times k)$-sized PSD constraints, and (7.3) can be solved more efficiently when $k \ll n$. For example, $k = 1$ yields a linear programming (LP) approximation and $k = 2$ produces a second-order cone programming (SOCP) approximation of the original SDP [6].

In this section, we consider a scheme to approximate $\boldsymbol{S}^n_+$ by enforcing $k \times k$ PSD constraints on particular subspaces. To be precise, we choose a fixed set of k-dimensional subspaces in $\mathbb{R}^n$ and define a cone of $n \times n$ symmetric matrices that are PSD when restricted to these subspaces. The cone associated with (7.3) is an example of this construction that is obtained by imposing PSD constraints on the $\binom{n}{k}$ subspaces of k-sparse vectors in $\mathbb{R}^n$, and will be referred to as the sparse k-PSD approximation of $\boldsymbol{S}^n_+$.

In Section 7.3.1, we formalize the definition of the k-PSD approximation and prove a lower bound on the number of $k \times k$ PSD constraints required. We show that when k is much smaller than n, it is necessary to impose PSD constraints on at least exponentially many subspaces to produce a cone that approximates $\boldsymbol{S}^n_+$ well. In Section 7.3.2, we consider a concrete example of the sparse k-PSD approximation and discuss the hardness of approximating $\boldsymbol{S}^n_+$. In Section 7.3.3, we revisit the example in more detail with a tailored analysis.

■ 7.3.1 A Lower Bound for k-PSD Approximations of S^n_+

We recall the definition of the k-PSD approximation of $\boldsymbol{S}^n_+$ from Definition 7.1.1.

Definition 7.3.1 (k-PSD approximation of $\boldsymbol{S}^n_+$; restatement of Definition 7.1.1)**.** Let $\mathcal{V} = \{V_1, \ldots, V_N\}$ be a set of k-dimensional subspaces of $\mathbb{R}^n$. The k-PSD approximation of $\boldsymbol{S}^n_+$

induced by $\mathcal{V}$ is the convex cone

$$\boldsymbol{S}_+^{n,k}(\mathcal{V}) := \{X \in \boldsymbol{S}^n : v^T X v \geq 0,\ \forall v \in V_i,\ \forall i = 1, \ldots, N\}.$$

Note that $\boldsymbol{S}_+^{n,k}(\mathcal{V}) \supseteq \boldsymbol{S}_+^n$ is the set of $n \times n$ symmetric matrices whose associated quadratic forms are positive semidefinite when restricted to $V_1 \cup \cdots \cup V_N$. Thus, if $U_i \in \mathbb{R}^{n \times k}$ is a matrix whose columns form a basis of V_i, then $\boldsymbol{S}_+^{n,k}(\mathcal{V}) = \{X \in \boldsymbol{S}^n : U_i^T X U_i \in \boldsymbol{S}_+^k,\ \forall i = 1, \ldots, N\}$.

Our first main theorem in the chapter provides an upper bound on the Gaussian width of the base of the dual cone of $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ as a function of k and $N = |\mathcal{V}|$.

Theorem 7.3.2. *Let n, $1 \leq k \leq n$ be positive integers and $\mathcal{V} = \{V_1, \ldots, V_N\}$ be any set of k-dimensional subspaces of $\mathbb{R}^n$. Then*

$$w_G\Big(B_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)\Big) \leq \sqrt{2k} + \sqrt{2\log N}.$$

Recall that $w_G\big(B_H^*\left(\boldsymbol{S}_+^n\right)\big) = w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big) \approx \sqrt{2n}$, cf. Remark 7.2.10. Comparing the upper bound in Theorem 7.3.2 against $\sqrt{2n}$, we can contrast the size of $B_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)$ relative to $B_H^*\left(\boldsymbol{S}_+^n\right)$. For example, when k and N are small, $\sqrt{2k} + \sqrt{2\log N} \ll \sqrt{2n}$, and we can intuitively see that the dual of the cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is much smaller than the original PSD cone $\boldsymbol{S}_+^n$. Therefore, the primal cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is too big to well approximate $\boldsymbol{S}_+^n$ in such a case.

Remark 7.3.3. Note that the upper bound in Theorem 7.3.2 holds regardless of the subspaces $V_1, \ldots, V_N$ in $\mathcal{V}$, i.e., it is oblivious to the configuration of the subspaces. That is, this upper bound is valid even for the "best" possible configuration of subspaces to imitate the expressive power of the full-sized PSD cone. We also note that this upper bound could conceivably be too conservative, especially when N is large, because it implicitly hinges on the union bound (through the use of Lemma 2.4.21).

Now we discuss how Theorem 7.3.2 implies the hardness of approximating $\boldsymbol{S}_+^n$ with a small number of $k \times k$ PSD constraints. In the next corollary, we show that if $N = |\mathcal{V}|$ is below a certain threshold determined by n, k, ϵ, then $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. Thus, it cannot be an ϵ-approximation of $\boldsymbol{S}_+^n$, either.

Corollary 7.3.4. *Let n, k be positive integers such that $1 \leq k \leq n$, and $\epsilon > 0$. If $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$, then $|\mathcal{V}| \geq \exp\big(n \cdot \varphi(n,k,\epsilon)\big)$ where*

$$\varphi(n,k,\epsilon) = \left[\ \frac{1}{1+\epsilon}\frac{w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)}{\sqrt{2n}} - \sqrt{\frac{k}{n}}\ \right]_+^2.$$

Remark 7.3.5. Recall from Remark 2.4.11 that $\lim_{n\to\infty} w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)/\sqrt{2n} = 1$. With $k =$

$\lfloor \delta n \rfloor$ for $0 < \delta < 1$,

$$\lim_{n\to\infty} \varphi\big(n, \lfloor \delta n \rfloor, \epsilon\big) = \left[\frac{1}{1+\epsilon} - \sqrt{\delta} \right]_+^2.$$

That is, when n is sufficiently large, $|\mathcal{V}| \geq \exp\left(n\big[1/(1+\epsilon) - \sqrt{\delta}\big]_+^2\right)$ is necessary for the cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ to be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

As discussed in Remark 7.3.3, our lower bound in Corollary 7.3.4 can be excessively conservative due to the union bound. In fact, we do not know whether our lower bound is tight. Thus, it is possible that even if $N \geq \exp\big(n \cdot \varphi(n,k,\epsilon)\big)$, there does not exist any $\mathcal{V}$ such that $|\mathcal{V}| = N$ and $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

■ 7.3.2 Example: the Sparse k-PSD Approximation

In this section, we consider the sparse k-PSD approximation, which is a concrete example of the k-PSD approximation of $\boldsymbol{S}_+^n$ (Definition 7.3.1) discussed in the previous section.

Definition 7.3.6 (Sparse k-PSD approximation of $\boldsymbol{S}_+^n$)**.** Given positive integers n and $1 \leq k \leq n$, the sparse k-PSD approximation of $\boldsymbol{S}_+^n$ is the set

$$\boldsymbol{S}_+^{n,k} := \big\{X \in \boldsymbol{S}^n : X_I \succeq 0, \quad \forall I \subset [n] \text{ with } |I| \leq k\big\}.$$

We observe that the sparse k-PSD approximation is an instance of the k-PSD approximation $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ such that $\mathcal{V} = \{V_I : I \in [n] \text{ with } |I| = k\}\}$ where $V_I = \{v \in \mathbb{R}^n : v_i = 0, \ \forall i \notin I\}$. Note that $|\mathcal{V}| = \binom{n}{k}$.

In this section, we examine the implications of Corollary 7.3.4 for the sparse k-PSD approximation of $\boldsymbol{S}_+^n$. Later in Section 7.3.3, we provide a more refined analysis that is tailored to $\boldsymbol{S}_+^{n,k}$, based on properties that are specific to $\boldsymbol{S}_+^{n,k}$. It turns out that we can derive stronger hardness results from the tailored approach.

A Weak Bound Using Corollary 7.3.4 First of all, we inspect what the lower bound obtained in Section 7.3.1 implies for the sparse k-PSD approximation of $\boldsymbol{S}_+^n$. According to the contrapositive of Corollary 7.3.4, when n and $\epsilon > 0$ are fixed, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$ if k satisfies the following inequality:

$$\binom{n}{k} < \exp\left(n \cdot \left[\frac{1}{1+\epsilon} \frac{w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)}{\sqrt{2n}} - \sqrt{\frac{k}{n}} \right]_+^2\right). \tag{7.4}$$

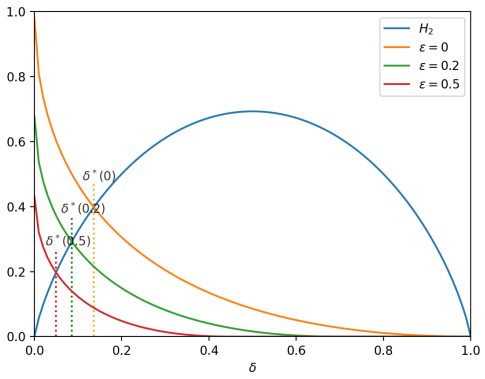
Assume $k = \delta n$ for some $0 < \delta < 1$ and n tends to infinity. By Stirling's approximation,

$$\log \binom{n}{k} = \big(1 + o_n(1)\big) H_2\left(\frac{k}{n}\right) n,$$

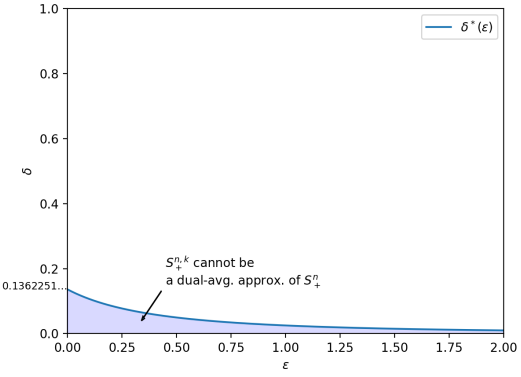
where $H_2(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function defined for $p \in [0,1]$. With this asymptotic approximation and the observation that $w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)/\sqrt{2n} \le 1$, we take logarithm of both sides of (7.4) to obtain the inequality (in the limit $n \to \infty$),

$$H_2(\delta) < \left[\ \frac{1}{1+\epsilon} - \sqrt{\delta}\ \right]_+^2. \tag{7.5}$$

Given $\epsilon \ge 0$, let $g_\epsilon(\delta) := \left[\ \frac{1}{1+\epsilon} - \sqrt{\delta}\ \right]_+^2 - H_2(\delta)$. Note that g_ϵ is strictly convex on the interval $\delta \in [0,1]$ and $g_\epsilon(0) > 0$. Moreover, if $\epsilon > 0$, then $g_\epsilon(1/(1+\epsilon)^2) < 0$. By the intermediate value theorem, there exists a unique $0 < \delta^*(\epsilon) < 1/(1+\epsilon)^2$ such that $g_\epsilon(\delta^*(\epsilon)) = 0$ and $g_\epsilon(\delta) > 0$ for all $0 \le \delta < \delta^*(\epsilon)$. Thus, if $k/n < \delta^*(\epsilon)$, then $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. The expressions on both sides of Eq. (7.5) are illustrated in Figure 7.4a for a few values of ϵ; the plot of $\delta^*(\epsilon)$ vs ϵ is depicted in Figure 7.4b.



(a) Plot of the expressions in (7.5): $H_2(\delta)$ (entropy) vs $\left[\ 1/(1+\epsilon) - \sqrt{\delta}\ \right]_+^2$ for $\epsilon = 0$, 0.2, and 0.5. The location of $\delta^*(\epsilon)$ are also annotated.



(b) Plot of $\delta^*(\epsilon)$ vs ϵ. For a fixed $\epsilon > 0$, if k/n is contained in the blue region, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

Figure 7.4: Hardness of the sparse k-PSD approximations implied by Corollary 7.3.4.

Recall that $\epsilon^*_{\text{dual-avg}}(P,S) := \inf\{\epsilon > 0 : S \text{ is a dual-average } \epsilon \text{ approximation of } P\}$, which indicates the 'best possible' (i.e., the smallest) $\epsilon > 0$ for which S is a dual-average ϵ-approximation of P. For fixed n and k, the preceding discussion leads to a lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ as

$$\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \ge \sup\left\{\epsilon > 0 : H_2\Big(\frac{k}{n}\Big) < \left[\ \frac{1}{1+\epsilon} - \sqrt{\frac{k}{n}}\ \right]_+^2\right\} =: \xi(k/n). \tag{7.6}$$

On the one hand, we can already see from the above discussion that for any fixed $\epsilon > 0$, $\boldsymbol{S}_+^{n,k}$ with $k = o_n(n)$ cannot be an ϵ-approximation of $\boldsymbol{S}_+^n$ (in the dual-average sense). That

is, k must scale linearly with respect to n for $\boldsymbol{S}_+^{n,k}$ to be a good approximation of $\boldsymbol{S}_+^n$. On the other hand, the lower bound on k from the discussion above, i.e., $k/n \geq \delta^*(\epsilon)$, becomes uninformative once k increases beyond a certain threshold because $\delta^*(\epsilon) < \delta^*(0) \approx 0.137$ for all $\epsilon > 0$. In other words, if $k/n > \delta^*(0)$, then we can only get a trivial lower bound $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) > -\infty$, and do not know whether $\boldsymbol{S}_+^{n,k}$ approximates $\boldsymbol{S}_+^n$ well or not.

We remark that this is possibly due to the conservative nature of inequality (7.4), which is inherited from Corollary 7.3.4. Recall that the cardinality lower bound from Corollary 7.3.4 is oblivious to the configuration of the subspaces $V_1, \ldots, V_N$ in $\mathcal{V}$. That is, it is valid even for the "best" possible configuration of subspaces to imitate the expressive power of the full-sized PSD cone. Nevertheless, the subspaces of k-sparse vectors have overlaps, and some of them could be redundant. Thus, the general lower bound from Corollary 7.3.4 can be excessively conservative to apply to the sparse k-PSD approximation of $\boldsymbol{S}_+^n$.

Indeed, we can acquire a tighter lower bound for $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ by using the knowledge about the subspaces of $\boldsymbol{S}_+^{n,k}$. This is the topic that will be discussed in Section 7.3.3.

■ 7.3.3 Tailored Analysis for the Sparse k-PSD Approximation

In this section, we derive lower bounds on $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ and $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ with an analysis that exploits specific properties of $\boldsymbol{S}_+^{n,k}$. More precisely, we construct a matrix on the boundary of $B_H(\boldsymbol{S}_+^{n,k})$ to argue a lower bound on $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$, and characterize $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ by observing that the Gaussian width of $B_H^*(\boldsymbol{S}_+^{n,k})$ is the expectation of the largest k-sparse eigenvalue of a standard Gaussian random matrix. The resulting lower bounds imply stronger hardness results for approximating $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^{n,k}$ than those discussed in Section 7.3.2.

Hardness of ϵ-approximation First of all, we discuss how hard it is to approximate $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^{n,k}$ in the ϵ-approximation sense (see Definition 7.2.1) when k is small. For that purpose, we consider a specific matrix on the line segment connecting $\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ and $\frac{1}{n}I_n$ where $\mathbf{1}_n \in \mathbb{R}^{n\times 1}$ denotes the $n \times 1$ column matrix with all entries equal to 1. Specifically, we construct a matrix $M \in B_H(\boldsymbol{S}_+^{n,k})$ that is far away from $B_H\left(\boldsymbol{S}_+^n\right)$, and prove a lower bound for $\epsilon > 0$ as a necessary condition for $M \in (1+\epsilon) \cdot B_H\left(\boldsymbol{S}_+^n\right)$.

Theorem 7.3.7. *If $\boldsymbol{S}_+^{n,k}$ is an ϵ-approximation of $\boldsymbol{S}_+^n$, then $k > \frac{n-1}{1+\epsilon}$.*

Proof. Let $P_1(n) := \mathbf{1}_n\mathbf{1}_n^T/n$ and $P_2(n) := I_n - P_1(n)$. Note that $P_1(n)$ and $P_2(n)$ are projection matrices. For $a, b \in \mathbb{R}$, we define

$$G(a, b; n) := aP_1(n) + bP_2(n).$$

It is easy to verify that the eigenvalues of $G(a, b; n)$ are a with multiplicity 1, and b with multiplicity $n - 1$.

Next, recall from Definition 7.3.6 that $G(a,b;n) \in \boldsymbol{S}_+^{n,k}$ if and only if $G(a,b;n)_{[k]} \succeq 0$. Observe that $G(a,b;n)_{[k]} = \frac{ka+(n-k)b}{n} P_1(k) + bP_2(k) \succeq 0$ if and only if $ka + (n-k)b \geq 0$ and $b \geq 0$. Letting $a = \frac{k-n}{n(k-1)}$ and $b = \frac{k}{n(k-1)}$, we observe that (1) $G(a,b;n) \in \boldsymbol{S}_+^{n,k}$ because $ka + (n-k)b = 0$ and $b \geq 0$; and (2) $G(a,b;n) \in H$ because $\operatorname{Tr} G(a,b;n) = a + b(n-1) = 1$. Next, we can also verify that $G(a,b;n) - \frac{1}{n} I_n \in (1+\epsilon) \cdot B_H\left(\boldsymbol{S}_+^n\right)$ if and only if $\epsilon \geq \frac{n-k}{k-1}$. It is because $G(a,b;n) + \frac{\epsilon}{n} I_n = G\big(a + \frac{\epsilon}{n}, b + \frac{\epsilon}{n}; n\big) \in \boldsymbol{S}_+^n$ if and only if $a + \frac{\epsilon}{n} \geq 0$. Rewriting $\epsilon \geq \frac{n-k}{k-1}$ as a condition for k in terms of ϵ, we obtain $k \geq \frac{n-1}{1+\epsilon} + 1$. ■

Alternatively, when k is fixed, Theorem 7.3.7 implies that

$$\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \geq \frac{n-k}{k-1} \geq \frac{1-k/n}{k/n} =: \zeta(k/n). \tag{7.7}$$

Hardness of dual average ϵ-approximation Next, we re-examine how well $\boldsymbol{S}_+^{n,k}$ can approximate $\boldsymbol{S}_+^n$ in the dual-average sense (Definition 7.2.6) to find a better lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$. We use the duality between $\boldsymbol{S}_+^{n,k}$ and its dual cone, $(\boldsymbol{S}_+^{n,k})^* = \operatorname{cone}\{vv^T : v \in \mathbb{R}^n \text{ with } \|v\|_0 \leq k\}$, which is the cone of matrices that have factor width at most k [22].

Observe that $\boldsymbol{S}_+^n \cap H = \{X \in \boldsymbol{S}_+^n : \operatorname{Tr}(X) = 1\} = \operatorname{conv}\{vv^T : x \in \mathbb{R}^n,\ \|v\|_2 = 1\}$. For any $G \in \boldsymbol{S}^n$, $\max_{X \in \boldsymbol{S}_+^n \cap H} \langle G, X\rangle = \lambda_1(G)$ and thus, $w_G\big(\boldsymbol{S}_+^n \cap H\big)$ is equal to the expectation of the largest eigenvalue of a random matrix that has the standard Gaussian distribution in $\boldsymbol{S}^n$ (Definition 2.4.2). Likewise, $(\boldsymbol{S}_+^{n,k})^* \cap H = \operatorname{conv}\{vv^T : x \in \mathbb{R}^n,\ \|v\|_2 = 1,\ \|v\|_0 \leq k\}$, and $\max_{X \in (\boldsymbol{S}_+^{n,k})^* \cap H} \langle G, X\rangle$ is the largest k-sparse eigenvalue of G. Based on these observations, we show an asymptotic upper bound on the ratio $w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big)/w_G(B_H^*(\boldsymbol{S}_+^n))$ in Theorem 7.3.8 that subsequently leads to a tighter lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ in (7.10).

Theorem 7.3.8. *Fix $0 < \delta < 1$ and let $k = \lfloor \delta n \rfloor$. Then*

$$\lim_{n\to\infty} \frac{w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big)}{w_G\big(B_H^*(\boldsymbol{S}_+^n)\big)} \leq \left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{1/2}, \tag{7.8}$$

where Q_{χ^2} denotes the quantile function[2] of the χ^2-distribution with one degree of freedom. Moreover,

$$\int_0^\delta Q_{\chi^2}(1-s)ds = \delta + \sqrt{\frac{2}{\pi}} \Phi^{-1}\left(1 - \frac{\delta}{2}\right) \exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(1 - \frac{\delta}{2}\right)\right]^2 \right) \tag{7.9}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

Before we prove Theorem 7.3.8, we note that it implies the following lower bound in the

[2]That is, $Q_{\chi^2}(s) := \inf\{x \in \mathbb{R} : F_{\chi^2}(x) \geq s\}$ for $0 < s \leq 1$ where F_{χ^2} be the cumulative distribution function of the χ^2-distribution with one degree of freedom.

asymptotic limit $n \to \infty$:

$$\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+) \geq \left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{-1/2} - 1 =: \psi(k/n). \tag{7.10}$$

See Figure 7.1 (left) in Section 7.1 to compare the three lower bounds, ξ (Corollary 7.3.4 and (7.6)), ζ (Theorem 7.3.7 and (7.7)), and ψ (Theorem 7.3.8 and (7.10)). We make two remarks: one on the advantage of tailored analysis for $\boldsymbol{S}^{n,k}_+$; and the other on comparing the rate of convergence for ζ vs ψ.

- (Generic vs tailored) The lower bound ψ gives a sharper lower bound than ξ. In particular, $\psi(\delta) > 0$ for all $0 < \delta < 1$ and ψ gracefully converges to 0 as $k/n \to 1$, whereas $\xi(\delta) = 0$ for all $\delta \geq \delta^*(0)$.
- (ϵ-approx. vs dual-avg. ϵ-approx.) We can see from the expression in (7.9) that $\psi(1-\delta) = \Theta_\delta(\delta^3)$ as $\delta \to 0$. This sharply contrasts with $\varphi(1-\delta) = \Theta_\delta(\delta)$. That is, $\boldsymbol{S}^{n,k}_+$ gets harder to approximate $\boldsymbol{S}^n_+$ in both senses as k diminishes from n, but at a much slower rate in the dual-average sense.

Hardness of average ϵ-approximation As a matter of fact, we can derive the following corollary from Theorem 7.3.8 by applying Urysohn's inequality (Lemma 2.3.5), thereby obtaining an asymptotic lower bound on $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+)$ (see Definition 7.2.2).

Corollary 7.3.9. *Fix $0 < \delta < 1$ and let $k = \lfloor \delta n \rfloor$. Then*

$$\lim_{n\to\infty} \frac{w_G\big(B_H(\boldsymbol{S}^{n,k}_+)\big)}{w_G\big(B_H(\boldsymbol{S}^n_+)\big)} \geq \frac{1}{4}\left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{-1/2}.$$

Corollary 7.3.9 implies that $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+) \geq \frac{1}{4}\big(\int_0^\delta Q_{\chi^2}(1-s)ds \big)^{-1/2} - 1$. Note that this lower bound is more conservative than the lower bound in (7.10), due to the additional multiplier $1/4$ that arises from the use of Urysohn's inequality. It might be possible to derive a better lower bound for $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+)$, which is beyond the scope of this thesis.

Proof of Corollary 7.3.9. By Lemma 2.3.1, we observe that $B_H(\boldsymbol{S}^{n,k}_+) = -\frac{1}{n} B^*_H(\boldsymbol{S}^{n,k}_+)^\circ$. It follows from Lemma 2.3.5 that $w\big(B_H(\boldsymbol{S}^{n,k}_+)\big) \geq \frac{1}{n \cdot w\left(B^*_H(\boldsymbol{S}^{n,k}_+)\right)}$, and therefore,

$$w_G\big(B_H(\boldsymbol{S}^{n,k}_+)\big) \geq \frac{\kappa_d^2}{n \cdot w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big)}$$

where $d = \binom{n+1}{2} - 1$ is the dimension of H. Since $\kappa_d^2 \geq d - \frac{1}{2}$, we obtain for any $0 < \delta < 1$,

$$\lim_{n\to\infty} \frac{w_G\big(B_H(\boldsymbol{S}_+^{n,k})\big)}{\sqrt{2n}} \geq \lim_{n\to\infty} \frac{\sqrt{2n}}{w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big)} \frac{1}{2n^2} \left\{ \binom{n+1}{2} - \frac{3}{2} \right\} = \frac{1}{4f(\delta)}.$$

■

■ 7.4 Approximate Extended Formulations of $\mathbf{S}_+^n$

Now we further extend our discussion beyond the k-PSD approximation. Specifically, we consider an arbitrary approximation of $\boldsymbol{S}_+^n$ through the lens of extended formulations. This defines a much broader class of approximations as we are allowed to introduce as many new variables as we want in the description of an approximating set. However, we show that even in this case, at least superpolynomially many $k \times k$ PSD constraints are required to approximate $\boldsymbol{S}_+^n$ when $k \ll n$.

In Section 7.4.1, we present two main theorems of this chapter about the extension complexity lower bounds that hold for any ϵ-approximation of $B_H(\boldsymbol{S}_+^n)$. Their proofs are deferred until Section 7.6.4 and Section 7.6.5, respectively. In Section 7.4.2, we discuss that these theorems imply the hardness of approximating a large-scale PSD cone in a stronger sense than the results from Section 7.3 as they apply to *any* constructions of approximating cones.

■ 7.4.1 Theorem Statements

Recall that $B_H\left(\boldsymbol{S}_+^n\right) := \boldsymbol{S}_+^n \cap H - \frac{1}{n} I_n$. In this section, we present two main theorems on the $\boldsymbol{S}_+^k$-extension complexity lower bounds that hold for any approximating set of $B_H\left(\boldsymbol{S}_+^n\right)$. These theorems imply the hardness of approximating $B_H\left(\boldsymbol{S}_+^n\right)$ with a small number of $k \times k$ PSD constraints.

Our first theorem is about an $\boldsymbol{S}_+^k$-extension complexity lower bound that holds for any ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$.

Theorem 7.4.1. *There exists a constant $C > 0$ such that if S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then*

$$\mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \min\left\{ \sqrt{\frac{n}{1+\epsilon}},\ \frac{1}{1+\epsilon}\ \frac{n}{k} \right\} \right).$$

Theorem 7.4.1 suggests that at least $\Omega_n(\exp(\sqrt{n}))$ copies of $\boldsymbol{S}_+^k$ are required to approximate $\boldsymbol{S}_+^n$ when $k = O_n(\sqrt{n})$. When $k = \Omega_n(\sqrt{n})$, this extension complexity lower bound gracefully decreases to 1 as k increases to n. We remark that Theorem 7.4.1 holds for arbitrary k, and thus, extends the result of Fawzi [56, Theorem 1] beyond the special case of $k = 1$. A more formal version of Theorem 7.4.1 and its proof can be found in Section 7.6.4.

Next, we consider the $\boldsymbol{S}_+^k$-extension complexity of an average ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$.

Theorem 7.4.2. *There exists a constant $C > 0$ such that if S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then*

$$\mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3},\ \frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}\right\}\right).$$

Theorem 7.4.2 is a stronger result than Theorem 7.4.1 in the sense that it provides an extension complexity lower bound for a broader range of sets that approximate $B_H\left(\boldsymbol{S}_+^n\right)$. Again, this result subsumes [56, Theorem 2] as a special case for $k = 1$. Specifically, Theorem 7.4.2 states that even if we relax the notion of approximation, we still need at least exponentially many number of $k \times k$ PSD constraints to approximate $\boldsymbol{S}_+^n$ when k is small, more precisely, when k is smaller than $\frac{1}{(1+\epsilon)^2}\frac{n}{\log^2 n}$. A more formal version of Theorem 7.4.2 and its proof can be found in Section 7.6.5.

Lastly, we mention that we do not know whether the extension complexity lower bounds in Theorem 7.4.1 and Theorem 7.4.2 are tight. For example, when $k = 1$, the lower bound from Theorem 7.4.1 reads as $\exp(C\sqrt{n})$, whereas the current best construction has size $\exp(cn)$. We are curious if it could be possible to close this gap between the upper and the lower bounds, either by proving a stronger exponential complexity lower bound or by inventing a clever construction scheme with smaller size, and leave it as an interesting open problem.

■ 7.4.2 Implication: Hardness of Approximating the PSD Cone

Theorem 7.4.1 and Theorem 7.4.2 imply that any convex set that well approximates $B_H\left(\boldsymbol{S}_+^n\right)$ must have $\boldsymbol{S}_+^k$-extension complexity at least superpolynomially large in n for all k much smaller than n. Thus, we conclude that it is impossible to approximate $B_H\left(\boldsymbol{S}_+^n\right)$ using only polynomially many $k \times k$ PSD constraints, for *any* constructions of the approximating set. In contrast, the results discussed in Section 7.3 only apply to the specific construction scheme of the k-PSD approximations.

It is noteworthy that the extension complexity lower bounds have a strong implication about the hardness of approximating $\boldsymbol{S}_+^n$ with $k \times k$ PSD constraints. Because they apply to *any* constructions of the approximating set, the lower bounds completely refute the possibility of *globally* approximating the large PSD cone $\boldsymbol{S}_+^n$ with a few, small-sized $k \times k$ PSD constraints.

Nevertheless, the extension complexity lower bounds do not rule out the possibility of approximating $\boldsymbol{S}_+^n$ *locally*. For example, one may still be able to approximate the feasible set of the original SDP in the direction of the objective C, e.g. by exploiting the problem data, cf. (7.1). Therefore, the results from this section seem to suggest that local and/or problem-data-adaptive techniques are essential to approximate SDPs.

■ 7.5 Discussion

Here we make a few comments on the results in this chapter and some related work.

- Blekherman et al. [21] also investigated the question of how well $\boldsymbol{S}_+^{n,k}$ approximates $\boldsymbol{S}_+^n$. They use the quantity $\overline{\operatorname{dist}}_F(\boldsymbol{S}_+^{n,k}, \boldsymbol{S}_+^n) := \sup_{X \in \boldsymbol{S}_+^{n,k}, \|X\|_F = 1} \inf_{Y \in \boldsymbol{S}_+^n} \|X - Y\|_F$ to measure the quality of approximation, and thus, their result has a connection with our result on ϵ-approximation. In this thesis, we extend the scope of the question in two directions: first, we consider the 'average' distance with the notion of average ϵ-approximation as well as the maximal distance; second, our result (Theorem 7.3.2) applies to not only $\boldsymbol{S}_+^{n,k}$, but also $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ with an arbitrary collection of k-dimensional subspaces $\mathcal{V}$.

- Fawzi [56] showed that any polytope that well approximates $B_H\left(\boldsymbol{S}_+^n\right)$ must have LP extension complexity at least exponentially large in n. Our Theorems 7.4.1 and 7.4.2 generalize their results beyond the special case of $k = 1$. We refine and adapt the ideas from [56] to prove a lower bound for arbitrary k. Specifically, we devise a different way of decomposing the component functions of the $\boldsymbol{S}_+^k$-factorization of the slack matrix into their sharp and flat parts, which allows us to apply Fawzi's argument even when $k > 1$. In addition, we compare the variance of two representations of the slack matrix instead of their tail probabilities to obtain a nontrivial $\boldsymbol{S}_+^k$-extension complexity lower bound even when $k = \Omega_n(\sqrt{n})$. See the proof of Theorem 7.4.1 in Section 7.6.4 for details.

- Here we compare our Theorem 7.4.1 with a back-of-the-envelope calculation based on known results about the LP extension complexity of $B_H\left(\boldsymbol{S}_+^n\right)$. Assume that there is a set S such that $\operatorname{xc}_{\boldsymbol{S}_+^k}(S) = N$ and S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$. On the one hand, each of the N cones of $k \times k$ PSD matrices can be approximated by $\exp(ck)$ facets of linear inequalities, where $c > 0$ is an absolute constant; see Aubrun and Szarek [11, Proposition 10]. Thus, the LP extension complexity of S is at most $N \exp(ck)$. On the other hand, the LP extension complexity of S is at least $\exp(c'\sqrt{n})$; see Fawzi [56, Theorem 1]. Therefore, we get $N \geq \exp(c'\sqrt{n} - ck)$. This lower bound becomes trivial when $k = \Omega_n(\sqrt{n})$. In contrast, the lower bound from Theorem 7.4.1 remains superpolynomial as long as $k = o_n(n/\log n)$.

- We also remark some works that studied lower bounds on the semidefinite extension complexity of polytopes associated with NP-Hard combinatorial problems. Fawzi and Parrilo [57] showed that the $\boldsymbol{S}_+^k$-extension complexity of the correlation polytope, $\operatorname{COR}(n) := \operatorname{conv}\{vv^T : v \in \{0,1\}^n\}$, is exponentially large in n for any fixed constant k. Their proof relies on a combinatorial argument that counts possible sparsity patterns of certain matrices with small PSD rank. Lee, Raghavendra, and Steurer [89] proved a stronger lower

bound on the semidefinite extension complexity of the correlation polytope, based on the notion of low-degree sum-of-squares proof. While these works consider a similar problem to ours, the object of study is different; in this work, we are interested in the approximate semidefinite extension complexity of the *spectrahedron* $B_H(\boldsymbol{S}^n_+)$.

- Let $D_k = (\boldsymbol{S}^{n,k}_+)^* \cap H$. In our analysis, $w_G(D_k)$ turns out to be the expectation of the largest k-sparse eigenvalue of the Gaussian Orthogonal Ensemble (GOE) (divided by $\sqrt{2}$). In this work, we only provide an asymptotic upper bound for $w_G(D_k)$ using Slepian's lemma (Lemma 2.4.9), however, it might be possible to prove a lower bound for $w_G(D_k)$ with tools from random matrix theory.

- We do not know whether our lower bounds in Theorems 7.4.1 and 7.4.2 are tight. We remark that our proof techniques only utilize partial information from the slack matrix, i.e., up to the second moment of the marginal distribution with respect to x. It may be possible to achieve a stronger lower bound by exploiting the full information, e.g. higher-order information or the moments of the joint distribution with respect to (x, y).

- In this work, we consider the question of approximating $\boldsymbol{S}^n_+ \cap H$ and show that at least superpolynomially many $k \times k$ PSD constraints are needed when $k \ll n$. However, if one is allowed to exploit the problem data (namely, $C, (A_i, b_i)_{i=1}^m$), it could be still possible to construct a good approximation F' of the feasible set $F = \{X \in \boldsymbol{S}^n_+ : \langle A_i, X\rangle = b_i\}$ with a smaller number of $\boldsymbol{S}^k_+$ so that the optimality gap $\sup_{X\in F'}\langle C, X\rangle - \sup_{X\in F}\langle C, X\rangle$ is small, as empirically evidenced in [6].

7.6 Proofs

7.6.1 Proof of Theorem 7.3.2

Proof. First of all, due to the translation invariance of the Gaussian width, we have

$$w_G\Big(B^*_H\Big(\boldsymbol{S}^{n,k}_+(\mathcal{V})\Big)\Big) = w_G\Big(\boldsymbol{S}^{n,k}_+(\mathcal{V})^* \cap H - \frac{1}{n}I_n\Big) = w_G\Big(\boldsymbol{S}^{n,k}_+(\mathcal{V})^* \cap H\Big).$$

Next, we let $U_i \in \mathbb{R}^{n\times k}$ be a matrix whose columns form an orthonormal basis of V_i for each $V_i \in \mathcal{V}$. We observe that $\boldsymbol{S}^{n,k}_+(\mathcal{V})^* = \operatorname{cl}\operatorname{cone}\big(\bigcup_{i\in[N]}\{U_i Z U_i^T : Z \in \boldsymbol{S}^k_+\}\big)$ because $(\boldsymbol{S}^k_+)^* = \boldsymbol{S}^k_+$ and $(C_1 \cap C_2)^* = \operatorname{cl}\operatorname{cone}(C_1^* \cup C_2^*)$, cf. the paragraph on duality in Section 2.3. Thus, it follows that $\boldsymbol{S}^{n,k}_+(\mathcal{V})^* \cap H = \operatorname{conv}\Big(\bigcup_{i\in[N]}\{U_i v v^T U_i^T : v \in \mathbb{S}^{k-1}\}\Big)$. Note that

$\langle G, U_i v v^T U_i^T \rangle = \langle U_i^T G U_i, v v^T \rangle$, and therefore,

$$w_G\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})^* \cap H\Big) = \mathbb{E}_G\Big[\sup_{\substack{i\in[N]\\ v\in\mathbb{S}^{n-1}}} \langle U_i^T G U_i, v v^T\rangle\Big] = \mathbb{E}_G\Big[\sup_{i\in[N]} \lambda_1(U_i^T G U_i)\Big]$$
$$\leq \sup_{i\in[N]} \mathbb{E}_G\big[\lambda_1(U_i^T G U_i)\big] + \mathbb{E}_G\Big[\sup_{i\in[N]} \Big(\lambda_1(U_i^T G U_i) - \mathbb{E}_G\big[\lambda_1(U_i^T G U_i)\big]\Big)\Big].$$

Note that for every $i \in [N]$, the random matrix $U_i^T G U_i \in \boldsymbol{S}^k$ has the standard Gaussian distribution in $\boldsymbol{S}^k$. By Lemma 2.4.10, $\mathbb{E}_G\big[\lambda_1(U_i^T G U_i)\big] \leq \sqrt{2k}$. Moreover, the function $G \mapsto \lambda_1(U_i^T G U_i)$ is 1-Lipschitz, and therefore, the random variable $\lambda_1(U_i^T G U_i) - \mathbb{E}_G\big[\lambda_1(U_i^T G U_i)\big]$ is sub-Gaussian with sub-Gaussian parameter 1 by Lemma 2.4.12. Lemma 2.4.21 implies that $\mathbb{E}_G\Big[\sup_{i\in[N]} \big(\lambda_1(U_i^T G U_i) - \mathbb{E}_G\big[\lambda_1(U_i^T G U_i)\big]\big)\Big] \leq \sqrt{2\log N}$.

■

■ 7.6.2 Proof of Corollary 7.3.4

Proof. Suppose that $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. Then, by definition of the dual-average approximation (see Definitions 7.2.6 and 7.2.9),

$$w_G\Big(B_H\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\Big)^\circ\Big) \geq \frac{1}{1+\epsilon} w_G\Big(B_H\big(\boldsymbol{S}_+^n\big)^\circ\Big). \tag{7.11}$$

By Lemma 2.3.1, we have $B_H\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\Big)^\circ = -nB_H^*\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\Big)$ and $B_H\big(\boldsymbol{S}_+^n \cap H\big)^\circ = -nB_H^*\big(\boldsymbol{S}_+^n\big) = -nB_H\big(\boldsymbol{S}_+^n\big)$ because $\boldsymbol{S}_+^n$ is self-dual. Thus, Theorem 7.3.2, combined with the inequality (7.11), implies

$$\frac{1}{1+\epsilon} w_G\Big(B_H\big(\boldsymbol{S}_+^n\big)\Big) \leq w_G\Big(B_H\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})^*\Big)\Big) \leq \sqrt{2k} + \sqrt{2\log|\mathcal{V}|}.$$

Note that this inequality holds if and only if

$$\sqrt{\log|\mathcal{V}|} \geq \frac{1}{\sqrt{2}(1+\epsilon)} w_G\Big(B_H\big(\boldsymbol{S}_+^n\big)\Big) - \sqrt{k},$$

which is again equivalent to

$$|\mathcal{V}| \geq \exp\left[\frac{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)}{\sqrt{2}(1+\epsilon)} - \sqrt{k}\right]_+^2 = \exp\big(n\cdot\varphi(n,k,\epsilon)\big).$$

■

■ 7.6.3 Proof of Theorem 7.3.8

Proof. Fix $k \in \{0, 1, \ldots, n\}$. Let $T = \{u \in \mathbb{R}^n : \|u\|_2 \le 1, \|u\|_0 \le k\}$ and observe that

$$w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big) = w_G\big((\boldsymbol{S}_+^{n,k})^* \cap H\big) = \mathbb{E}_G\bigg[\sup_{u\in T} \big\langle G, uu^T\big\rangle\bigg].$$

We consider a Gaussian process $(X_u)_{u\in T}$ such that $X_u = u^T G u + \gamma$ with G being standard Gaussian in $\boldsymbol{S}^n$ and $\gamma \sim N(0,1)$ independent of G. It is easy to verify that

$$\mathbb{E}_G\bigg[\sup_{u\in T} \big\langle G, uu^T\big\rangle\bigg] = \mathbb{E}_{G,\gamma}\bigg[\sup_{u\in T} \big\{u^T G u + \gamma\big\}\bigg] = \mathbb{E}_{G,\gamma}\bigg[\sup_{u\in T} X_u\bigg].$$

Next, we introduce an instrumental Gaussian process $(Y_u)_{u\in T}$ such that $Y_u = g^T u$ with $g \sim N(0, 2I_n)$. It is easy to check that for all $u, v \in T$, (1) $\mathbb{E}X_u = \mathbb{E}Y_u = 0$; (2) $\mathbb{E}X_u^2 = \mathbb{E}Y_u^2 = 2$; and (3) $\mathbb{E}X_uX_v - \mathbb{E}Y_uY_v = (1 - u^Tv)^2 \ge 0$. Now we can apply Slepian's lemma (Lemma 2.4.9) to obtain $\mathbb{E}_{G,\gamma}\big[\sup_{u\in T} X_u\big] \le \mathbb{E}_{g\sim N(0,2I_n)}\big[\sup_{u\in T} Y_u\big]$. Then it follows that

$$w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big) = \mathbb{E}_{G,\gamma}\bigg[\sup_{u\in T} X_u\bigg] \le \mathbb{E}_{g\sim N(0,2I_n)}\big[\sup_{u\in T} Y_u\big] = \mathbb{E}_{g\sim N(0,2I_n)} \sup_{\substack{u\in\mathbb{R}^n \\ \|u\|_2\le 1, \|u\|_0\le k}} g^T u.$$

Therefore,

$$\frac{1}{\sqrt{2n}} w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big) \le \mathbb{E}_{g\sim N(0,2I_n)}\left[\frac{\|g\|_2}{\sqrt{2n}} \sup_{\substack{u\in\mathbb{R}^n \\ \|u\|_2\le 1, \|u\|_0\le k}} \frac{g^T u}{\|g\|_2}\right].$$

Note that when $g \sim N(0, 2I_n)$, $\frac{\|g\|_2}{\sqrt{2n}} \to 1$ in probability as $n \to \infty$. To compute the expectation on the right-hand side, it suffices to identify the limit of $\sup_{\substack{u\in\mathbb{R}^n \\ \|u\|_2\le 1, \|u\|_0\le k}} \frac{g^T u}{\|g\|_2}$ (in probability).

Given $x \in \mathbb{R}^n$, we let $(x_i^2)^\downarrow$ denote the i-th largest element in the set $\{x_1^2, x_2^2, \ldots, x_n^2\}$. Observe that

$$\sup_{\substack{u\in\mathbb{R}^n \\ \|u\|_2\le 1, \|u\|_0\le k}} \frac{g^T u}{\|g\|_2} = \frac{1}{\|g\|_2} \frac{\sum_{i=1}^k (g_i^2)^\downarrow}{\sqrt{\sum_{i=1}^k (g_i^2)^\downarrow}} = \left(\frac{1}{\|g\|_2^2} \sum_{i=1}^k (g_i^2)^\downarrow\right)^{1/2}$$

and that $(g_1^2)^\downarrow \ge (g_2^2)^\downarrow \ge \cdots \ge (g_n^2)^\downarrow$ are χ^2 order statistics of degree 1, multiplied by a factor of 2. It is well known from literature on extreme order statistics (e.g., [109, Theorem 2.7]) that for any fixed $0 < \delta < 1$,

$$\frac{1}{\|g\|_2^2} \sum_{i=1}^{\lfloor \delta n \rfloor} (g_i^2)^\downarrow \longrightarrow \int_0^\delta Q_{\chi^2}(1-s)ds \qquad \text{in probability as } n \to \infty.$$

Combining these observations and the well-known fact that $\lim_{n\to\infty} \frac{w_G(B_H^*(\boldsymbol{S}_+^n))}{\sqrt{2n}} = 1$, cf. Remark 7.2.10, we obtain the desired inequality:

$$\lim_{n\to\infty} \frac{w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big)}{w_G\big(B_H^*(\boldsymbol{S}_+^n)\big)} = \lim_{n\to\infty} \frac{\sqrt{2n}}{w_G\big(B_H^*(\boldsymbol{S}_+^n)\big)} \lim_{n\to\infty} \frac{w_G\big(B_H^*(\boldsymbol{S}_+^{n,k})\big)}{\sqrt{2n}} \le \bigg(\int_0^\delta Q_{\chi^2}(1-s)ds \bigg)^{1/2}.$$

We conclude the proof by computing the integral in the upper bound. An explicit formula for the integral is well known; see [109, Remark 2.8], for example.

$$\int_0^\delta Q_{\chi^2}(1-s)ds = 2\int_{\Phi^{-1}(1-\frac{\delta}{2})}^\infty s^2\Phi'(s)ds = \delta + \sqrt{\frac{2}{\pi}}\Phi^{-1}\Big(1-\frac{\delta}{2}\Big)\exp\bigg(-\frac{1}{2}\Big[\Phi^{-1}\Big(1-\frac{\delta}{2}\Big)\Big]^2\bigg).$$

■

■ 7.6.4 Proof of Theorem 7.4.1

Let $c = \max\big\{\sqrt{c_1/2\log 3},\ \sqrt{2}c_2\big\}$ denote an absolute constant where $c_1, c_2 > 0$ are the constants that appear in Lemma 2.4.20 (MGF of sub-Gaussian chaos of order 2, also known as Hanson-Wright inequality). We state a full version of Theorem 7.4.1 as follows.

Theorem 7.6.1. *If S is an ϵ-approximation of $B_H\big(\boldsymbol{S}_+^n\big)$, then for all positive integer $1 \le k \le n$,*

$$\log \mathrm{xc}_{\boldsymbol{S}_+^k}(S) \ge -\frac{\alpha+\beta}{2} + \sqrt{\Big(\frac{\alpha-\beta}{2}\Big)^2 + \gamma}$$

where

$$\alpha = 2k\log 3, \qquad \beta = \log\Big(\frac{n^3}{8\cdot c\log 3}\Big), \qquad \gamma = \frac{1}{2e\cdot c}\frac{n-1}{(1+\epsilon)}.$$

Now we discuss how Theorem 7.4.1 can be derived from Theorem 7.6.1. Suppose that n is sufficiently large, tending to infinity.

- When $k = o_n(\sqrt{\frac{n}{1+\epsilon}})$, we observe that $\gamma \gg \max\{\alpha^2, \beta^2\}$, and therefore, $-\frac{\alpha+\beta}{2} + \big\{(\frac{\alpha-\beta}{2})^2 + \gamma\big\}^{1/2} \approx \sqrt{\gamma}$.
- When $k = \omega_n(\sqrt{\frac{n}{1+\epsilon}})$, $\alpha \gg \max\{\beta, \sqrt{\gamma}\}$. Thus, $\big\{(\frac{\alpha-\beta}{2})^2 + \gamma\big\}^{1/2} \approx \frac{\alpha}{2}\big(1+\frac{4\gamma}{\alpha^2}\big)^{1/2} \approx \frac{\alpha}{2}\big(1+\frac{2\gamma}{\alpha^2}\big)$. As a result, $-\frac{\alpha+\beta}{2} + \big\{(\frac{\alpha-\beta}{2})^2 + \gamma\big\}^{1/2} \approx \frac{\gamma}{\alpha}$.

In the rest of this section, we prove Theorem 7.6.1. Our proof is based on similar arguments to those found in the proof of Fawzi [56, Theorem 1], but with appropriate adaptations. Indeed, our results can be seen as an extension of Fawzi's beyond the special case with

$k = 1$, which is made possible by introducing different notions of normalization, (7.16), and decomposition of $\boldsymbol{S}_+^k$-factors into sharp and flat components, (7.17).

Proof of Theorem 7.6.1. We begin this proof with a rough sketch of the main ideas used in the proof. First, we consider the generalized slack matrix s of the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ (1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)$ restricted to the hypercube H_n. In light of the generalized Yannakakis theorem (Lemma 6.2.2), the $\boldsymbol{S}_+^k$-extension complexity of S is bounded from below by the $\boldsymbol{S}_+^k$-rank of the slack matrix s, cf. (6.4). Thus, it suffices to prove a lower bound for $\operatorname{rank}_{\boldsymbol{S}_+^k}(s)$.

To this end, we express the slack matrix s in two different ways: one obtained from the definition of the slack operator using the knowledge about the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$, and the other obtained by assuming that s admits a $\boldsymbol{S}_+^k$-factorization having N factors. Interpreting the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$ and $(1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)$ as formal variables, x and y, we may view the two expressions of the slack matrix as bivariate polynomials. Next, we 'smooth out' the two expressions with respect to one variable, x, by taking projection onto the harmonic subspace of degree 2; and then take expectation with respect to the other variable, y. Comparing the two resulting expressions, we derive a lower bound on the number of factors N, which implies a lower bound on the $\boldsymbol{S}_+^k$-extension complexity of S.

Step 1. Slack Matrix and $\boldsymbol{S}_+^k$-Factorization We consider the (generalized) slack operator associated to the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ (1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)$. Let $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. Observe that the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$ are $\tilde{x}\tilde{x}^T - \frac{1}{n}I_n$ for $\tilde{x} \in \mathbb{S}^{n-1}$, and that $\left((1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)^\circ = -\frac{n}{1+\epsilon}B_H\left(\boldsymbol{S}_+^n\right)$. Thus, we are led to study the following infinite matrix:

$$(\tilde{x},\tilde{y}) \in \mathbb{S}^{n-1}\times\mathbb{S}^{n-1} \mapsto 1 - \left\langle \tilde{x}\tilde{x}^T - \frac{1}{n}I_n, -\frac{n}{1+\epsilon}\left(\tilde{y}\tilde{y}^T - \frac{1}{n}I_n\right)\right\rangle = \frac{n}{1+\epsilon}(\tilde{x}^T\tilde{y})^2 + \frac{\epsilon}{1+\epsilon}.$$

We consider the PSD rank ($\boldsymbol{S}_+^k$-rank) of the finite submatrix restricted to $\tilde{x},\tilde{y} \in \left\{-\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right\}^n \subset \mathbb{S}^{n-1}$. Specifically, we consider the following matrix s defined on the n-dimensional hypercube, with a proper reparametrization ($x = \sqrt{n}\tilde{x}$ and $y = \sqrt{n}\tilde{y}$):

$$s : (x,y) \in \{-1,1\}^n \times \{-1,1\}^n \mapsto \frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2 + \epsilon\right). \tag{7.12}$$

Assuming that we can write the matrix (7.12) as a sum of N trace inner products of $\boldsymbol{S}_+^k$ factors, we have

$$\frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2 + \epsilon\right) = s(x,y) = \sum_{i=1}^N \langle f_i(x), g_i(y)\rangle, \qquad \forall x,y \in H_n \tag{7.13}$$

where $f_i, g_i : H_n \to \boldsymbol{S}_+^k$ are some matrix-valued functions on H_n.

In this proof, we use the two expressions of $s(x, y)$ in (7.13) to derive a lower bound on N. First, we fix $y \in H_n$ and 'smooth out' the expressions on both sides of (7.13) with respect to x by taking projection onto the space of harmonic polynomials of degree 2. Then we plug $x = y$ and consider the expectation of the smoothed functions with respect to $y \in H_n$.

More precisely, for each fixed $y \in H_n$, we let $q_y(x) = (x^T y)^2 - n$. Also, let μ denote the uniform probability measure on H_n. The inner product of any two functions $f, g : H_n \to \mathbb{R}$ is defined as $\langle f, g\rangle_\mu = \mathbb{E}_{x\sim\mu}\big[f(x)g(x)\big]$. We observe that $\langle f(x), q_y(x)\rangle_\mu = 2\ \mathrm{proj}_2 f(y)$.

Taking the inner product of both sides of (7.13) with $q_y(x)$, we obtain

$$\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n} q_y(x)^2 + q_y(x)\right] = \sum_{i=1}^{N} \left\langle \mathbb{E}_{x\sim\mu}\big[q_y(x) f_i(x)\big], g_i(y)\right\rangle .$$

Subsequently, we get the following equation by taking expectation over $y \sim \mu$:

$$\underbrace{\mathbb{E}_{y\sim\mu}\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n} q_y(x)^2 + q_y(x)\right]}_{=:LHS} = \underbrace{\mathbb{E}_{y\sim\mu}\sum_{i=1}^{N} \left\langle \mathbb{E}_{x\sim\mu}\big[q_y(x) f_i(x)\big], g_i(y)\right\rangle}_{=:RHS} . \tag{7.14}$$

The rest of the proof is organized as follows. In Step 2, we compute the expectation on the left-hand side exactly. In Step 3, we derive an upper bound on the expectation on the right-hand side as a function of N. In the end, we obtain the desired lower bound on N in Step 4 by comparing these two quantities.

Step 2. The Left-hand Side of (7.14). We evaluate the left-hand side of (7.14) based on the following observations:

$$\begin{aligned}
\mathbb{E}_{x\sim\mu}\big[(x^T y)^2\big] &= \mathbb{E}_{x\sim\mu}\left[\left(\sum_{i=1}^{n} x_i y_i\right)^2\right] = \sum_{i,j=1}^{n} y_i y_j \mathbb{E}_{x\sim\mu}[x_i x_j] = \sum_{i=1}^{n} \mathbb{E}_{x\sim\mu}[x_i^2] \\
&= n, \\
\mathbb{E}_{x\sim\mu}\Big[(x^T y)^4\Big] &= \mathbb{E}_{x\sim\mu}\left[\left(\sum_{i=1}^{n} x_i y_i\right)^4\right] = \sum_{i=1}^{n} \mathbb{E}_{x\sim\mu}[x_i^4] + 3\sum_{\substack{i=1\\ j\neq i}}^{n} \mathbb{E}_{x\sim\mu}[x_i^2]\cdot\mathbb{E}_{x\sim\mu}[x_j^2] \\
&= n + 3n(n-1).
\end{aligned}$$

Therefore, $\mathbb{E}_{x\sim\mu}[q_y(x)] = \mathbb{E}_{x\sim\mu}\big[(x^T y)^2 - n\big] = 0$ and $\mathbb{E}_{x\sim\mu}[q_y(x)^2] = \mathbb{E}_{x\sim\mu}\big[(x^T y)^4 - 2n(x^T y)^2 + n^2\big] = 2n(n-1)$. It follows that for any $y \in H_n$,

$$\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n} q_y(x)^2 + q_y(x)\right] = \frac{2}{1+\epsilon}(n-1). \tag{7.15}$$

This does not depend on y, and therefore, LHS in (7.14) $= \frac{2}{1+\epsilon}(n-1)$.

Step 3. An Upper Bound for the Right-hand Side of (7.14). Now, we prove an upper bound on the right-hand side of (7.14), which has the form of an increasing function of N. This step is the most technical part of the proof, and is composed of four mini-steps.

First of all, we claim that we may assume without loss of generality that the factor functions f_i, g_i satisfy

$$\left\|\mathbb{E}_{x\sim\mu}[f_i(x)]\right\|_{op} = 1, \quad \forall i \in [N] \qquad \text{and} \qquad \sum_{i=1}^{N} \mathrm{Tr}\big(g_i(y)\big) = 1, \quad \forall y \in H_n. \tag{7.16}$$

Next, in Step 3-B, we decompose each f_i into its sharp component $f_i^\sharp$ and flat component $f_i^\flat$ with a fixed threshold $\Lambda \geq e$ whose value will be determined later in Step 4 of the proof; see (7.17).Then, we observe that $RHS = 2\big[\mathbb{E}_y \sum_{i=1}^N \mathrm{proj}_2 f_i^\sharp(y) g_i(y) + \mathbb{E}_y \sum_{i=1}^N \mathrm{proj}_2 f_i^\flat(y) g_i(y)\big]$ due to linearity of expectation. Lastly, we prove upper bounds for the two terms separately in Step 3-C and Step 3-D.

The key idea is that for all $i \in [N]$, $f_i^\sharp$ is supported only on a set of small measure due to the normalization, and $\|\mathrm{proj}_2(v^T f_i^\flat v)\|_2 \leq e \log \Lambda$ for all $v \in \mathbb{S}^{k-1}$ due to hypercontractivity (Lemma 6.2.4).

Step 3-A: Normalization of Factor Functions f_i, g_i We claim that if $s(x,y)$ admits a $(\boldsymbol{S}_+^k)^N$-factorization, then we may assume (7.16) without loss of generality. More precisely, we show it is possible to normalize arbitrary factor functions $\{(\tilde{f}_i, \tilde{g}_i)\}_{i=1}^N$ to $\{(f_i, g_i)\}_{i=1}^N$ so that the conditions in (7.16) are satisfied.

Suppose that $s(x,y)$, defined in (7.12), admits a factorization $\{(\tilde{f}_i, \tilde{g}_i)\}_{i=1}^N$ such that $\tilde{f}_i, \tilde{g}_i : H_n \to \boldsymbol{S}_+^k$ and $s(x,y) = \sum_{i=1}^N \left\langle \tilde{f}_i(x), \tilde{g}_i(y) \right\rangle$ for all $x, y \in H_n$. For each $i \in [N]$, we can see that $\mathbb{E}_x[\tilde{f}_i(x)] \in \boldsymbol{S}_+^k$, and therefore, we may define $W_i = \mathbb{E}_x[\tilde{f}_i(x)]^{1/2}$ to be the principal square root of $\mathbb{E}_x[\tilde{f}_i(x)]$. Let $W_i^\dagger$ denote the Moore-Penrose pseudoinverse of W_i.

Now for each $i \in [N]$, we let $f_i(x) = W_i^\dagger \tilde{f}_i(x) W_i^\dagger$ and $g_i(y) = W_i \tilde{g}_i(y) W_i$. It is easy to verify that $\langle f_i(x), g_i(y)\rangle = \mathrm{Tr}\big(W_i^\dagger \tilde{f}_i(x) W_i^\dagger W_i \tilde{g}_i(y) W_i\big) = \mathrm{Tr}\big(\tilde{f}_i(x)\tilde{g}_i(y)\big) = \big\langle \tilde{f}_i(x),\ \tilde{g}_i(y)\big\rangle$ for all $x, y \in H_n$. Therefore, $\{(f_i, g_i)\}_{i=1}^N$ also constitutes a valid $\boldsymbol{S}_+^k$-factorization of $s(x,y)$.

It remains to check if $\{(f_i, g_i)\}_{i=1}^N$ satisfies (7.16). First, we can easily observe that

$$\mathbb{E}_{x\sim\mu}[f_i(x)] = W_i^\dagger \mathbb{E}_{x\sim\mu}[\tilde{f}_i(x)] W_i^\dagger = W_i^\dagger W_i^2 W_i^\dagger = \Pi_{\mathcal{R}(W_i)}$$

where $\mathcal{R}(W_i)$ is the range of W_i and $\Pi_{\mathcal{R}(W_i)}$ is the projection matrix onto $\mathcal{R}(W_i)$. Thus, $\left\|\mathbb{E}_{x\sim\mu}[f_i(x)]\right\|_{op} = \|\Pi_{\mathcal{R}(W_i)}\|_{op} = 1$. Next, we revisit (7.13), fix any $y \in H_n$, and take expectation with respect to $x \sim \mu$. On the left-hand side, we obtain $\mathbb{E}_{x\sim\mu}\big[\frac{1}{1+\epsilon}\big(\frac{1}{n}(x^T y)^2 + \epsilon\big)\big] =$

1 because $\mathbb{E}_{x\sim\mu}\big[(x^Ty)^2\big] = n$ (see Step 2). On the right-hand side, we have

$$\mathbb{E}_{x\sim\mu}\sum_{i=1}^N \langle f_i(x), g_i(y)\rangle = \sum_{i=1}^N \langle \mathbb{E}_{x\sim\mu}[f_i(x)], g_i(y)\rangle = \sum_{i=1}^N \big\langle \Pi_{\mathcal{R}(W_i)}, g_i(y)\big\rangle = \sum_{i=1}^N \operatorname{Tr}\, g_i(y)$$

because $\Pi_{\mathcal{R}(W_i)} g_i(y) = g_i(y)$ for all $y \in H_n$, by definition of g_i. Therefore, $\sum_{i=1}^N \operatorname{Tr}\, g_i(y) = 1$ for all $y \in H_n$.

Step 3-B: Decomposition of f_i We decompose each f_i into its 'sharp' (spiky) component $f_i^\sharp$ and the 'flat' component $f_i^\flat$ using a fixed threshold Λ whose value will be determined later in Step 4 of the proof. To be specific, for each $i \in [N]$, we define the component functions $f_i^\sharp, f_i^\flat : H_n \to \boldsymbol{S}_+^k$ as follows. Given $x \in H_n$, let $f_i(x) = \sum_{a=1}^k \lambda_a u_a u_a^T$ be the eigendecomposition of $f_i(x)$. Then we let

$$f_i^\sharp(x) = \sum_{a=1}^k \lambda_a \mathbf{1}_{\{\lambda_a > \Lambda\}} u_a u_a^T, \qquad \text{and} \qquad f_i^\flat(x) = \sum_{a=1}^k \lambda_a \mathbf{1}_{\{\lambda_a \le \Lambda\}} u_a u_a^T. \tag{7.17}$$

Observe that $f_i = f_i^\sharp + f_i^\flat$ and $\big\langle f_i^\sharp(x),\ f_i^\flat(x)\big\rangle = \operatorname{Tr}\big(f_i^\sharp(x) f_i^\flat(x)\big) = 0$ for all $x \in H_n$. From now on, we may refer to $f_i^\sharp$ ($f_i^\flat$, resp.) as the sharp component (flat component, resp.) of f_i.

By linearity of expectation, we can decompose the expression on the right-hand side of (7.14) as follows:

$$RHS \text{ in } (7.14) = \mathbb{E}_{y\sim\mu}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\sharp(x)], g_i(y)\Big\rangle + \mathbb{E}_{y\sim\mu}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\flat(x)], g_i(y)\Big\rangle. \tag{7.18}$$

Step 3-C. Upper Bound on the Contribution of Sharp Components in (7.18) In this paragraph, we argue that the first term on the right hand side of (7.18) is bounded from above by $N\frac{k}{\Lambda} n^3$. Our argument is based on the following three observations.

- Let $\operatorname{supp} f_i^\sharp = \{x \in H_n : f_i^\sharp(x) \neq 0\}$. Then $|\operatorname{supp} f_i^\sharp| < \frac{k}{\Lambda}\, 2^n$ for all $i \in [N]$. It is because (i) $\operatorname{Tr}\, \mathbb{E}_{x\sim\mu}[f_i^\sharp(x)] \le \operatorname{Tr}\, \mathbb{E}_{x\sim\mu}[f_i(x)] \le k \left\|\mathbb{E}_{x\sim\mu}[f_i(x)]\right\|_{op} = k$, cf. (7.16); (ii) $\operatorname{Tr}\, \mathbb{E}_{x\sim\mu}[f_i^\sharp(x)] = \frac{1}{2^n}\sum_{x\in H_n} \operatorname{Tr}\, f_i^\sharp(x)$; and (iii) $\operatorname{Tr}\, f_i^\sharp(x) > \Lambda$ for all $x \in \operatorname{supp} f_i^\sharp$ by definition of $f_i^\sharp$.

- For each $i \in [N]$, $\big\langle f_i^\sharp(x),\ g_i(y)\big\rangle \le n$ for all $x, y \in H_n$. This follows from Eq. (7.13) because

$$\Big\langle f_i^\sharp(x), g_i(y)\Big\rangle \le \sum_{i=1}^N \langle f_i(x), g_i(y)\rangle = \frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2 + \epsilon\right) \le \frac{1}{1+\epsilon}(n+\epsilon) \le n.$$

- Lastly, $|q_y(x)| \le n(n-1)$ for all $x, y \in H_n$ because $q_y(x) = (x^T y)^2 - n \le n(n-1)$ and $q_y(x) \ge -n$.

Combining the three observations above, we can see that for any $y \in H_n$,

$$\begin{aligned}
\sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\sharp(x)], g_i(y) \right\rangle &= \sum_{i=1}^N \mathbb{E}_{x\sim\mu}\Big[q_y(x) \left\langle f_i^\sharp(x), g_i(y) \right\rangle \Big] \\
&= \sum_{i=1}^N \sum_{x \in \operatorname{supp} f_i^\sharp} \frac{1}{2^n} q_y(x) \left\langle f_i^\sharp(x), g_i(y) \right\rangle \\
&\le \sum_{i=1}^N \frac{|\operatorname{supp} f_i^\sharp|}{2^n} \Big(\max_{x,y\in H_n} |q_y(x)| \Big) \Big(\max_{x,y\in H_n} \left\langle f_i^\sharp(x), g_i(y) \right\rangle \Big) \\
&\le \frac{k}{\Lambda} n^2 (n-1) N.
\end{aligned}$$

Taking expectation with respect to $y \sim \mu$, we obtain

$$\mathbb{E}_{y\sim\mu} \sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\sharp(x)], g_i(y) \right\rangle \le \frac{k}{\Lambda} n^2 (n-1) N \le \frac{k}{\Lambda} n^3 N. \tag{7.19}$$

Step 3-D. Upper Bound on the Contribution of Flat Components in (7.18) Next, we prove an upper bound for the second term on the right hand side of (7.18). Our proof is based on the concentration of the degree-2 harmonic components of bounded functions and the usual ϵ-net argument.

First, we reduce the matrix-valued function $f_i^\flat$'s to the supremum of multiple scalar-valued functions indexed over a finite set. Given $\epsilon_{\text{net}} > 0$, let $\mathcal{N}$ be an ϵ_{net}-net of $\mathbb{S}^{k-1}$ with the smallest possible cardinality. Note that $|\mathcal{N}| \le \big(1 + \frac{2}{\epsilon_{\text{net}}}\big)^k$ by the well-known upper bound on the ϵ_{net}-covering number of $\mathbb{S}^{k-1}$; see Lemma 2.2.6. Then

$$\begin{aligned}
\mathbb{E}_{y\sim\mu}\left[\sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\flat(x)], g_i(y) \right\rangle \right] &\overset{(a)}{\le} \mathbb{E}_{y\sim\mu}\left[\sum_{i=1}^N \Big\| \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\flat(x)] \Big\|_{op} \operatorname{Tr}\, g_i(y) \right] \\
&\overset{(b)}{\le} \mathbb{E}_{y\sim\mu}\left[\max_{i\in[N]} \Big\| \mathbb{E}_{x\sim\mu}[q_y(x) f_i^\flat(x)] \Big\|_{op} \right] \\
&\overset{(c)}{\le} \frac{1}{1-2\epsilon_{\text{net}}} \mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N] \\ v\in\mathcal{N}}} \Big| v^T \mathbb{E}_{x\sim\mu}\big[q_y(x) f_i^\flat(x)\big] v \Big| \right] \\
&= \frac{1}{1-2\epsilon_{\text{net}}} \mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N] \\ v\in\mathcal{N}}} \Big| \mathbb{E}_{x\sim\mu}\big[q_y(x)\ v^T f_i^\flat(x) v\big] \Big| \right].
\end{aligned}$$

In the above lines, (a) follows from Cauchy-Schwarz inequality; (b) is due to the normalization

$\sum_{i=1}^N \operatorname{Tr} g_i(y) \equiv 1$; and (c) is obtained by the ϵ-net argument, i.e., if $\mathcal{N}$ is an ϵ_{net}-net of $\mathbb{S}^{k-1}$, then for any $M \in \boldsymbol{S}_+^k$, $\|M\|_{op} \le \frac{1}{1-2\epsilon_{\text{net}}} \sup_{v\in\mathcal{N}} v^T M v$. Now it remains to evaluate the expectation in the last line.

Recall that $\mathbb{E}_{x\sim\mu}\big[q_y(x) v^T f_i^\flat(x) v\big] = \big\langle q_y(x),\ v^T f_i^\flat(x) v\big\rangle_\mu = 2\mathrm{proj}_2\big(v^T f_i^\flat v\big)(y)$. We observe that for each $(i,v) \in [N]\times\mathcal{N}$, the derived random variable $\mathrm{proj}_2\big(v^T f_i^\flat v\big)(y)$ is sub-exponential with parameters $(c_1\|\mathrm{proj}_2(v^T f_i^\flat v)\|_2^2/2,\ c_2\|\mathrm{proj}_2(v^T f_i^\flat v)\|_2/\sqrt{2})$, due to Lemma 6.2.5. Here, $c_1, c_2 > 0$ are the same absolute constants that appear in Lemma 2.4.20.

Next, we find a common upper bound on $\|\mathrm{proj}_2(v^T f_i^\flat v)\|_2$ that holds for all (i,v). Note that for all (i,v), $\mathbb{E}_{y\in\mu}[v^T f_i^\flat(y) v] \le \mathbb{E}_{y\in\mu}\|f_i^\flat(y)\|_{op} \le \mathbb{E}_{y\in\mu}\|f_i(y)\|_{op} = 1$ due to the normalization in (7.16), and $0 \le v^T f_i^\flat v \le \Lambda$ by definition of $f_i^\flat$. Thus, we can apply Lemma 6.2.4 to get $\|\mathrm{proj}_2(v^T f_i^\flat v)\|_2 \le e\log\Lambda$ for all (i,v), provided that we will choose the threshold $\Lambda \ge e$.

Now we can use a result on the expected maximum of $N|\mathcal{N}|$ sub-exponential random variables (Lemma 2.4.21) to obtain

$$\begin{aligned}
\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|\mathbb{E}_{x\sim\mu}\big[q_y(x)\ v^T f_i^\flat(x) v\big]\right|\right] &= 2\cdot\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|\mathrm{proj}_2\big(v^T f_i^\flat v\big)(y)\right|\right] \\
&\le 2\cdot\frac{e\log\Lambda}{\sqrt{2}}\max\left\{\sqrt{2c_1\log(N|\mathcal{N}|)},\ 2c_2\log(N|\mathcal{N}|)\right\} \\
&\le 2e\log\Lambda\cdot\max\left\{\sqrt{c_1\log(N|\mathcal{N}|)},\ \sqrt{2}c_2\log(N|\mathcal{N}|)\right\}.
\end{aligned}$$

Collecting the pieces in this step, we obtain the following upper bound:

$$\begin{aligned}
&\mathbb{E}_{y\sim\mu}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\in H_n}[q_y(x) f_i^\flat(x)], g_i(y)\Big\rangle \\
&\quad\le \frac{2e\log\Lambda}{1-2\epsilon_{\text{net}}}\max\left\{\sqrt{c_1\log\left[N\Big(1+\frac{2}{\epsilon_{\text{net}}}\Big)^k\right]},\ \sqrt{2}c_2\log\left[N\Big(1+\frac{2}{\epsilon_{\text{net}}}\Big)^k\right]\right\}.
\end{aligned} \tag{7.20}$$

Step 4. Concluding the Proof Lastly, we revisit Eq. (7.14) to conclude the proof. Recall that we obtained the value of the left-hand side in Step 2, cf. (7.15), and derived an upper bound for the right-hand side in Step 3, cf. (7.18), (7.19), and (7.20). Putting these together, we have the following inequality that holds for any choice of parameters ϵ_{net}, Λ such that $0 < \epsilon_{\text{net}} < \frac{1}{2}$ and $\Lambda \ge e$:

$$\frac{2}{1+\epsilon}(n-1) \le \frac{k}{\Lambda}n^3 N + \frac{2e\log\Lambda}{1-2\epsilon_{\text{net}}}\max\left\{\sqrt{c_1\log\left[N\Big(1+\frac{2}{\epsilon_{\text{net}}}\Big)^k\right]},\ \sqrt{2}c_2\log\left[N\Big(1+\frac{2}{\epsilon_{\text{net}}}\Big)^k\right]\right\}. \tag{7.21}$$

We choose $\epsilon_{\text{net}} = 1/4$ for simplicity because optimizing ϵ_{net} does not make much difference.

Observe that $\log(9^k N) \geq 2\log 3$ for all $k, N \geq 1$. Thus, $\sqrt{c_1 \log(9^k N)} \leq \sqrt{c_1/2\log 3}\log(9^k N)$ for all $k, N \geq 1$. Therefore, $\max\{\sqrt{c_1 \log(9^k N)},\ \sqrt{2}c_2 \log(9^k N)\} \leq c\log(9^k N)$ where $c = \max\{\sqrt{c_1/2\log 3},\ \sqrt{2}c_2\}$.

Then, we select Λ that minimizes the right-hand side of (7.21). It is easy to see that the upper bound is minimized (w.r.t. Λ) at $\Lambda^* = \frac{kn^3 N}{4ec\log(9^k N)}$. Noticing that $\Lambda^* \leq \frac{kn^3 N}{4ec\log(9^k)}$ (because $N \geq 1$), we get the following quadratic inequality in $\log N$ as a necessary condition for (7.21):

$$\begin{aligned}\frac{2}{1+\epsilon}(n-1) &\leq 4e\cdot c\log(9^k N)\big(1+\log\Lambda^*\big)\\ &\leq 4e\cdot c\big[\log N + 2k\log 3\big]\left[\log N + \log\left(\frac{n^3}{8\cdot c\log 3}\right)\right]. \end{aligned} \tag{7.22}$$

Letting $z = \log N \geq 0$, we note that (7.22) is a quadratic inequality of the form $(z+\alpha)(z+\beta) \geq \gamma$ where

$$\alpha = 2k\log 3, \qquad \beta = \log\left(\frac{n^3}{8\cdot c\log 3}\right), \qquad \gamma = \frac{1}{2e\cdot c}\frac{n-1}{(1+\epsilon)}.$$

We want to solve this quadratic inequality with an implicit constraint $z \geq 0$ because $N \geq 1$. Observe that its discriminant $D = (\alpha-\beta)^2 + 4\gamma > 0$, regardless of n, k, ϵ. Therefore, the set of solutions is given as $\big\{z \in \mathbb{R} :\ (z+\alpha)(z+\beta) \geq \gamma,\ z \geq 0\big\} = \big\{z \in \mathbb{R} :\ z \geq \big[\frac{-(\alpha+\beta)+\sqrt{D}}{2}\big]_+\big\}$ where $[x]_+ = \max\{x, 0\}$. ■

■ 7.6.5 Proof of Theorem 7.4.2

The following is a formal version of Theorem 7.4.2, which will be proved later in this section.

Theorem 7.6.2. *If S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then for all positive integer $1 \leq k \leq n$,*

$$\log \mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq \left\{\left(\alpha + \sqrt{\alpha^2+\beta^3}\right)^{1/3} + \left(\alpha - \sqrt{\alpha^2+\beta^3}\right)^{1/3}\right\}^2 - 2k\log 3$$

where

$$\alpha = \frac{\sqrt{n}}{22000e(1+\epsilon)} \qquad \textit{and} \qquad \beta = \frac{1}{3}\left\{\log\left(\frac{16(1+\epsilon)k^{1/2}n^{3/2}}{5\sqrt{2\log 3}}\right) - 2k\log 3\right\}.$$

Now we discuss how Theorem 7.4.2 can be derived from Theorem 7.6.2. For notational brevity in our derivation, we let $T_+ = \left(\alpha + \sqrt{\alpha^2+\beta^3}\right)^{1/3}$ and $T_- = \left(\alpha - \sqrt{\alpha^2+\beta^3}\right)^{1/3}$. Suppose that n is sufficiently large, tending to infinity.

- When $k = o_n\big(\big(\frac{n}{(1+\epsilon)^2}\big)^{1/3}\big)$, we can see that $\alpha^2 \gg |\beta|^3$ and thus, $\sqrt{\alpha^2+\beta^3} \approx \alpha$. Therefore, $T_+ + T_- \approx (2\alpha)^{1/3}$, and in the end, $(T_+ + T_-)^2 - 2k\log 3 \approx (2\alpha)^{2/3} \approx C \cdot \frac{n^{1/3}}{(1+\epsilon)^{2/3}}$ for some constant C.

- When $k = \omega_n\big(\big(\frac{n}{(1+\epsilon)^2}\big)^{1/3}\big)$, note that $\beta < 0$ and $\alpha^2 \ll |\beta|^3$. Let $\gamma := \sqrt{\alpha^2+\beta^3}$. We observe that $\gamma \approx |\beta|^{3/2}i$, and thus, $|\gamma| \approx |\beta|^{3/2} \gg \alpha$. Then, we can see that $T_+ = (\alpha+\gamma)^{1/3} \approx \gamma^{1/3}\big(1+\frac{\alpha}{3\gamma}\big)$, and likewise, $T_- \approx \bar{\gamma}^{1/3}\big(1+\frac{\alpha}{3\bar{\gamma}}\big)$ where $\bar{\gamma}$ is the complex conjugate of γ. Then it follows that

$$\begin{aligned} T_+ + T_- &\approx \gamma^{1/3} + \bar{\gamma}^{1/3} + \frac{\alpha}{3}\Big(\frac{1}{\gamma^{2/3}} + \frac{1}{\bar{\gamma}^{2/3}}\Big) \approx |\gamma|^{1/3}\big(e^{i\frac{\pi}{6}} + e^{-i\frac{\pi}{6}}\big) + \frac{\alpha}{3|\gamma|^{2/3}}\big(e^{-i\frac{\pi}{3}} + e^{i\frac{\pi}{3}}\big) \\ &= \sqrt{3}|\gamma|^{1/3}\left(1 + \frac{\alpha}{3\sqrt{3}|\gamma|}\right). \end{aligned}$$

  Therefore, $(T_+ + T_-)^2 \approx 3|\gamma|^{2/3}\big(1 + \frac{2\alpha}{3\sqrt{3}|\gamma|}\big) = 3|\gamma|^{2/3} + \frac{2}{\sqrt{3}}\frac{\alpha}{|\gamma|^{1/3}}$. Lastly, noticing that $|\gamma|^{2/3} \approx |\beta| \approx \frac{2}{3}k\log 3$, we can conclude that $(T_+ + T_-)^2 - 2k\log 3 \approx C \cdot \frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}$ for some constant C.

Proof of Theorem 7.6.2. We follow a similar strategy to that of Theorem 7.6.1 with some modifications. Here, we assume S is ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$ only in the average sense, and thus, S can be arbitrarily shaped and $S \subseteq (1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)$ is not necessarily true. Instead, we define a set Q – to be precise, we let $Q = 10(1+\epsilon)\sqrt{n}\mathbb{G}_S^\circ$ for $\mathbb{G}_S$ to be defined in (7.23) – that contains S in an adaptive manner. Then we consider the generalized slack matrix of the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ Q\right)$. We express the slack matrix in two equivalent ways: one is obtained from the knowledge about the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$, and the other is obtained by assuming the existence of a $\boldsymbol{S}_+^k$-factorization having N factors. Interpreting the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$ and Q° as formal variables, x and G, we may view the two expressions of the slack matrix as bivariate polynomials. As already done in the proof of Theorem 7.6.1, we 'smooth out' the two expressions with respect to one variable, x; and then take expectation with respect to the other variable, G. Comparing the two resulting expressions, we derive a lower bound on the number of factors N, which implies a lower bound on the $\boldsymbol{S}_+^k$-extension complexity of S.

Step 1. Gaussian Surrogate for S° and the Associated Slack Matrix Let $\boldsymbol{S}_0^n$ denote the set of $n \times n$ symmetric matrices with trace zero, endowed with the trace inner product. Let $\mathcal{N}_0$ denote the standard Gaussian distribution associated to $\boldsymbol{S}_0^n$, i.e., $G_0 \sim \mathcal{N}_0$ if $G_0 = G - \frac{\mathrm{Tr}G}{n}I_n$ where G has the standard Gaussian distribution in S^n. Then we define a set

$$\mathbb{G}_S = \left\{G \in \boldsymbol{S}_0^n : \big|\,\langle G, X\rangle\,\big| \le 5\sqrt{2}w_G(S),\ \forall X \in S\right\}. \tag{7.23}$$

The number $5\sqrt{2}$ is chosen for the convenience of our analysis, and has no special meaning. Observe that $w_G(S) \le (1+\epsilon)\cdot w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big) \le (1+\epsilon)\sqrt{2n}$, cf. Remark 7.2.10.

Then, we can see that

$$-10(1+\epsilon)\sqrt{n} \le \langle G, X\rangle \le 10(1+\epsilon)\sqrt{n}, \qquad \forall (X,G) \in S \times \mathbb{G}_S.$$

This implies that $\frac{1}{10(1+\epsilon)\sqrt{n}}\mathbb{G}_S \subseteq S^\circ$, or equivalently, $S \subseteq 10(1+\epsilon)\sqrt{n}\mathbb{G}_S^\circ$.

Now we consider the slack operator associated to the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ 10(1+\epsilon)\sqrt{n}\mathbb{G}_S^\circ\right)$, treating $\frac{1}{10(1+\epsilon)\sqrt{n}}\mathbb{G}_S$ as a surrogate for S°. Specifically, we are led to study the following infinite matrix:

$$(\tilde{x}, G) \in \mathbb{S}^{n-1} \times \mathbb{G}_S \mapsto 1 - \left\langle \tilde{x}\tilde{x}^T - \frac{1}{n}I_n, \frac{1}{10(1+\epsilon)\sqrt{n}}G \right\rangle = 1 - \frac{1}{10(1+\epsilon)\sqrt{n}}\tilde{x}^T G \tilde{x}.$$

We consider the PSD rank ($\boldsymbol{S}_+^k$-rank) of the submatrix restricted to $\tilde{x} \in \left\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right\}^n \subset \mathbb{S}^{n-1}$, with a proper reparametrization ($x = \sqrt{n}\tilde{x}$), namely,

$$s : (x, G) \in H_n \times \mathbb{G}_S \mapsto 1 - \frac{1}{10(1+\epsilon)n\sqrt{n}}x^T G x. \tag{7.24}$$

Assuming that we can write the matrix (7.24) as a sum of N trace inner products of $\boldsymbol{S}_+^k$ factors, we have

$$1 - \frac{1}{10(1+\epsilon)n\sqrt{n}}x^T G x = s(x,G) = \sum_{i=1}^N \langle f_i(x), g_i(G)\rangle, \qquad \forall (x,G) \in H_n \times \mathbb{G}_S \tag{7.25}$$

where $f_i : H_n \to \boldsymbol{S}_+^k$ and $g_i : \mathbb{G}_S \to \boldsymbol{S}_+^k$ are some matrix-valued functions.

Again, we 'smooth out' the two expressions of $s(x,G)$ in (7.25) and compare them to derive a lower bound for N. To be precise, for each fixed $G \in \mathbb{G}_S$, we let $q_G(x) = -x^T G x$. Recall that we let μ denote the uniform probability measure on H_n, and observe that for any function $f : H_n \to \mathbb{R}$, the inner product, $\langle f, q_G(x)\rangle_\mu = \mathbb{E}_{x\sim\mu}[f(x)q_G(x)]$ is a centered Gaussian random variable.

Taking the inner product of both sides of (7.25) with $q_G(x)$, we get

$$\mathbb{E}_{x\sim\mu}\left[q_G(x) + \frac{1}{10(1+\epsilon)n\sqrt{n}}q_G(x)^2\right] = \sum_{i=1}^N \langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i(x)], g_i(G)\rangle.$$

Letting $\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}[\ \cdot\ ]$ denote the conditional expectation with respect to $G \sim \mathcal{N}_0$ given

$G \in \mathbb{G}_S$, we can see that

$$\underbrace{\frac{1}{10(1+\epsilon)n\sqrt{n}} \cdot \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right]}_{=:LHS} = \underbrace{\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^{N}\left\langle\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i(x)], g_i(G)\right\rangle}_{=:RHS}. \tag{7.26}$$

The rest of the proof is organized as follows. In Step 2, we prove a lower bound for the expectation on the left-hand side. In Step 3, we derive an upper bound on the expectation on the right-hand side as a function of N. In the end, we obtain the desired lower bound on N in Step 4 by comparing these bounds.

Step 2. A Lower Bound for the Left-hand side of (7.26). We additionally define a set

$$\mathbb{G}_{1/2} = \left\{G \in \boldsymbol{S}_0^n : \mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \frac{1}{5}n(n-1)\right\}.$$

The constant $1/5$ is chosen for the convenience of analysis, and has no special meaning. By the law of total probability, we can see that

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S\cap\mathbb{G}_{1/2}}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right]\cdot\Pr[G\in\mathbb{G}_S\cap\mathbb{G}_{1/2}\mid G\in\mathbb{G}_S].$$

Note that $\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S\cap\mathbb{G}_{1/2}}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \frac{1}{5}n(n-1)$ by definition of $\mathbb{G}_{1/2}$. Thus, it suffices to find a lower bound for the conditional probability, $\Pr[G\in\mathbb{G}_S\cap\mathbb{G}_{1/2}\mid G\in\mathbb{G}_S]$.

It is easy to see that

$$\Pr[G\in\mathbb{G}_S\cap\mathbb{G}_{1/2}\mid G\in\mathbb{G}_S] = \frac{\Pr[G\in\mathbb{G}_S\cap\mathbb{G}_{1/2}]}{\Pr[G\in\mathbb{G}_S]} \geq \frac{\Pr[G\in\mathbb{G}_S]-\Pr[G\notin\mathbb{G}_{1/2}]}{\Pr[G\in\mathbb{G}_S]}.$$

Observe that $\Pr[G\in\mathbb{G}_S] \geq 1-\exp\left(-\frac{(5\sqrt{2}-1)^2}{4\pi}\right) > 0.893$ by Lemma 2.4.13. Now it remains to show an upper bound for $\Pr[G\notin\mathbb{G}_{1/2}]$.

We use standard concentration results for the chi-square distribution. Note that if $G\sim\mathcal{N}_0$, then $q_G(x) = -\mathrm{Tr}\,G - 2\sum_{i<j}G_{ij}x_ix_j = -2\sum_{i<j}G_{ij}x_ix_j$, and therefore, $\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] = 4\sum_{i<j}G_{ij}^2$. Thus, we have $\mathbb{E}_{G\sim\mathcal{N}_0}\mathbb{E}_{x\sim\mu}[q_G(x)^2] = 4\binom{n}{2}\frac{1}{2} = n(n-1)$. Using an exponential inequality for chi-square distribution (e.g., [87, Lemma 1]), we obtain $\Pr[G\notin\mathbb{G}_{1/2}] \leq \exp\left(-\frac{2}{25}n(n-1)\right) \leq 0.8522$ for all $n\geq 1$.

All in all, we obtain

$$LHS \text{ in } (7.26) \geq \frac{1}{10(1+\epsilon)n\sqrt{n}}\cdot\frac{1}{5}n(n-1)\cdot\frac{\Pr[G\in\mathbb{G}_S]-\Pr[G\notin\mathbb{G}_{1/2}]}{\Pr[G\in\mathbb{G}_S]} \geq \frac{\sqrt{n}}{2200(1+\epsilon)} \tag{7.27}$$

because $\frac{0.893-0.8522}{0.893} \geq 1/22$ and $n-1\geq n/2$ for all $n\geq 1$.

Step 3. An Upper Bound for the Right-hand side of (7.26). Next, we prove an upper bound on the right-hand side of (7.26), which is a function of N. Note that for the same reason as discussed in Step 3-A of the proof of Theorem 7.6.1, we may assume without loss of generality that the factor functions f_i, g_i satisfy

$$\|\mathbb{E}_{x\sim\mu}[f_i(x)]\|_{op} = 1, \quad \forall i \in [N] \qquad \text{and} \qquad \sum_{i=1}^{n} \mathrm{Tr}\big(g_i(G)\big) = 1, \quad \forall G \in \mathbb{G}_S. \tag{7.28}$$

For each $i \in [N]$, we define the component functions $f_i^\sharp, f_i^\flat : H_n \to \boldsymbol{S}_+^k$ in the same way as in (7.17), using a fixed threshold Λ whose value will be determined later in this proof, cf. Step 3-B of the proof of Theorem 7.6.1.

By linearity of expectation, we can decompose the expression on the right-hand side of (7.26) as

$$\begin{aligned} RHS \text{ in } (7.26) = \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S} \sum_{i=1}^{N} \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\sharp(x)], g_i(G)\Big\rangle \\ + \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S} \sum_{i=1}^{N} \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)], g_i(G)\Big\rangle . \end{aligned} \tag{7.29}$$

In the two sub-steps below, we prove upper bounds for the two terms on the right hand side separately.

Step 3-A. Upper Bound on the Contribution of Sharp Components in (7.29) Here we argue that the first term on the right hand side of (7.29) is bounded from above by $\frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N$. Our argument is based on the following three observations.

- Let $\mathrm{supp}\ f_i^\sharp = \{x \in H_n : f_i^\sharp(x) \neq 0\}$. Then $|\mathrm{supp}\ f_i^\sharp| < \frac{k}{\Lambda}\, 2^n$ for all $i \in [N]$, cf. Step 3-C of the proof of Theorem 7.6.1.
- Observe that $\langle f_i^\sharp(x),\ g_i(G)\rangle \le \langle f_i(x), g_i(G)\rangle \le s(x,G) \le 2$ for all $i \in [N]$ and for all $(x,G) \in H_n \times \mathbb{G}_S$.
- $q_G(x) = -x^T G x = 8(1+\epsilon)n\sqrt{n}\big(s(x,G)-1\big) \le 8(1+\epsilon)n\sqrt{n}$ for all $(x,G) \in H_n \times \mathbb{G}_S$.

Combining these observations, we can see that for every $G \in \mathbb{G}_S$,

$$\begin{aligned} \sum_{i=1}^{N} \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x) f_i^\sharp(x)], g_i(G)\Big\rangle &= \sum_{i=1}^{N} \mathbb{E}_{x\sim\mu}\Big[q_G(x)\Big\langle f_i^\sharp(x), g_i(G)\Big\rangle\Big] \\ &\le \sum_{i=1}^{N} \frac{\big|\mathrm{supp}\ f_i^\sharp\big|}{2^n} \cdot 16(1+\epsilon)n\sqrt{n} \\ &\le \frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N. \end{aligned}$$

This upper bound is independent of G, and thus, we get

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^{N}\Big\langle\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\sharp(x)],g_i(G)\Big\rangle\leq\frac{16(1+\epsilon)}{\Lambda}kn\sqrt{n}N. \tag{7.30}$$

Step 3-B. Upper Bound on the Contribution of Flat Components in (7.29) Here we prove an upper bound for the second term in (7.29). We observe that for every $G\in\mathbb{G}_S$,

$$\begin{aligned}\sum_{i=1}^{N}\Big\langle\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)],g_i(G)\Big\rangle&\leq\sum_{i=1}^{N}\Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)]\Big\|_{op}\ \mathrm{Tr}\ g_i(G)\\&\leq\max_{i\in[N]}\Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)]\Big\|_{op}\end{aligned}$$

due to Cauchy-Schwarz inequality and the normalization assumption (7.28) that $\sum_{i=1}^{N}\mathrm{Tr}\ g_i(G)=1,\ \forall G\in\mathbb{G}_S$.

Given $\epsilon_{\text{net}}>0$, let $\mathcal{N}$ be an ϵ_{net}-net of $\mathbb{S}^{k-1}$ with the smallest possible cardinality. It follows from the standard ϵ-net argument that for each $i\in[N]$,

$$\Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)]\Big\|_{op}=\sup_{v\in\mathbb{S}^{k-1}}v^T\mathbb{E}_{x\sim\mu}\big[q_G(x)\cdot f_i^\flat(x)\big]v\leq\frac{1}{1-2\epsilon_{\text{net}}}\max_{v\in\mathcal{N}}\mathbb{E}_{x\sim\mu}\big[q_G(x)\cdot v^T f_i^\flat(x)v\big].$$

Next, we observe that if $G\sim\mathcal{N}_0$, then for every function $f:H_n\to\mathbb{R}$, the derived random variable $\langle f,\ q_G(x)\rangle_\mu$ is a centered Gaussian random variable with variance

$$\begin{aligned}\mathbb{E}_{G\sim\mathcal{N}_0}\Big[\langle f,\ q_G(x)\rangle_\mu^2\Big]&=\mathbb{E}_{G\sim\mathcal{N}_0}\Big[\mathbb{E}_{x\sim\mu}\big[f(x)\cdot x^TGx\big]^2\Big]\\&=\mathbb{E}_{G\sim\mathcal{N}_0}\left[\mathbb{E}_{x\sim\mu}\left[f(x)\cdot\Big(\mathrm{Tr}\ G+2\sum_{i<j}G_{ij}x_ix_j\Big)\right]^2\right]\\&=4\sum_{i<j}\mathbb{E}_{G\sim\mathcal{N}_0}[G_{ij}^2]\cdot\mathbb{E}_{x\sim\mu}[f(x)x_ix_j]^2=2\sum_{i<j}\langle f(x),\ x_ix_j\rangle_\mu^2\\&=2\|\mathrm{proj}_2 f\|_2^2.\end{aligned}$$

Then we use Lemma 2.4.21 to obtain the following inequalities:

$$\begin{aligned}&\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^{N}\Big\langle\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)],g_i(G)\Big\rangle\\&\quad\leq\frac{1}{\Pr[G\in\mathbb{G}_S]}\mathbb{E}_{G\sim\mathcal{N}_0}\sum_{i=1}^{N}\Big\langle\mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)],g_i(G)\Big\rangle\end{aligned}$$

$$\leq \frac{1}{\Pr[G \in \mathbb{G}_S]} \frac{1}{1-2\epsilon_{\text{net}}} \mathbb{E}_{G\sim\mathcal{N}_0}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{X}}} \mathbb{E}_{x\sim\mu}[q_G(x)\ v^T f_i^\flat(x) v]\right]$$
$$\leq \frac{1}{\Pr[G \in \mathbb{G}_S]} \frac{2}{1-2\epsilon_{\text{net}}} \left(\max_{\substack{i\in[N]\\ v\in\mathcal{X}}} \|\text{proj}_2(v^T f_i^\flat v)\|_2\right) \sqrt{\log(N|\mathcal{X}|)}.$$

We have seen in Step 2 that $\Pr[G \in \mathbb{G}_S] \geq 1 - \exp\left(-\frac{(5\sqrt{2}-1)^2}{4\pi}\right) \geq 4/5$. Also, Lemma 6.2.4 ensures that $\|\text{proj}_2(v^T f_i^\flat v)\|_2 \leq e \log \Lambda$ for all (i, v), provided that we will choose the threshold $\Lambda \geq e$. Lastly, it is well known that $|\mathcal{X}| \leq \left(1 + \frac{2}{\epsilon_{\text{net}}}\right)^k$; e.g. see Lemma 2.2.6. In conclusion, we obtain

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S} \sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot f_i^\flat(x)], g_i(G)\right\rangle \leq \frac{5e\log\Lambda}{2(1-2\epsilon_{\text{net}})} \sqrt{\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]}. \tag{7.31}$$

Step 4. Concluding the Proof Lastly, we revisit Eq. (7.26) to conclude the proof. Recall that we obtained a lower bound for the left-hand side in Step 2, cf. (7.27), and derived an upper bound for the right-hand side in Step 3, cf. (7.29), (7.30), and (7.31). Putting these together, we obtain the following inequality that holds for any choice of parameters ϵ_{net}, Λ such that $0 < \epsilon_{\text{net}} < \frac{1}{2}$ and $\Lambda \geq e$:

$$\frac{1}{2200(1+\epsilon)}\sqrt{n} \leq \frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N + \frac{5e\log\Lambda}{2(1-2\epsilon_{\text{net}})} \sqrt{\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]}. \tag{7.32}$$

We choose $\epsilon_{\text{net}} = 1/4$ for simplicity because optimizing ϵ_{net} does not make much difference. Next, we find Λ that minimizes the right-hand side of (7.32). It is easy to see that the upper bound is minimized (w.r.t. Λ) at $\Lambda^* = \frac{16(1+\epsilon)kn\sqrt{n}N}{5e\sqrt{\log(9^k N)}}$. As a result, we get the following inequality from (7.32) by choosing $\Lambda = \Lambda^*$ and noticing $N \geq 1$:

$$\frac{1}{11000e(1+\epsilon)}\sqrt{n} \leq \sqrt{\log(9^k N)} \cdot \log\left(\frac{16(1+\epsilon)kn\sqrt{n}N}{5\sqrt{\log(9^k N)}}\right)$$
$$\leq \sqrt{\log(9^k N)} \cdot \left[\log N + \log\left(\frac{16(1+\epsilon)kn\sqrt{n}}{5\sqrt{\log(9^k)}}\right)\right]. \tag{7.33}$$

Letting $z = \sqrt{\log(9^k N)}$, we can see that (7.33) is a cubic inequality of the form $z^3 + 3\beta z \geq 2\alpha$ where

$$\alpha = \frac{\sqrt{n}}{22000e(1+\epsilon)} \quad \text{and} \quad \beta = \frac{1}{3}\log\left(\frac{16(1+\epsilon)kn\sqrt{n}}{5\cdot 9^k\sqrt{\log(9^k)}}\right).$$

We want to solve the cubic inequality with an implicit constraint $z > 0$ because $\log(9^k N) > 0$ for all $k, N \geq 1$.

Note that $\alpha > 0$ for all $\epsilon \geq 0$, $n \geq 1$. Observe that the cubic equation $z^3 + 3\beta z - 2\alpha = 0$ always has a unique positive real root when $\alpha > 0$, regardless of the value of β. Letting z_* denote the positive real root, we can see that $\{z \in \mathbb{R} : \ z^3 + 3\beta z \geq 2\alpha, \ z > 0\} = \{z \in \mathbb{R} : \ z \geq z_*\}$. Indeed, we can explicitly write z_* as $z_* = \big(\alpha + \sqrt{\alpha^2 + \beta^3}\big)^{1/3} + \big(\alpha - \sqrt{\alpha^2 + \beta^3}\big)^{1/3}$, due to the general cubic formula, commonly referred to as Cardano's formula. See Appendix 7.7.2 for more details.

Consequently, we obtain the following lower bound for N by solving (7.22):

$$\log N \geq \left\{\big(\alpha + \sqrt{\alpha^2 + \beta^3}\big)^{1/3} + \big(\alpha - \sqrt{\alpha^2 + \beta^3}\big)^{1/3}\right\}^2 - 2k \log 3$$

because $\sqrt{\log(9^k N)} \geq z_*$ if and only if $\log N \geq z_*^2 - 2k \log 3$.

■

■ 7.7 Appendix to the Chapter

■ 7.7.1 More on Example 7.2.8 (Ball, Needle, and Pancake)

Let $B_2^d := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ denote the d-dimensional unit ℓ_2-ball, and let $B = B_2^d$. Fix $0 < \delta < 1$, and let $N = \mathrm{conv}\big\{B_2^d(0,1) \cup \{\pm \frac{1}{\delta} e_1\}\big\}$ be the 'needle' where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$. Lastly, we define the 'pancake' $P = \{x \in B : -\delta \leq x_1 \leq \delta\}$ where x_1 is the first coordinate of $x \in \mathbb{R}^d$. Observe that N and P are the polars of each other, and B is the polar of itself.

First of all, $w_G(B) = \mathbb{E}_g \|g\|_2 = \kappa_d$ and it is known that $\sqrt{d - 1/2} \leq \kappa_d \leq \sqrt{d - d/(2d+1)}$, cf. the paragraph below Definition 2.3.4. Next, we can see that $w_G(N) \geq \frac{1}{\delta}\sqrt{2/\pi}$ because $\{\pm \frac{1}{\delta} e_1\} \subseteq N$ and thus, $w_G(N) \geq w_G\big(\{\pm \frac{1}{\delta} e_1\}\big) = \frac{1}{\delta} \mathbb{E}_{g \sim \mathcal{N}(0,1)} |g| = \frac{1}{\delta}\sqrt{2/\pi}$. Lastly, observe that $w_G(P) \geq \kappa_{d-1} \geq \sqrt{d - 3/2}$ because $\{0\} \times B_2^{d-1}(0,1) \subseteq P$ and $w_G(P) \geq w_G\big(\{0\} \times B_2^{d-1}(0,1)\big) = w_G\big(B_2^{d-1}(0,1)\big) = \kappa_{d-1}$.

It follows that B is an ϵ-approximation of P in the average sense for $\epsilon = \kappa_d/\kappa_{d-1} - 1 \leq 3/(2d-3)$. Nevertheless, B is not an ϵ'-approximation of P in the dual-average sense unless $\epsilon' \geq \frac{1}{\delta}\sqrt{2/\pi}/\kappa_d - 1 \geq \frac{2}{\delta\sqrt{\pi(2d-1)}} - 1$, which can be made arbitrarily large by choosing small δ. For example, if we choose $\delta \leq 1/\sqrt{\pi(2d-1)}$, then $\epsilon^*_{\text{dual-avg}}(P, S) \geq 1$ whereas $\epsilon^*_{\text{avg}}(P, S) \leq 3/(2d-3)$ regardless of δ.

■ 7.7.2 Solving the Cubic Inequality $z^3 + \alpha z \geq \beta$ with $\beta > 0$

Consider a cubic equation of the form $z^3 + \alpha z - \beta = 0$, which is commonly referred to as a depressed cubic. Note that when $\beta > 0$, this cubic equation always has a positive real root. The other two roots can be either negative real roots (when $D \leq 0$), or a pair of complex conjugate roots (when $D > 0$), depending on the sign of its discriminant, $D = (\alpha/3)^3 + (\beta/2)^2$.

Indeed, we can find the roots with a generic cubic formula, known as Cardano's formula. Let $i = \sqrt{-1}$ denote the imaginary unit, $\omega = \frac{-1+\sqrt{3}i}{2}$ be a primitive 3rd of unity, and

$$T_+ = \sqrt[3]{\frac{\beta}{2} + \sqrt{\left(\frac{\beta}{2}\right)^2 + \left(\frac{\alpha}{3}\right)^3}} \quad \text{and} \quad T_- = \sqrt[3]{\frac{\beta}{2} - \sqrt{\left(\frac{\beta}{2}\right)^2 + \left(\frac{\alpha}{3}\right)^3}}. \tag{7.34}$$

Case 1: $D > 0$. When $D > 0$, the cubic equation $z^3 + \alpha z - \beta = 0$ with $\beta > 0$ has only one real root, $z^* = T_+ + T_-$, which turns out to be positive. Thus, the set of real solutions for the cubic inequality $z^3 + \alpha z \geq \beta$ is $\{z \in \mathbb{R} : z \geq T_+ + T_-\}$.

Case 2: $D \leq 0$. There are three real roots for the cubic equation $z^3 + \alpha z - \beta = 0$, which can be written as

$$z_1 = T_+ + T_-, \qquad z_2 = \omega T_+ + \omega^2 T_-, \qquad z_3 = \omega^2 T_+ + \omega T_-.$$

One of these three real roots is positive, and the other two are negative.

Note that (7.34) now involves complex roots, and the choice of branches might affect the order of the roots, z_1, z_2, z_3, however, the choice will not change the values of the roots. To avoid any ambiguity in our description, we choose the principal branch so that $\operatorname{Arg}(\sqrt[m]{z}) \in (-\frac{\pi}{m}, \frac{\pi}{m}]$ for any complex number z and any positive integer m.

Observe that $T_+ = \sqrt[3]{\beta/2 + \sqrt{|D|}i}$ and $\operatorname{Arg}(T_+) \in [0, \pi/3)$. Similarly, we can see that $\operatorname{Arg}(T_-) \in (-\pi/3, 0]$. It follows that $T_+ + T_-$ is a positive real number, and thus, the largest real root. Thus, the set of real solutions for the cubic inequality $z^3 + \alpha z \geq \beta$ is $\{z \in \mathbb{R} : z \geq T_+ + T_-\}$.

Part IV

Conclusions

Chapter 8

Concluding Remarks

To conclude the thesis, we briefly summarize our main contributions, and outline some directions for future research.

8.1 Summary of the Thesis

Part II of the Thesis In Part II of the thesis, we explore the idea of imputing missing values in data using low-rank matrix completion techniques for the purpose of predictive modeling. In Chapter 3, after briefly reviewing existing matrix completion techniques and their error analysis in the literature, we establish novel error guarantees for the singular value thresholding algorithm. In Chapters 4 and 5, we consider two concrete problem setups – supervised learning and reinforcement learning – and argue the utility of data imputation for predictive modeling, based on the error guarantees for matrix completion. To be specific, in Chapter 4, we study the errors-in-variables regression problem and prove an upper bound for the bias in prediction that arises from having the missing data in covariates imputed by matrix completion. In Chapter 5, we show that one can provably reduce the sample complexity in Q-learning algorithms by exploring only a small subset of state-action pairs and extrapolating the gathered information, e.g. by utilizing the low-rank structure in the Q^*-function.

Part III of the Thesis In Part III of the thesis, we investigate the approximability of the cone of positive semidefinite (PSD) matrices by a few, smaller-sized PSD constraints. As discussed in Chapter 6, we are motivated to study this question by the scalability issue of semidefinite programming. In Chapter 7, we prove lower bounds on the number of $k \times k$ PSD constraints required to approximate the $n \times n$ PSD cone. In particular, our extension complexity lower bounds do not rely on any specific construction; they apply to *all* sets that approximate $\boldsymbol{S}^n_+$. As a result, these lower bounds refute the possibility of *globally* approximating the large PSD cone $\boldsymbol{S}^n_+$ with a few, small-sized $k \times k$ PSD constraints. Nevertheless, it may still be possible to *locally* approximate the feasible set of the original SDP, e.g. by exploiting the problem data such as the objective and the constraints. Indeed, our results in this thesis suggest that local and/or problem-adaptive techniques are essential to approximate SDPs.

8.2 Future Directions

8.2.1 Handling Missing Data

Beyond Low-rank Matrix Completion In this thesis, we attempted to address the challenge of missing data by 'guessing' what the missing values would be if they were available. For that purpose, we utilize matrix completion techniques to impute the missing values in the data by exploiting the low-rank structure inherent in the data matrix. While the rank is a natural notion of 'simplicity' for matrix-formatted data, this notion crucially relies on certain linearity assumptions; thus, there may be more appropriate notions of the simplicity, depending on the specific problem settings and the types of data. For example, if the data vectors are contained in a low-dimensional manifold rather than in a low-dimensional subspace, the rank of the data matrix can be much larger than the intrinsic dimensionality of the data. Moreover, matrices may not be the best format to represent some types of data; a certain type of data may be better represented by graphs, or higher-order tensors, etc. Therefore, it would be an interesting question to identify the counterpart notions of the rank for data beyond matrices, and to develop algorithmic tools for imputing the missing values in data of such types.

Parametric vs Non-parametric Modeling Assumptions While we impose a modeling assumption that the data matrix has low rank, our approaches to handle the missing data are basically non-parametric as we do not assume any parametric models. In contrast, the likelihood-based approaches to errors-in-variables regression (reviewed in Section 4.2.1) make parametric assumptions to handle the missing/noisy data. When correct assumptions are made, parametric approaches lead to more accurate estimates, requiring a less amount of data. However, when the modeling assumptions are incorrect, non-parametric approaches are usually less vulnerable to the mismatch between the model and the reality. Thus, it would be an interesting research direction to investigate if we can achieve the best of both worlds, i.e., the efficiency of parametric modeling and the flexibility of the non-parametric modeling.

Overcoming Missing Data without Imputation A more fundamental question of interest would be to ask if we can overcome the missing data problem without imputing the missing values. In this thesis, we focus on data imputation to see if we can continue to use standard statistical techniques. However, at the core of our analysis is the robust recovery of certain statistical quantities[1] that contain sufficient information for the task of interest – e.g. the column space of X in the errors-in-variables regression (Chapter 4), and $\max_{a \in \mathcal{A}} Q^*(s, a)$ in the Q-learning (Chapter 5) – and data imputation is just one option to estimate the quantities. Thus, it is of interest to ask if we can directly compute the "sufficient statistics," avoiding the computation and storage overheads incurred by imputing the missing values in data.

[1] They play a role similar to that of sufficient statistics in parameter estimation.

■ 8.2.2 Approximating SDP for Enhanced Scalability

Sharper Characterization of the $\boldsymbol{S}_+^k$-Extension Complexity to Approximate $B_H\left(\boldsymbol{S}_+^n\right)$ In Chapter 7, we presented an $\boldsymbol{S}_+^k$-extension complexity lower bound (Theorem 7.4.1) that holds for any set that approximates $B_H\left(\boldsymbol{S}_+^n\right)$. Nevertheless, we do not know whether this lower bound is tight, or can be further improved. For example, for $k = 1$, our lower bound implies that $\Omega_n\big(\exp(\sqrt{n})\big)$ number of linear constraints (i.e., $k \times k$ PSD constraints with $k = 1$) are required to approximate $B_H\left(\boldsymbol{S}_+^n\right)$, whereas the best construction known in the literature so far has size of $O_n\big(\exp(n)\big)$. We leave it as an open question to resolve this discrepancy, either by proving a stronger exponential complexity lower bound or by inventing a construction scheme with smaller size.

We conjecture that the 'plateau' observed in our lower bound (Theorem 7.4.1) at the level of $\exp(\sqrt{n})$ for all $k = O_n(\sqrt{n})$ (see Figure 7.1b) is an artifact of our analysis. We are presently working to refine our proof techniques in hopes of achieving an improved lower bound. Specifically, we conjecture that if S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then S must have extension complexity bounded from below as

$$\mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \frac{1}{1+\epsilon}\,\frac{n}{k}\right).$$

Local, Adaptive Approximation Methods for Scalable SDP In this thesis, we argue that it is hard to *globally* approximate the large PSD cone $\boldsymbol{S}_+^n$ with a few, small-sized $k \times k$ PSD constraints. However, if one is allowed to exploit the problem data (namely, $C, (A_i, b_i)_{i=1}^m$), it could be still possible to construct a good approximation F' of the original feasible set $F = \{X \in \boldsymbol{S}_+^n : \langle A_i, X\rangle = b_i\}$ by using a much smaller number of $k \times k$ PSD constraints so that the optimality gap $\sup_{X\in F'}\langle C, X\rangle - \sup_{X\in F}\langle C, X\rangle$ is small. For instance, Ahmadi et al. [7, 5] proposed iterative procedures to improve DSOS/SDSOS approximations to an SDP, based on the ideas of re-centering and column generation. Despite their promising empirical performance, there are no theoretical guarantees available for such procedures yet, to the best of our knowledge. Thus, it is of interest to develop a practical framework to approximate a given large-scale SDP with provable guarantees for the computational gain as well as the approximation accuracy.

Bibliography

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3): 1452–1474, 2020.

[2] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(3), 2009.

[3] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[4] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Journal of the American Statistical Association*, To appear, 2021.

[5] Amir Ali Ahmadi and Georgina Hall. Sum of squares basis pursuit with linear and second order cone programming. *Algebraic and Geometric Methods in Discrete Mathematics*, 685:27–53, 2017.

[6] Amir Ali Ahmadi and Anirudha Majumdar. DSOS and SDSOS optimization: More tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2):193–230, 2019.

[7] Amir Ali Ahmadi, Sanjeeb Dash, and Georgina Hall. Optimization over structured subsets of positive semidefinite matrices via column generation. *Discrete Optimization*, 24:129–151, 2017.

[8] Paul D Allison. *Missing data*. Sage Publications, 2001.

[9] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous

action-space MDPs. In *Advances in Neural Information Processing Systems*, volume 20, pages 9–16, 2007.

[10] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.

[11] Guillaume Aubrun and Stanislaw Szarek. Dvoretzky's theorem and the complexity of entanglement detection. *Discrete Analysis*, 2017.

[12] Guillaume Aubrun and Stanisław J Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Society, 2017.

[13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

[14] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473): 119–137, 2006.

[15] Evelyn ML Beale and Roderick JA Little. Missing values in multivariate analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):129–145, 1975.

[16] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, pages 159–182, 1975.

[17] Robert M Bell and Yehuda Koren. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[18] Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956, 2017.

[19] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific, 4th edition, 2017.

[20] Åke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[21] Grigoriy Blekherman, Santanu S Dey, Marco Molinaro, and Shengding Sun. Sparse PSD approximation of the PSD cone. *Mathematical Programming*, pages 1–24, 2020.

[22] Erik G Boman, Doron Chen, Ojas Parekh, and Sivan Toledo. On factor width and symmetric H-matrices. *Linear Algebra and Its Applications*, 405:239–248, 2005.

[23] Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. In *Annales de l'Institut Fourier*, volume 20, pages 335–402, 1970.

[24] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[25] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[26] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[27] V Buldygin and K Moskvichova. The sub-gaussian norm of a binary random variable. *Theory of Probability and Mathematical Statistics*, 86:33–49, 2013.

[28] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2): 329–357, 2003.

[29] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[30] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

[31] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[32] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[33] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[34] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[35] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: A modern perspective*. CRC Press, 2nd edition, 2006.

[36] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[37] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214, 2015.

[38] 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. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

[39] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

[40] Nir Cohen, Charles R Johnson, Leiba Rodman, and Hugo J Woerdeman. Ranks of completions of partial matrices. In *The Gohberg Anniversary Collection*, pages 165–185. Springer, 1989.

[41] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[42] Abhirup Datta and Hui Zou. CoCoLasso for high-dimensional error-in-variables regression. *Annals of Statistics*, 45(6):2400–2426, 2017.

[43] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[44] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[45] Arthur P. Dempster and Donald B Rubin. Introduction pp 3-10. In *Incomplete Data in Sample Surveys: Theory and Bibliographies*, volume 2. Academic Press, 1983.

[46] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[47] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[48] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52 (4):1289–1306, 2006.

[49] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of International Conference on Machine Learning 2019*, pages 1665–1674. PMLR, 2019.

[50] Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. *arXiv preprint arXiv:2103.10897*, 2021.

[51] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[52] Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: Spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory*, pages 321–358. PMLR, 2018.

[53] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010.

[54] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(1), 2003.

[55] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[56] Hamza Fawzi. On polyhedral approximations of the positive semidefinite cone. *Mathematics of Operations Research*, 2021.

[57] Hamza Fawzi and Pablo A Parrilo. Exponential lower bounds on fixed-size PSD rank and semidefinite extension complexity. *arXiv preprint arXiv:1311.2571*, 2013.

[58] Hamza Fawzi, João Gouveia, Pablo A Parrilo, James Saunderson, and Rekha R Thomas. Lifting for simplicity: Concise descriptions of convex sets. *arXiv preprint arXiv:2002.09788*, 2020.

[59] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734–4739. IEEE, 2001.

[60] Abel Folch-Fortuny, Francisco Arteaga, and Alberto Ferrer. PCA model building with missing data: New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 146:77–88, 2015.

[61] Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[62] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[63] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[64] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[65] Gene H Golub and Hongyuan Zha. The canonical correlations of matrix pairs and their numerical computation. In *Linear Algebra for Signal Processing*, pages 27–49. Springer, 1995.

[66] Joao Gouveia, Pablo A Parrilo, and Rekha R Thomas. Lifts of convex sets and cone factorizations. *Mathematics of Operations Research*, 38(2):248–264, 2013.

[67] Bjørn Grung and Rolf Manne. Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2):125–139, 1998.

[68] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of International Conference on Machine Learning 2018*, pages 1861–1870. PMLR, 2018.

[69] Joe Harris. *Algebraic geometry: A first course*, volume 133. Springer Science & Business Media, 2013.

[70] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[71] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[72] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

[73] 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*, pages 665–674, 2013.

[74] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of International Conference on Machine Learning 2017*, pages 1704–1713. PMLR, 2017.

[75] Ian T Jolliffe. *Principal component analysis*. Springer, 1986.

[76] Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique de France*, 3:103–174, 1875.

[77] Sham Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

[78] Michael J Kearns, Yishay Mansour, and Andrew Y Ng. Approximate planning in large pomdps via reusable trajectories. In *Advances in Neural Information Processing Systems*, volume 12, pages 1001–1007, 1999.

[79] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[80] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078, 2010.

[81] Olga Klopp. Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5:1161–1183, 2011.

[82] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[83] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[84] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, volume 29, pages 1848–1856, 2016.

[85] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[86] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[87] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[88] Christina E Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*, volume 29, pages 2155–2163, 2016.

[89] James R Lee, Prasad Raghavendra, and David Steurer. Lower bounds on the size of semidefinite programming relaxations. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, pages 567–576, 2015.

[90] Jason D Lee, Ben Recht, Ruslan Salakhutdinov, Nathan Srebro, and Joel A Tropp. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[91] Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, 2019.

[92] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[93] Alan Wee-Chung Liew, Ngai-Fong Law, and Hong Yan. Missing value imputation for gene expression data: Computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12(5):498–513, 2011.

[94] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of International Conference on Learning Representations*, 2016.

[95] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 3rd edition, 2019.

[96] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3): 1637–1664, 2012.

[97] László Lovász. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25(1):1–7, 1979.

[98] Hamid R Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 719–726, 2010.

[99] Anirudha Majumdar, Georgina Hall, and Amir Ali Ahmadi. Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:331–360, 2020.

[100] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[101] Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, pages 664–671, 2008.

[102] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[103] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

[104] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *Proceedings of International Conference on Artificial Intelligence and Statistics 2020*, pages 2010–2020. PMLR, 2020.

[105] Nagarajan Natarajan and Inderjit S Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

[106] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

[107] Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019.

[108] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[109] Sean O'Rourke, Van Vu, and Ke Wang. Eigenvectors of random matrices: A survey. *Journal of Combinatorial Theory, Series A*, 144:361–442, 2016.

[110] Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 752–759, 2008.

[111] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[112] Pablo A Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

[113] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[114] James L Peugh and Craig K Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556, 2004.

[115] Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in Neural Information Processing Systems*, volume 28, pages 2107–2115, 2015.

[116] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3): 471–501, 2010.

[117] Oded Regev and Bo'az Klartag. Quantum one-way communication can be exponentially stronger than classical communication. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, pages 31–40, 2011.

[118] R Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[119] Lior Rokach, Francesco Ricci, and Bracha Shapira. *Recommender systems handbook*. Springer, 2015.

[120] Mathieu Rosenbaum and Alexandre B Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[121] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[122] Donald B Rubin. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

[123] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 1987.

[124] Joseph L Schafer and John W Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147, 2002.

[125] Devavrat Shah and Dogyoon Song. Learning RUMs: Reducing mixture to single component via PCA. *arXiv preprint arXiv:1812.11917*, 2018.

[126] Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, volume 31, pages 3111–3121, 2018.

[127] Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[128] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[129] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, volume 31, pages 5186–5196, 2018.

[130] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018.

[131] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[132] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[133] Sasha Slijepcevic, Seapahn Megerian, and Miodrag Potkonjak. Location errors in wireless embedded sensor networks: sources, models, and effects on applications. *ACM SIGMOBILE Mobile Computing and Communications Review*, 6(3):67–78, 2002.

[134] Dogyoon Song and Pablo A Parrilo. On approximations of the PSD cone by a polynomial number of smaller-sized PSD cones. *arXiv preprint arXiv:2105.02080*, 2021.

[135] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

[136] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 18, pages 1329–1336, 2005.

[137] Nicolas Städler and Peter Bühlmann. Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235, 2012.

[138] Gilbert W Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977.

[139] Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic Press, 1990.

[140] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

[141] Richard E Strauss, Momchil N Atanassov, and João Alves De Oliveira. Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. *Journal of Vertebrate Paleontology*, 23(2):284–296, 2003.

[142] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory 2019*, pages 2898–2933. PMLR, 2019.

[143] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[144] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2nd edition, 2018.

[145] Richard Stuart Sutton. *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts, Amherst, 1984.

[146] Craig A Tracy and Harold Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

[147] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[148] John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

[149] John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1):59–94, 1996.

[150] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[151] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[152] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[153] Stef Van Buuren. *Flexible imputation of missing data*. CRC Press, 2nd edition, 2018.

[154] Hado Van Hasselt. Reinforcement learning in continuous state and action spaces. In *Reinforcement Learning*, pages 207–251. Springer, 2012.

[155] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38 (1):49–95, 1996.

[156] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[157] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[158] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4): 279–292, 1992.

[159] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards.* PhD thesis, University of Cambridge, 1989.

[160] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[161] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[162] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[163] Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of International Conference on Machine Learning 2019*, pages 6995–7004. PMLR, 2019.

[164] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Harnessing structures for value-based planning and reinforcement learning. In *Proceedings of International Conference on Learning Representations*, 2020.

[165] Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441–466, 1991.