THE UNIVERSITY OF CHICAGO

FIRST ORDER METHODS FOR NONCONVEX OPTIMIZATION VIA SYMMETRIC FACTORIZATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
QINQING ZHENG

CHICAGO, ILLINOIS
AUGUST 2017

To my parents

It’s simple, but not easy.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

ACKNOWLEDGMENTS ix

ABSTRACT x

1 INTRODUCTION 1

I LOW RANK MATRICES

2 AFFINE RANK MINIMIZATION 5

3 SEMIDEFINITE PROGRAMMING FROM RANDOM LINEAR MEASUREMENTS . 7
- 3.1 Semidefinite Programming and Rank Minimization 8
- 3.2 A Gradient Descent Algorithm for Rank Minimization 9
- 3.3 Convergence Analysis 12
- 3.4 Related Work 15
- 3.5 Experiments 17
  - 3.5.1 Computational Complexity 17
  - 3.5.2 Runtime Comparison 18
  - 3.5.3 Sample Complexity 19
- 3.6 Discussion 20
- 3.7 Proofs 21
  - 3.7.1 Proof of Lemma 3.1 21
  - 3.7.2 Technical Lemmas 22
  - 3.7.3 Linear Convergence 27
  - 3.7.4 Regularity Condition 28
  - 3.7.5 Initialization 33
  - 3.7.6 Sample Complexity 35
  - 3.7.7 ADMM for Nuclear Norm Minimization 44

4 RECTANGULAR MATRIX COMPLETION 46
- 4.1 Semidefinite Lifting, Factorization, and Gradient Descent 47
- 4.2 Main Result: Convergence Analysis 51
  - 4.2.1 Proof Sketch 52
- 4.3 Related Work 54
- 4.4 Experiments 57
  - 4.4.1 Computational Complexity 58
  - 4.4.2 Runtime Comparison 59
  - 4.4.3 Sample Complexity 60
- 4.5 Discussion 60

4.6 Proofs . . . 61
4.6.1 Technical Lemmas . . . 61
4.6.2 Initialization . . . 69
4.6.3 Regularity Condition . . . 72
4.6.4 Linear Convergence . . . 80

II SPARSE GRAPHS

5 FASTEST MIXING MARKOV CHAIN . . . 83
5.1 Graph Laplacian and Cholesky Factorization . . . 83
5.2 Problem Statement . . . 86
5.3 Nonconvex Formulation and First Order Method . . . 87
5.3.1 Sparsity Pattern of Cholesky Factor . . . 87
5.3.2 Algorithm: A Variant of ADMM . . . 89
5.3.3 Computational Complexity . . . 93
5.4 Related Work . . . 95
5.5 Experiments . . . 97
5.5.1 Initialization of Nonconvex ADMM . . . 97
5.5.2 Comparison with Other Methods . . . 99
5.6 Discussion . . . 100
5.7 Proofs . . . 101
5.7.1 Proof of Theorem 5.1 . . . 101

III CONCLUSION, EXTENSIONS AND FUTURE WORK

6 CONCLUSION . . . 106
6.1 Summary . . . 106
6.2 Future Work . . . 106

REFERENCES . . . 109

LIST OF FIGURES

3.1 An instance of $f(Z)$ where $X^\star \in \mathbb{R}^{2\times 2}$ is rank-1 and $Z \in \mathbb{R}^2$. The underlying truth is $Z^\star = [1, 1]^\top$. Both $Z^\star$ and $-Z^\star$ are minimizers. 10
3.2 Linear convergence of the gradient scheme, for $n = 200$, $m = 1000$ and $r = 2$. The distance metric is given in Definition 3.1. 11
3.3 Runtime comparison where $X^\star \in \mathbb{R}^{400\times 400}$ is rank-2 and A_is are dense. 19
3.4 Runtime comparison where $X^\star \in \mathbb{R}^{600\times 600}$ is rank-2 and A_is are sparse. 19
3.5 Sample complexity comparison. 20

4.1 (a) Runtime comparison where $X^\star$ is 4000×2000 and of rank 3. 199057 entries are observed. (b) Magnified plots to compare other methods except `nuclear`. 60
4.2 (a) Runtime growth of `AltMin`, `trustRegresion`, `GD` and `SVP`. (b) Sample complexity of gradient scheme. 61

5.1 For a randomly generated Erdős–Rényi graph, the sparsity patterns of adjacency matrix and Cholesky factors of a few graph Laplacian matrices. 88
5.2 (a) Optimal SLEM and the output of our approach using heuristic initialization. (b) Comparing with random initialization. 98
5.3 Convergence comparison for a random Erdős–Rényi graph. 99
5.4 Convergence comparison for a random graph in stochastic block model. 100

LIST OF TABLES

3.1 Matrix sensing: per-iteration computational complexities of different methods. 17

4.1 Matrix completion: per-iteration computational complexities of different methods. . . . 58

ACKNOWLEDGMENTS

"Happy graduate students come from good advisors and the happiest graduate students come from John Lafferty." When I received the PhD offer from UChicago, I was anxious about math and reluctant to accept it. John encouraged me to join the program, and his endless support over the years has being invaluable to me. He taught me how to enjoy research, and how to tolerate the frustration. This is a memorable journey.

Ryota Tomioka, you are an integral part of my graduate student experience. Working with you is always fun and rewarding, thanks for having taught me so much. I am also fortunate to have Risi Kondor and Lek-Heng Lim, together with their constructive input, on my thesis committee.

I am very grateful to Annie Marsden and Matt Bonakdarpour for collaborating with me on the FMMC problem, which becomes the second part of this thesis. The earlier version of Theorem 5.1 and its proof was mainly contributed by Annie. Our Regenstein hackathons are unforgettable memories to me.

The weekly HELIOS meeting made my research not a lonely ride. It is fortunate to have great peers sharing research ideas and discussing ongoing work. Thanks to the rest of the members of HELIOS, especially the organizers: Rina Foygel Barber, Tracy Ke and Chao Gao.

The passionate students in John's LSDA course gave me wonderful teaching experience. I have enjoyed exchanging knowledge and hearing crazy ideas with you. The jokes and notes you left in the exam are indeed shining moments of my PhD life.

I also would like to thank the rest of the faculty and staff in CS department, who made our small department a lovely family.

My dear friends, Nikita Mishra, Aiman Fang, Yuancheng Zhu, Bumeng Zhuo, Liwen Zhang, Jiajun Shen, David Kim, Wooseok Ha, Min Xu and many many others — you know who you are, I have the good luck of sharing my PhD adventure with you. Special thanks go to Nikita, who helped me on various aspects of my life. Min, I quoted the beautiful sentence from you — so true!

Last, I thank my parents for their unconditional love.

ABSTRACT

In this thesis, we discuss three positive semidefinite matrix estimation problems. We recast them by decomposing the semidefinite variable into symmetric factors, and investigate first-order methods for optimizing the transformed nonconvex objectives.

The central theme of our methods is to exploit the structure of the factors for computational efficiency. The first part of this thesis focuses on low rank structure. We first consider a family of random semidefinite programs. We reformulate the problem as minimizing a fourth order objective function, and propose a simple gradient descent algorithm. With $O(r^3\kappa^2 n \log n)$ random measurements of a positive semidefinite $n \times n$ matrix of rank r and condition number κ, our method is guaranteed to converge linearly to the global optimum.

Similarly, we address the rectangular matrix completion problem by lifting the unknown matrix to a positive semidefinite matrix in higher dimension, and optimizing a fourth order objective over the factor using a simple gradient descent scheme. With $O(\mu r^2\kappa^2 n \max(\mu, \log n))$ random observations of a $n_1 \times n_2$ μ-incoherent matrix of rank r and condition number κ, where $n = \max(n_1, n_2)$, the algorithm linearly converges to the global optimum with high probability.

Sparsity is the other structure we study. In the second part of this thesis, we consider the problem of computing the fastest mixing Markov chain on a given graph. The task is to choose the edge weights so that a function of the eigenvalues of the associated graph Laplacian matrix is minimized. We rewrite this problem so that the search space is over the sparse Cholesky factor of the associated graph Laplacian, and develop a nonconvex ADMM algorithm. Experiments are conducted to demonstrate the convergence of this approach.

CHAPTER 1

INTRODUCTION

A growing body of recent research is shedding new light on the role of nonconvex optimization for tackling large scale problems in machine learning, signal processing, and convex programming. A parallel development is the surprising effectiveness of simple classical procedures such as gradient descent for problems with exploded size and complexity, as explored in the recent literature [Bach and Moulines, 2011; Bach, 2014; Hoffman et al., 2013]. This thesis is devoted to develop relatively simple first-order algorithms for certain nonconvex approaches and explain the remarkable effectiveness and efficiency of them.

Optimizing a nonconvex function is in general hard due to the presence of local minima and saddle points. For the past few decays, there has been extensive studies that focuses on convex relaxation of nonconvex functions [Goemans and Williamson, 1995; Candès, 2006; Donoho, 2006; Recht et al., 2010; Chandrasekaran et al., 2012]. In particular, semidefinite programming has become a key surrogate optimization tool of difficult combinatorial problems [d'Aspremont et al., 2004; Amini and Wainwright, 2009; Goemans and Williamson, 1995]. In spite of the importance of SDPs in principle—promising efficient algorithms with polynomial runtime guarantees—it is widely recognized that current optimization algorithms based on interior point methods can handle only relatively small problems. Thus, a considerable gap exists between the theory and applicability of SDP formulations. Scalable algorithms for semidefinite programming, and closely related families of nonconvex programs more generally, are greatly needed.

The motivating result of this thesis is recent work for *phase retrieval* by Candès et al. [2015b]. The phase retrieval problem is to recover a complex vector $z \in \mathcal{C}^n$ from squared magnitudes of its linear measurements

$$y_i = |\langle a_i, z\rangle|^2, \quad i = 1, \dots m.$$

The authors propose a gradient descent algorithm to optimize a fourth order nonconvex objective

function

$$f(z) = \frac{1}{4m} \sum_{i=1}^{m} \left(y_i - |\langle a_i, z \rangle|^2 \right)^2 . \tag{1.1}$$

Under mild assumptions, using carefully constructed initialization and step size, the iterates converge to global optimum at a linear rate.

If we assume that $a_1, \ldots, a_m, z \in \mathbb{R}^n$, an interesting reparameterization of f is

$$\frac{1}{4m} \sum_{i=1}^{m} \left(y_i - \langle A_i, X \rangle \right)^2 , \tag{1.2}$$

where $A_i = a_i a_i^\top$ and $X = zz^\top$ is a semidefinite variable. This observation has inspired our thinking in two aspects:

1. Local searching such as gradient descent can be effective and computationally efficient for certain nonconvex problems of symmetric structure.

2. For certain families of SDPs, one can nonconvexify the problem by taking symmetric factorization $X = ZZ^\top$ and then solve the resulting nonconvex problem via gradient descent over the factor Z.

In this thesis, we study several problems where the symmetric factorization technique can apply. The first part of our work focuses on utilizing the low rank structure of the semidefinite variable. When Z is of low rank, this can be viewed as part of a framework for solving general low rank semidefinite programs proposed by Burer and Monteiro [2003, 2005]. In Chapter 2, we introduce the affine rank minimization problem. It provides a unified characterization of problems we study in the next two chapters up to certain transformations. In Chapter 3, we consider a class of *SDPs with random linear constraints* where the solution matrix is of low rank. We prove that a simple gradient scheme linearly converges to global optimum with high probability. This work was presented at NIPS 2015 [Zheng and Lafferty, 2015].

As a generalization, Chapter 4 studies a projected gradient descent algorithm for solving *low rank rectangular matrix completion* problem. We introduce a *lifting* method that transforms the

rectangular matrix into a positive semidefinite matrix in higher dimension, so that it can be decomposed in the same way as before. This work is reported in a technical report [Zheng and Lafferty, 2016]. It extends the results in previous chapter in two directions that are of more practical interest: the target matrix is rectangular and the observation is incomplete.

In addition to the low rank model, the second part of this thesis considers sparse structure. Chapter 5 discusses the *fastest mixing Markov chain* problem: finding edge weights of a given graph to achieve the fastest mixing rate. It can be written as a matrix eigenvalue optimization problem, whose variable is the graph Laplacian matrix. The graph Laplacian is positive semidefinite and of nearly full rank, but it has a sparse Cholesky factor. We propose a variant of the ADMM algorithm that optimizes a nonconvex objective over the Cholesky factor with a fixed sparsity pattern.

Finally, we conclude in Chapter 6. Some directions for future work are also provided in this chapter.

Part I

Low Rank Matrices

CHAPTER 2

AFFINE RANK MINIMIZATION

We consider the problem of recovering a unknown low rank matrix $X^\star \in \mathbb{R}^{n_1 \times n_2}$ from m linear measurements

$$b_i = \langle A_i, X^\star \rangle, \quad i = 1, \ldots, m.$$

Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \longrightarrow \mathbb{R}^m$ be the affine transformation such that $\mathcal{A}(\cdot) = \langle A_i, \cdot \rangle$. Our goal is to find a matrix X^* of minimum rank satisfying $\mathcal{A}(X^*) = b$. The underdetermined case where $m \ll n_1 n_2$ is of particular interest, and can be formulated as the optimization

$$\begin{aligned} \min_{X \in \mathbb{R}^{n_1 \times n_2}} \quad & \text{rank}(X) \\ \text{subject to} \quad & \mathcal{A}(X) = b. \end{aligned} \tag{2.1}$$

This problem is a direct generalization of compressed sensing, and subsumes many machine learning problems such as image compression, low rank matrix completion and low-dimensional metric embedding [Recht et al., 2010; Jain et al., 2013].

The challenges that are both statistical and computational in nature.

- *Computationally*, while the problem is natural and has many applications, the objective function is nonconvex. Without conditions on the transformation $\mathcal{A}$ or the minimum rank solution X^*, it is generally NP hard [Meka et al., 2008]. We would like to have an algorithm which converges to X^* in polynomial time, meanwhile has fast convergence rate and low per-iteration cost.

- *Statstically*, we want to achieve exact recovery $X^* = X^\star$ with as few measurements as possible.

We study two instances of Problem 2.1 in this thesis. In Chapter 3, we assume that

(i) $X^\star$ is positive semidefinite, which implies $n_1 = n_2 = n$;

(ii) Each A_i is a random $n \times n$ symmetric matrix from the Gaussian Orthogonal Ensemble (GOE), with $(A_i)_{jj} \sim \mathcal{N}(0, 2)$ and $(A_i)_{jk} \sim \mathcal{N}(0, 1)$ for $j \neq k$.

We shall refer to this problem as *low rank positive semidefinite matrix sensing*. In addition to the wide applicability of affine rank minimization, this problem is also closely connected to a class of semidefinite programs. In Section 3.1, we show that the minimizer of a particular class of SDP can be obtained by a linear transformation of $X^\star$. Thus, efficient algorithms for problem (3.1) can be applied in this setting as well.

In Chapter 4, we consider the rectangular matrix completion problem, a common model for recommendation system. There the transformation $\mathcal{A}$ represents a random sampling operator and b consists of entries of $X^\star$ that are observed.

CHAPTER 3

SEMIDEFINITE PROGRAMMING FROM RANDOM LINEAR MEASUREMENTS

We would like to a reconstruct a positive semidefinite matrix $X^\star$ of minimum rank that satisfies a group of linear constraints. The task is to solve the nonconvex optimization problem

$$\begin{array}{ll} \underset{X \succeq 0}{\min} & \text{rank}(X) \\ \text{subject to} & \text{tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \end{array} \tag{3.1}$$

where A_is are i.i.d variables generated from GOE.

As mentioned in Chapter 1, noting that a rank-r matrix $X^\star$ can be decomposed as $X^\star = Z^\star Z^{\star\top}$ where $Z^\star$ is a n by r matrix, our approach is based on minimizing the squared residual

$$f(Z) = \frac{1}{4m} \left\| \mathcal{A}(ZZ^\top) - b \right\|^2 = \frac{1}{4m} \sum_{i=1}^{m} \left(\text{tr}(Z^\top A_i Z) - b_i \right)^2. \tag{3.2}$$

While this is a nonconvex function, we develop a gradient descent algorithm for optimizing $f(Z)$. Our main contributions concerning this algorithm are as follows.

- We prove that with $O(r^3 n \log n)$ constraints our gradient descent scheme can exactly recover $X^\star$ with high probability. Empirical experiments show that this bound may potentially be improved to $O(rn \log n)$.

- We show that our method converges linearly, and has lower computational cost compared with previous methods.

- We carry out a detailed comparison of rank minimization algorithms, and demonstrate that when the measurement matrices A_i are sparse, our gradient method significantly outperforms alternative approaches.

Later sections are organized as follows. Before presenting our algorithm, we explain the connection between semidefinite programming and rank minimization in Section 3.1. This connection enables our scalable gradient descent algorithm to be applied and analyzed for certain classes of SDPs. In Section 3.2 we discuss the gradient scheme in detail. Our main analytical results are presented in Section 3.3, with detailed proofs contained in the Section 3.7. In Section 3.4 we review related work. Our experimental results are presented in Section 3.5, and we conclude with a brief discussion of future work in Section 3.6.

3.1 Semidefinite Programming and Rank Minimization

Consider a standard form semidefinite program

$$\begin{array}{ll} \underset{\widetilde{X} \succeq 0}{\min} & \text{tr}(\widetilde{C}\widetilde{X}) \\ \text{subject to} & \text{tr}(\widetilde{A}_i \widetilde{X}) = b_i, \;\; i = 1, \ldots, m \end{array} \tag{3.3}$$

where $\widetilde{C}, \widetilde{A}_1, \ldots, \widetilde{A}_m \in \mathbb{S}^n$. If $\widetilde{C}$ is positive definite, then we can write $\widetilde{C} = LL^\top$ where $L \in \mathbb{R}^{n \times n}$ is invertible. It follows that the minimum of problem (3.3) is the same as

$$\begin{array}{ll} \underset{X \succeq 0}{\min} & \text{tr}(X) \\ \text{subject to} & \text{tr}(A_i X) = b_i, \;\; i = 1, \ldots, m \end{array} \tag{3.4}$$

where $A_i = L^{-1}\widetilde{A}_i {L^{-1}}^\top$. In particular, minimizers $\widetilde{X}^*$ of Problem (3.3) are obtained from minimizers X^* of Problem (3.4) via the transformation

$$\widetilde{X}^* = {L^{-1}}^\top X^* L^{-1}.$$

Since X is positive semidefinite, $\text{tr}(X)$ is equal to $\|X\|_*$. Hence, problem (3.4) is the nuclear norm relaxation of Problem (3.1). Next, we characterize the specific cases where $X^* = X^\star$, so that the

SDP and rank minimization solutions coincide.

Theorem 3.1 (Recht et al. [2010]). *Let $\mathcal{A} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^m$ be a linear map. For every integer k with $1 \leq k \leq n$, define the k-restricted isometry constant to be the smallest value δ_k such that*

$$(1 - \delta_k) \|X\|_F \leq \|\mathcal{A}(X)\| \leq (1 + \delta_k) \|X\|_F$$

holds for any matrix X of rank at most k. Suppose that there exists a rank r matrix $X^\star$ such that $\mathcal{A}(X^\star) = b$. If $\delta_{2r} < 1$, then $X^\star$ is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$. Furthermore, if $\delta_{5r} < 1/10$, then $X^\star$ can be attained by minimizing $\|X\|_*$ over the affine subset.

In other words, since $\delta_{2r} \leq \delta_{5r}$, if $\delta_{5r} < 1/10$ holds for the transformation $\mathcal{A}$ and one finds a matrix X of rank r satisfying the affine constraint, then X must be positive semidefinite. Hence, one can ignore the semidefinite constraint $X \succeq 0$ when solving the rank minimization (3.1). The resulting problem then can be exactly solved by nuclear norm relaxation. Since the minimum rank solution is positive semidefinite, it then coincides with the solution of the SDP (3.4), which is a constrained nuclear norm optimization.

3.2 A Gradient Descent Algorithm for Rank Minimization

Our method is described in Algorithm 1. It is parallel to the *Wirtinger Flow* (`WF`) algorithm for phase retrieval [Candès et al., 2015b]. To recover a complex vector $z^\star \in \mathcal{C}^n$ given the squared magnitudes of its linear measurements $b_i = |\langle a_i, z^\star\rangle|^2,\ i \in [m]$, where $a_1, \ldots, a_m \in \mathcal{C}^n$. Candès et al. [2015b] propose a first-order method to minimize the sum of squared residuals

$$f_{\texttt{WF}}(z) = \sum_{i=1}^{n} \left(|\langle a_i, z\rangle|^2 - b_i\right)^2. \tag{3.5}$$

The authors establish the convergence of `WF` to the global optimum—given sufficient measurements, the iterates of `WF` converge linearly to x up to a global phase, with high probability.

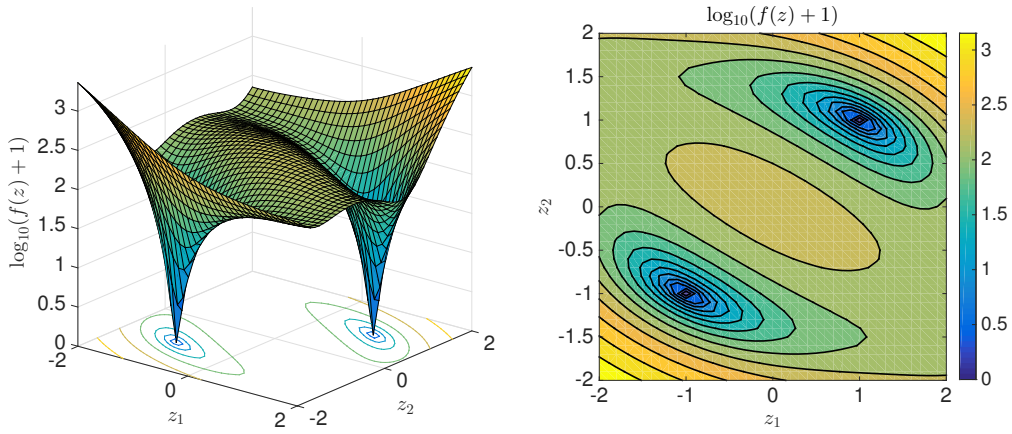


Figure 3.1: An instance of $f(Z)$ where $X^\star \in \mathbb{R}^{2\times 2}$ is rank-1 and $Z \in \mathbb{R}^2$. The underlying truth is $Z^\star = [1, 1]^\top$. Both $Z^\star$ and $-Z^\star$ are minimizers.

If z and the a_is are real-valued, the function $f_{\texttt{WF}}(z)$ can be expressed as

$$f_{\texttt{WF}}(z) = \sum_{i=1}^{n} \left(z^\top a_i a_i^\top z - x^\top a_i a_i^\top x \right)^2,$$

which is a special case of $f(Z)$ in Equation (3.2), where $A_i = a_i a_i^\top$ and each of Z and $X^\star$ are rank one. See Figure 3.1 for an illustration; Figure 3.2 shows the convergence rate of our method. Our methods and results are thus generalizations of Wirtinger flow for phase retrieval.

Before turning to the presentation of our technical results in the following section, we present some intuition and remarks about how and why this algorithm works. For simplicity, let us assume that the rank is specified correctly.

Initialization is of course crucial in nonconvex optimization, as many local minima may be present. To obtain a sufficiently accurate initialization, we use a spectral method, similar to those used in [Netrapalli et al., 2013; Candès et al., 2015b]. The starting point is the observation that a linear combination of the constraint values and matrices yields an unbiased estimate of the solution.

Lemma 3.1. *Let $M = \frac{1}{m}\sum_{i=1}^m b_i A_i$. Then $\frac{1}{2}\mathbb{E}(M) = X^\star$, where the expectation is with respect to the randomness in the measurement matrices A_i.*

Based on this fact, let $X^\star = U^\star \Sigma U^{\star\top}$ be the eigenvalue decomposition of $X^\star$, where $U^\star = [u_1^\star, \ldots, u_r^\star]$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ such that $\sigma_1 \geq \ldots \geq \sigma_r$ are the nonzero eigenvalues of

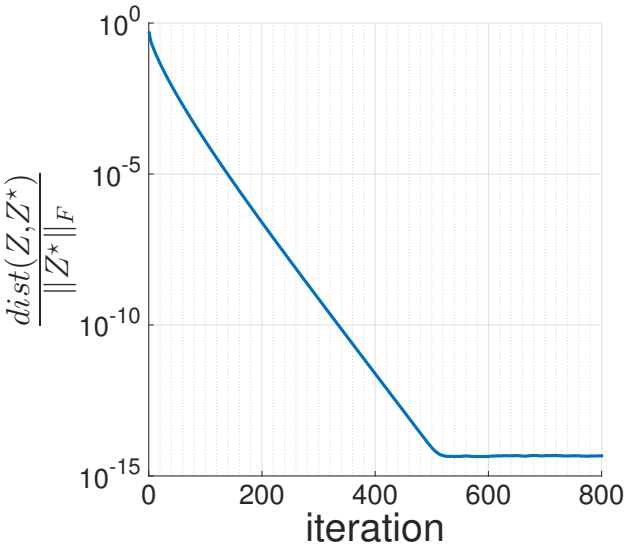


Figure 3.2: Linear convergence of the gradient scheme, for $n = 200$, $m = 1000$ and $r = 2$. The distance metric is given in Definition 3.1.

$X^\star$. Let $Z^\star = U^\star \Sigma^{\frac{1}{2}}$. Clearly, $u_s^\star = z_s^\star / \|z_s^\star\|$ is the top sth eigenvector of $\mathbb{E}(M)$ associated with eigenvalue $2\|z_s^\star\|^2$. Therefore, we initialize according to $z_s^0 = \sqrt{\frac{|\lambda_s|}{2}} v_s$ where (v_s, λ_s) is the top sth eigenpair of M. For sufficiently large m, it is reasonable to expect that Z^0 is close to $Z^\star$; this is confirmed by concentration of measure arguments.

Certain key properties of $f(Z)$ will be seen to yield a linear rate of convergence. In the analysis of convex functions, Nesterov [2004] shows that for unconstrained optimization, the gradient descent scheme with sufficiently small step size will converge linearly to the optimum if the objective function is strongly convex and has a Lipschitz continuous gradient. However, these two properties are global and do not hold for our objective function $f(Z)$. Nevertheless, we expect that similar conditions hold for the local area near $Z^\star$. If so, then if we start close enough to $Z^\star$, we can achieve the global optimum.

In our subsequent analysis, we establish the convergence of Algorithm 1 with a constant step size of the form $\mu / \|Z^\star\|_F^2$, where μ is a small constant. Since $\|Z^\star\|_F$ is unknown, we replace it by $\|Z^0\|_F$.

Algorithm 1: Gradient descent for rank minimization

input: $\{A_i, b_i\}_{i=1}^m, r, \mu$

initialization

Set $(v_1, \lambda_1), \ldots, (v_r, \lambda_r)$ to the top r eigenpairs of $\frac{1}{m}\sum_{i=1}^m b_i A_i$ s.t. $|\lambda_1| \geq \cdots \geq |\lambda_r|$

$Z^0 = [z_1^0, \ldots, z_r^0]$ where $z_s^0 = \sqrt{\frac{|\lambda_s|}{2}} \cdot v_s$, $s \in [r]$

$k \leftarrow 0$

repeat

$\nabla f(Z^k) = \frac{1}{m}\sum_{i=1}^m \left(\text{tr}({Z^k}^\top A_i Z^k) - b_i\right) A_i Z^k$

$Z^{k+1} = Z^k - \frac{\mu}{\sum_{s=1}^r |\lambda_s|/2}\nabla f(Z^k)$

$k \leftarrow k+1$

until *convergence*;

output: $\widehat{X} = Z^k {Z^k}^\top$

3.3 Convergence Analysis

In this section we present our main result analyzing the gradient descent algorithm, and give a sketch of the proof. To begin, note that the symmetric decomposition of $X^\star$ is not unique, since $X^\star = (Z^\star U)(Z^\star U)^\top$ for any $r \times r$ orthonormal matrix U. Thus, the solution set is

$$\mathcal{S} = \left\{ \widetilde{Z} \in \mathbb{R}^{n \times r} \mid \widetilde{Z} = Z^\star U \ \text{ for some } U \text{ with } UU^\top = U^\top U = I \right\}.$$

Note that $\|\widetilde{Z}\|_F^2 = \|X^\star\|_*$ for any $\widetilde{Z} \in \mathcal{S}$. We define the distance to the optimal solution in terms of this set.

Definition 3.1. *Define the distance between Z and $Z^\star$ as*

$$d(Z, Z^\star) = \min_{UU^\top = U^\top U = I} \|Z - Z^\star U\|_F = \min_{\widetilde{Z} \in \mathcal{S}} \left\|Z - \widetilde{Z}\right\|_F.$$

Our main result for exact recovery is stated below, assuming that the rank is correctly specified. Since the true rank is typically unknown in practice, one can start from a very low rank and gradually increase it.

Theorem 3.2. *Let the condition number $\kappa = \sigma_1/\sigma_r$ denote the ratio of the largest to the smallest nonzero eigenvalues of $X^\star$. There exists a universal constant c_0 such that if $m \geq c_0\kappa^2 r^3 n \log n$, with high probability the initialization Z^0 satisfies*

$$d(Z^0, Z^\star) \leq \sqrt{\frac{3}{16}\sigma_r}. \tag{3.6}$$

Moreover, there exists a universal constant c_1 such that when using constant step size $\mu/\left\|Z^\star\right\|_F^2$ with $\mu \leq \frac{c_1}{\kappa n}$ and initial value Z^0 obeying (3.6)*, the kth step of Algorithm 1 satisfies*

$$d(Z^k, Z^\star) \leq \sqrt{\frac{3}{16}\sigma_r}\left(1 - \frac{\mu}{12\kappa r}\right)^{k/2}$$

with high probability.

We now outline the proof, giving full details in the supplementary material. The proof has four main steps. The first step is to give a regularity condition under which the algorithm converges linearly if we start close enough to $Z^\star$. This provides a local regularity property that is similar to the Nesterov [2004] criteria that the objective function is strongly convex and has a Lipschitz continuous gradient.

Definition 3.2. *Let $\overline{Z} = \arg\min_{\widetilde{Z}\in\mathcal{S}} \left\|Z - \widetilde{Z}\right\|_F$ denote the matrix closest to Z in the solution set. We say that f satisfies the regularity condition $RC(\varepsilon, \alpha, \beta)$ if there exist constants α, β such that for any Z satisfying $d(Z, Z^\star) \leq \varepsilon$, we have*

$$\langle \nabla f(Z), Z - \overline{Z}\rangle \geq \frac{1}{\alpha}\sigma_r \left\|Z - \overline{Z}\right\|_F^2 + \frac{1}{\beta\left\|Z^\star\right\|_F^2}\left\|\nabla f(Z)\right\|_F^2.$$

Using this regularity condition, we show that the iterative step of the algorithm moves closer to the optimum, if the current iterate is sufficiently close.

Lemma 3.2. *Consider the update $Z^{k+1} = Z^k - \frac{\mu}{\left\|Z^\star\right\|_F^2}\nabla f(Z^k)$. If f satisfies $RC(\varepsilon, \alpha, \beta)$,*

$d(Z^k, Z^\star) \le \varepsilon$, *and* $0 < \mu < \min(\alpha/2, 2/\beta)$, *then*

$$d(Z^{k+1}, Z^\star) \le \sqrt{1 - \frac{2\mu}{\alpha\kappa r}} d(Z^k, Z^\star).$$

In the next step of the proof, we condition on two events that will be shown to hold with high probability using concentration results. Let δ denote a small value to be specified later.

A1 For any $u \in \mathbb{R}^n$ such that $\|u\| \le \sqrt{\sigma_1}$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (u^\top A_i u) A_i - 2uu^\top \right\| \le \frac{\delta}{r}.$$

A2 For any $\widetilde{Z} \in \mathcal{S}$,

$$\left\| \frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} - \mathbb{E}\left[\frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} \right] \right\| \le \frac{\delta}{r}, \quad \text{for all } s, k \in [r].$$

Here the expectations are with respect to the random measurement matrices. Under these assumptions, we can show that the objective satisfies the regularity condition with high probability.

Lemma 3.3. *Suppose that* ***A1*** *and* ***A2*** *hold. If* $\delta \le \frac{1}{16}\sigma_r$, *then* f *satisfies the regularity condition* $RC(\sqrt{\frac{3}{16}\sigma_r}, 24, 513\kappa n)$ *with probability at least* $1 - mCe^{-\rho n}$, *where* C, ρ *are universal constants.*

Next we show that under **A1**, a good initialization can be found.

Lemma 3.4. *Suppose that* ***A1*** *holds. Let* $\{v_s, \lambda_s\}_{s=1}^r$ *be the top* r *eigenpairs of* $M = \frac{1}{m}\sum_{i=1}^{m} b_i A_i$ *such that* $|\lambda_1| \ge \cdots \ge |\lambda_r|$. *Let* $Z^0 = [z_1, \ldots, z_r]$ *where* $z_s = \sqrt{\frac{|\lambda_s|}{2}} \cdot v_s$, $s \in [r]$. *If* $\delta \le \frac{\sigma_r}{4\sqrt{r}}$, *then*

$$d(Z^0, Z^\star) \le \sqrt{3\sigma_r/16}.$$

Finally, we show that conditioning on **A1** and **A2** is valid since these events have high probability as long as m is sufficiently large.

Lemma 3.5. *If the number of samples* $m \geq \dfrac{42}{\min(\delta^2/r^2\sigma_1^2,\ \delta/r\sigma_1)} n \log n$*, then for any* $u \in \mathbb{R}^n$ *satisfying* $\|u\| \leq \sqrt{\sigma_1}$*,*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (u^\top A_i u) A_i - 2uu^\top \right\| \leq \frac{\delta}{r}$$

holds with probability at least $1 - mCe^{-\rho n} - \frac{2}{n^2}$*, where* C *and* ρ *are universal constants.*

Lemma 3.6. *For any* $x \in \mathbb{R}^n$*, if* $m \geq \dfrac{128}{\min(\delta^2/4r^2\sigma_1^2,\ \delta/2r\sigma_1)} n \log n$*, then for any* $\widetilde{Z} \in \mathcal{S}$

$$\left\| \frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} - \mathbb{E}\left[\frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} \right] \right\| \leq \frac{\delta}{r}, \quad \textit{for all } s, k \in [r],$$

with probability at least $1 - 6me^{-n} - \frac{4}{n^2}$.

Note that since we need $\delta \leq \min\left(\frac{1}{16}, \frac{1}{4\sqrt{r}}\right)\sigma_r$, we have $\frac{\delta}{r\sigma_1} \leq 1$, and the number of measurements required by our algorithm scales as $O(r^3\kappa^2 n \log n)$, while only $O(r^2\kappa^2 n \log n)$ samples are required by the regularity condition. We conjecture this bound could be further improved to be $O(rn \log n)$; this is supported by the experimental results presented below.

Recently, Tu et al. [2016] establish a tighter $O(r^2\kappa^2 n)$ bound overall. Specifically, when only one single `SVP` step is used in preprocessing, the initialization of `PF` is also the spectral decomposition of $\frac{1}{2}M$. The authors show that $O(r^2\kappa^2 n)$ measurements are sufficient for the initial solution to satisfy $d(Z^0, Z^\star) \leq O(\sqrt{\sigma_r})$ with high probability, and demonstrate an $O(rn)$ sample complexity for the regularity condition.

3.4 Related Work

Burer and Monteiro [2003] proposed a general approach for solving semidefinite programs using factored, nonconvex optimization, giving mostly experimental support for the convergence of the algorithms. The first nontrivial guarantee for solving affine rank minimization problem is given by Recht et al. [2010], based on replacing the rank function by the convex surrogate nuclear norm, as already mentioned in the previous section. While this is a convex problem, solving it in practice is

nontrivial, and a variety of methods have been developed for efficient nuclear norm minimization. The most popular algorithms are proximal methods that perform singular value thresholding [Cai et al., 2010] at every iteration. While effective for small problem instances, the computational expense of the SVD prevents the method from being useful for large scale problems.

Recently, Jain et al. [2010] proposed a projected gradient descent algorithm `SVP` (Singular Value Projection) that solves

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times p}} \quad & \|\mathcal{A}(X) - b\|^2 \\ \text{subject to} \quad & \text{rank}(X) \leq r, \end{aligned}$$

where $\|\cdot\|$ is the ℓ_2 vector norm and r is the input rank. In the $(t+1)$th iteration, `SVP` updates X^{t+1} as the best rank r approximation to the gradient update $X^t - \mu \mathcal{A}^\top(\mathcal{A}(X^t) - b)$, which is constructed from the SVD. If $\text{rank}(X^\star) = r$, then `SVP` can recover $X^\star$ under a similar RIP condition as the nuclear norm heuristic, and enjoys a linear numerical rate of convergence. Yet `SVP` suffers from the expensive per-iteration SVD for large problem instances.

Subsequent work of Jain et al. [2013] proposes an alternating least squares algorithm `AltMinSense` that avoids the per-iteration SVD. `AltMinSense` factorizes X into two factors $U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{p \times r}$ such that $X = UV^\top$ and minimizes the squared residual $\left\|\mathcal{A}(UV^\top) - b\right\|^2$ by updating U and V alternately. Each update is a least squares problem. The authors show that the iterates obtained by `AltMinSense` converge to $X^\star$ linearly under a RIP condition. However, the least squares problems are often ill-conditioned, it is difficult to observe `AltMinSense` converging to $X^\star$ in practice.

As described above, considerable progress has been made on algorithms for rank minimization and certain semidefinite programming problems. Yet truly efficient, scalable and provably convergent algorithms have not yet been obtained. In the specific setting that $X^\star$ is positive semidefinite, our algorithm exploits this structure to achieve these goals. We note that recent and independent work of Tu et al. [2016] proposes a hybrid algorithm called *Procrustes Flow* (`PF`), which uses a few iterations of `SVP` as initialization, and then applies gradient descent. Similar algorithms and related problems are also analyzied in Chen and Wainwright [2015]; Bhojanapalli et al. [2016a].

Method	Complexity
nuclear norm minimization via ADMM	$O(mn^2\rho + m^2 + n^3)$
gradient descent	$O(mn^2\rho) + 2n^2r$
`SVP`	$O(mn^2\rho + n^2r)$
`AltMinSense`	$O(mn^2r^2 + n^3r^3 + mn^2r\rho)$

Table 3.1: Matrix sensing: per-iteration computational complexities of different methods.

3.5 Experiments

In this section we report the results of experiments on synthetic datasets. We compare our gradient descent algorithm with nuclear norm relaxation, `SVP` and `AltMinSense` for which we drop the positive semidefiniteness constraint, as justified by the observation in Section 3.1. We use ADMM for the nuclear norm minimization, based on the algorithm for the mixture approach in Tomioka et al. [2010]; see Section 3.7.7. For simplicity, we assume that `AltMinSense`, `SVP` and the gradient scheme know the true rank. Krylov subspace techniques such as the Lanczos method could be used compute the partial eigendecomposition; we use the randomized algorithm of Halko et al. [2011] to compute the low rank SVD. All methods are implemented in MATLAB and the experiments were run on a MacBook Pro with a 2.5GHz Intel Core i7 processor and 16 GB memory.

3.5.1 Computational Complexity

It is instructive to compare the per-iteration cost of the different approaches; see Table 3.1. Suppose that the density (fraction of nonzero entries) of each A_i is ρ. For `AltMinSense`, the cost of solving the least squares problem is $O(mn^2r^2 + n^3r^3 + mn^2r\rho)$. The other three methods have $O(mn^2\rho)$ cost to compute the affine transformation. For the nuclear norm approach, the $O(n^3)$ cost is from the SVD and the $O(m^2)$ cost is due to the update of the dual variables. The gradient scheme requires $2n^2r$ operations to compute $Z^k {Z^k}^\top$ and to multiply Z^k by $n \times n$ matrix to obtain the gradient. `SVP` needs $O(n^2r)$ operations to compute the top r singular vectors. However, in practice this partial SVD is more expensive than the $2n^2r$ cost required for the matrix multiplies

in the gradient scheme.

Clearly, `AltMinSense` is the least efficient. For the other approaches, in the dense case (ρ large), the affine transformation dominates the computation. Our method removes the overhead caused by the SVD. In the sparse case (ρ small), the other parts dominate and our method enjoys a low cost.

3.5.2 Runtime Comparison

We conduct experiments for both dense and sparse measurement matrices. `AltMinSense` is indeed slow, so we do not include it here.

In the first scenario, we randomly generate a 400×400 rank-2 matrix $X^\star = xx^\top + yy^\top$ where $x, y \sim \mathcal{N}(0, I)$. We also generate $m = 6n$ matrices $A_1, \ldots, A_m$ from the GOE, and then take $b = \mathcal{A}(X^\star)$. We report the relative error measured in the Frobenius norm defined as $\|\widehat{X} - X^\star\|_F / \|X^\star\|_F$. For the nuclear norm approach, we set the regularization parameter to $\lambda = 10^{-5}$. We test three values $\eta = 10, 100, 200$ for the penalty parameter and select $\eta = 100$ as it leads to the fastest convergence. Similarly, for `SVP` we evaluate the three values $5\times10^{-5}, 10^{-4}, 2\times10^{-4}$ for the step size, and select 10^{-4} as the largest for which `SVP` converges. For our approach, we test the three values $0.6, 0.8, 1.0$ for μ and select 0.8 in the same way.

In the second scenario, we use a more general and practical setting. We randomly generate a rank-2 matrix $X^\star \in \mathbb{R}^{600\times600}$ as before. We generate $m = 7n$ sparse A_is whose entries are i.i.d. Bernoulli:

$$(A_i)_{jk} = \begin{cases} 1 & \text{with probability } \rho, \\ 0 & \text{with probability } 1-\rho, \end{cases}$$

where we use $\rho = 0.001$. For all the methods we use the same strategies as before to select parameters. For the nuclear norm approach, we try three values $\eta = 10, 100, 200$ and select $\eta = 100$. For `SVP`, we test the three values $5 \times 10^{-3}, 2 \times 10^{-3}, 10^{-3}$ for the step size and select 10^{-3}. For the gradient algorithm, we check the three values $0.8, 1, 1.5$ for μ and choose 1.

The results are shown in Figures 3.3 and 3.4. In the dense case, our method is faster than the

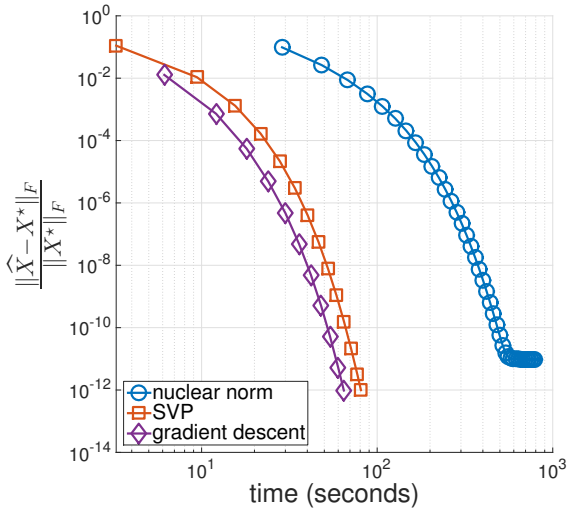


Figure 3.3: Runtime comparison where $X^\star \in \mathbb{R}^{400\times 400}$ is rank-2 and A_is are dense.

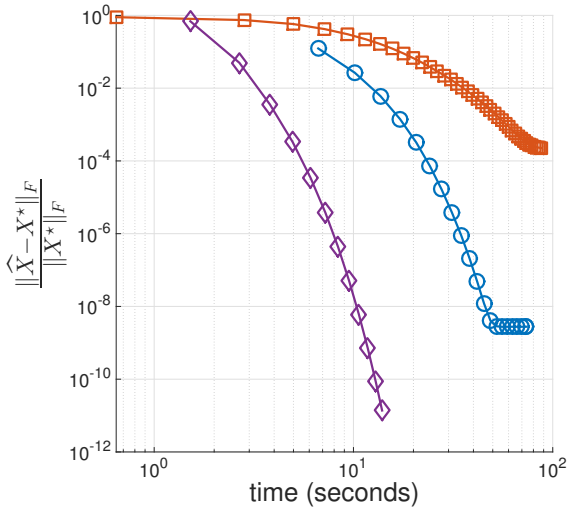


Figure 3.4: Runtime comparison where $X^\star \in \mathbb{R}^{600\times 600}$ is rank-2 and A_is are sparse.

nuclear norm approach and slightly outperforms `SVP`. In the sparse case, it is significantly faster than the other approaches.

3.5.3 Sample Complexity

We also evaluate the number of measurements required by each method to exactly recover $X^\star$, which we refer to as the *sample complexity*. We randomly generate the true matrix $X^\star \in \mathbb{R}^{n\times n}$ and compute the solutions of each method given m measurements, where the A_is are randomly drawn from the GOE. A solution with relative error below 10^{-5} is considered to be successful. We

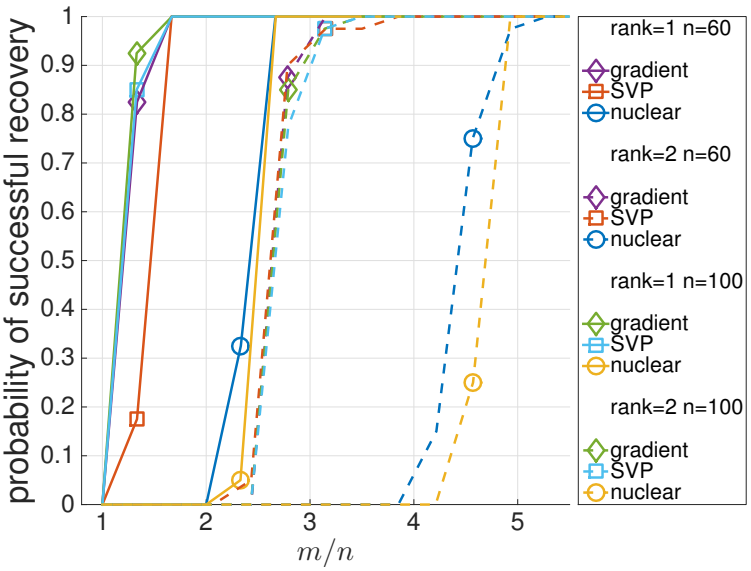


Figure 3.5: Sample complexity comparison.

run 40 trials and compute the empirical probability of successful recovery.

We consider cases where $n = 60$ or 100 and $X^\star$ is of rank one or two. The results are shown in Figure 3.5. For `SVP` and our approach, the phase transitions happen around $m = 1.5n$ when $X^\star$ is rank-1 and $m = 2.5n$ when $X^\star$ is rank-2. This scaling is close to the number of degrees of freedom in each case; this confirms that the sample complexity scales linearly with the rank r. The phase transition for the nuclear norm approach occurs later. The results suggest that the sample complexity of our method should also scale as $O(rn \log n)$ as for `SVP` and the nuclear norm approach [Jain et al., 2010; Recht et al., 2010].

3.6 Discussion

We connect a special case of affine rank minimization to a class of semidefinite programs with random constraints. Building on a recently proposed first-order algorithm for phase retrieval [Candès et al., 2015b], we develop a gradient descent procedure for rank minimization and establish convergence to the optimal solution with $O(r^3 n \log n)$ measurements. We conjecture that $O(rn \log n)$ measurements are sufficient for the method to converge, and that the conditions on the sampling matrices A_i can be significantly weakened. More broadly, the technique used in this paper—

factoring the semidefinite matrix variable, recasting the convex optimization as a nonconvex optimization, and applying first-order algorithms—first proposed by Burer and Monteiro [2003], may be effective for a much wider class of SDPs, and deserves further study.

3.7 Proofs

3.7.1 Proof of Lemma 3.1

Let $A = (a_{ij})$ be a random matrix that is GOE distributed; thus $a_{ij} \sim \mathcal{N}(0,1)$ for $i \neq j$ and $a_{ii} \sim \mathcal{N}(0,2)$. We have $\mathbb{E}(M) = \sum_{s=1}^{r} \mathbb{E}((z_s^\star{}^\top A z_s^\star)A)$. Hence, it suffices to show that $\mathbb{E}((x^\top Ax)A) = 2xx^\top$ for any $x \in \mathbb{R}^n$. The (i,j) entry of $(x^\top Ax)A$ has expected value

$$\begin{aligned}
\mathbb{E}((x^\top Ax)a_{ij}) &= \mathbb{E}\left(\sum_k \sum_l x_k x_l a_{kl} a_{ij}\right) \\
&= \sum_k \sum_l x_k x_l \mathbb{E}(a_{kl} a_{ij}) \\
&= \sum_k \sum_l x_k x_l \cdot \begin{cases} 0 & \text{if } (k,l) \neq (i,j) \wedge (k,l) \neq (j,i) \\ \mathbb{E}(a_{kl}^2) & \text{otherwise} \end{cases} \\
&= \begin{cases} 2x_i x_j \mathbb{E}(a_{ij}^2) & \text{if } i \neq j \\ x_i^2 \mathbb{E}(a_{ii}^2) & \text{otherwise} \end{cases} \\
&= \begin{cases} 2x_i x_j & \text{if } i \neq j, \\ 2x_i^2 & \text{otherwise,} \end{cases}
\end{aligned}$$

where we use that the variance of a_{ii} is 2 and the variance of a_{ij} is 1 for any $i \neq j$. In matrix form, this is $\mathbb{E}((x^\top Ax)A) = 2xx^\top$.

3.7.2 Technical Lemmas

We first present some technical lemmas that will be needed later. Recall Definition 3.2 that for any Z, $\overline{Z} = \arg\min_{\widetilde{Z}\in\mathcal{S}} \left\|Z - \widetilde{Z}\right\|_F$. Let $H = Z - \overline{Z}$. The sth column of Z, $\overline{Z}$, $Z^\star$, H are denoted by z_s, $\bar{z}_s$, $z_s^\star$, h_s respectively. We shall use the following formulas for the gradient and second order partial derivatives:

$$\begin{aligned}
\nabla f(Z) &= \frac{1}{m}\sum_{i=1}^{m}\Big(\mathrm{tr}(H^\top A_i H) + 2\,\mathrm{tr}(\overline{Z}^\top A_i H)\Big)\,(A_i H + A_i\overline{Z}),\\
\frac{\partial^2 f(Z)}{\partial z_s \partial z_s^\top} &= \frac{1}{m}\sum_{i=1}^{m}\Big(2A_i z_s z_s^\top A_i^\top + \Big(\mathrm{tr}(Z^\top A_i Z) - b_i\Big)A_i\Big), \quad \forall s\in[r],\\
\frac{\partial^2 f(Z)}{\partial z_s \partial z_k^\top} &= \frac{1}{m}\sum_{i=1}^{m} 2A_i z_s z_k^\top A_i^\top, \quad \forall s,k\in[r] \text{ such that } s\neq k.
\end{aligned}$$

The next ingredient we need is the expectation of the second order partial derivatives with respect to the random measurement matrices.

Lemma 3.7. *Let $A = (a_{ij})$ be a GOE distributed random matrix. For any two fixed vectors x and y, we have $\mathbb{E}\left[AxyA\right] = x^\top y I + yx^\top$.*

Proof. The expectation of (i,j) entry of $Axy^\top A$ is

$$\mathbb{E}[(Axy^\top A)_{ij}] = \mathbb{E}\left(\sum_{k\,l} a_{ik}a_{jk}x_k y_l\right).$$

If $i = j$, then we have

$$\mathbb{E}[(Axy^\top A)_{ii}] = \mathbb{E}\left(\sum_k a_{ik}^2 x_k y_k\right) = \sum_k x_k y_k + x_i y_i,$$

since $\mathbb{V}\mathrm{ar}(a_{ii}^2) = 2$ and $\mathbb{V}\mathrm{ar}(a_{ik}^2) = 1$ if $k\neq i$. On the other hand, if $i\neq j$, then

$$\mathbb{E}[(Axy^\top A)_{ij}] = \mathbb{E}\left(\sum_{kl} a_{ik}a_{jl}x_k y_l\right) = \mathbb{E}(a_{ij}^2 x_j y_i) = x_j y_i.$$

Therefore, $\mathbb{E}(Axy^\top A) = x^\top y I + yx^\top$. □

Lemma 3.8. *For all $s \in [r]$, it holds that $\mathbb{E}\left[\frac{\partial^2 f(Z)}{\partial z_s \partial z_s^\top}\right] = 2\|z_s\|^2 I + 2z_s z_s^\top + 2ZZ^\top - 2X^\star$ and $\mathbb{E}\left[\frac{\partial^2 f(Z)}{\partial z_s \partial z_k^\top}\right] = 2z_s^\top z_k I + 2z_k z_s^\top$ for all $k \in [r]$ such that $k \neq s$, where the expectation is over the random measurement matrices.*

Proof. The case where $k \neq s$ is a direct result of Lemma 3.7. For the other case, let $A = (a_{ij})$ be a GOE distributed random matrix. It follows from Lemma 3.1 that

$$\mathbb{E}\left[\frac{\partial^2 f(Z)}{\partial z_s \partial z_s^\top}\right] = 2\mathbb{E}(Az_s z_s^\top A) + 2ZZ^\top - 2X^\star.$$

By Lemma 3.7, we have

$$\mathbb{E}(Az_s z_s^\top A) = \|z_s\|^2 I + z_s z_s^\top.$$

Substituting this back into the above equation, we obtain the lemma. □

We next recall a concentration result for the operator (spectral) norm of the random measurement matrices.

Lemma 3.9. *(Ledoux and Rider [2010, Theorem 1]) There exists two absolute constants C and $\rho = \frac{1}{\sqrt{8C}}$ such that with probability at least $1 - Ce^{-\rho n}$,*

$$\|A_i\| \leq 3\sqrt{n}.$$

A tighter upper bound is actually given in the *Tracy-Widow law*: w.h.p. $\|A_i\| = O(2\sqrt{n} + n^{1/6})$.

Corollary 3.1. *With probability at least $1 - mCe^{-\rho n}$, the average of the squared operator norm of the random measurement matrices is upper bounded by $9n$.*

Proof. Applying a union bound we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\|A_i\|^2 \le 9n\right) &\ge \mathbb{P}\left(\forall i,\ \|A_i\| \le 3\sqrt{n}\right)\\ &\ge 1-\sum_{i=1}^{m}\mathbb{P}\left(\|A_i\| > 3\sqrt{n}\right)\\ &\ge 1 - mCe^{-\rho n},\end{aligned}$$

where we use Lemma 3.9 in the last line. □

The following two technical lemmas are important tools for us. Define the set

$$E(\varepsilon) = \{Z \mid d(Z, Z^\star) \le \varepsilon\}.$$

Lemma 3.10. *Suppose that **A1** holds:* $\left\|\frac{1}{m}\sum_{i=1}^{m}(u^\top A_i u)A_i - 2uu^\top\right\| \le \frac{\delta}{r}$*, for any* u *such that* $\|u\| \le \sqrt{\sigma_1}$*. If* $\delta \le \frac{1}{16}\sigma_r$*, then for any* $Z \in E\left(\sqrt{\frac{3}{16}\sigma_r}\right)$ *it holds that*

$$2\left\|HH^\top\right\|_F^2 - \delta\|H\|_F^2 \le \frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(H^\top A_i H)^2 \le \delta\|H\|_F^2 + 2\left\|HH^\top\right\|_F^2.$$

Proof. Let h_s be the sth column of H. Since $\max_{s\in[r]}\|h_s\|_2 \le \|H\|_F \le \sqrt{\frac{3}{16}\sigma_r} \le \sqrt{\sigma_1}$, it follows from the assumption of the lemma that

$$\left\|\frac{1}{m}\sum_{i=1}^{m}(h_s^\top A_i h_s)A_i - 2h_s h_s^\top\right\| \le \frac{\delta}{r}, \quad s = 1,\dots,r.$$

By the triangle inequality, we have

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{s=1}^{r}(h_s^\top A_i h_s)A_i - 2\sum_{s=1}^{r}h_s h_s^\top\right\| \le \delta$$

and consequently

$$-\delta \|h_s\|^2 \leq h_s^\top \left(\frac{1}{m} \sum_{i=1}^m \mathrm{tr}(H^\top A_i H) A_i - 2HH^\top \right) h_s^\top \leq \delta \|h_s\|^2 , \ s = 1, \ldots, r,$$

where we replace $\sum_{s=1}^r h_s^\top A_i h_s$ by $\mathrm{tr}(H^\top A_i H)$ and $\sum_{s=1}^r h_s h_s^\top$ by $HH^\top$. Taking the sum of the above inequalities, we obtain

$$-\delta \|H\|_F^2 \leq \frac{1}{m} \sum_{i=1}^m \mathrm{tr}(H^\top A_i H)^2 - 2\,\mathrm{tr}(H^\top HH^\top H) \leq \delta \|H\|_F^2 .$$

Note that $\mathrm{tr}(H^\top HH^\top H) = \left\| HH^\top \right\|_F^2$. Therefore,

$$2 \left\| HH^\top \right\|_F^2 - \delta \|H\|_F^2 \leq \frac{1}{m} \sum_{i=1}^m \mathrm{tr}(H^\top A_i H)^2 \leq \delta \|H\|_F^2 + 2 \left\| HH^\top \right\|_F^2 .$$

□

Lemma 3.11. *Suppose that **A2** holds: for any $\widetilde{Z}$ such that $\widetilde{Z}\widetilde{Z}^\top = X^\star$ we have*

$$\left\| \frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} - \mathbb{E}\left[\frac{\partial^2 f(\widetilde{Z})}{\partial \widetilde{z}_s \partial \widetilde{z}_k^\top} \right] \right\| \leq \frac{\delta}{r}, \quad s, k = 1, \ldots, r. \tag{3.7}$$

Then

$$\left(\sigma_r - \frac{\delta}{2} \right) \|H\|_F^2 + \left\| H^\top \overline{Z} \right\|_F^2 \leq \frac{1}{m} \sum_{i=1}^m tr(H^\top A_i \overline{Z})^2 \leq \left(\sigma_1 + \frac{\delta}{2} \right) \|H\|_F^2 + \left\| H^\top \overline{Z} \right\|_F^2 .$$

Proof. Our goal is to bound $\frac{1}{m} \sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})^2$. This can be expanded as

$$\frac{1}{m} \sum_{i=1}^m \left(\sum_{s=1}^r (h_s^\top A_i \bar{z}_s) \right)^2 = \frac{1}{m} \sum_{i=1}^m \sum_{s=1}^r (h_s^\top A_i x_s)^2 + \frac{1}{m} \sum_{i=1}^m \sum_{s<k} 2(h_s^\top A_i x_s)(h_k^\top A_i x_k).$$

We first bound the sum of the quadratic terms. For any $s \in [r]$, we have

$$\frac{\partial^2 f(\overline{Z})}{\partial \bar{z}_s \partial \bar{z}_s^\top} = \frac{1}{m}\sum_{i=1}^m 2A_i \bar{z}_s \bar{z}_s^\top A_i,$$
$$\mathbb{E}\left[\frac{\partial^2 f(\overline{Z})}{\partial \bar{z}_s \partial \bar{z}_s^\top}\right] = 2\left\|\bar{z}_s\right\|^2 I + 2\bar{z}_s\bar{z}_s^\top.$$

It follows from assumption (3.7) that for any $s \in [r]$,

$$-\frac{\delta}{r}\left\|h_s\right\|^2 \le \frac{1}{m}\sum_{i=1}^m 2(h_s^\top A_i \bar{z}_s)^2 - 2\left\|\bar{z}_s\right\|^2\left\|h_s\right\|^2 - 2(h_s^\top \bar{z}_s)^2 \le \frac{\delta}{r}\left\|h_s\right\|^2.$$

Taking the sum of above inequalities, we obtain

$$-\frac{\delta}{2r}\sum_{s=1}^r \left\|h_s\right\|^2 \le \frac{1}{m}\sum_{i=1}^m\sum_{s=1}^r (h_s^\top A_i \bar{z}_s)^2 - \sum_{s=1}^r \left\|\bar{z}_s\right\|^2\left\|h_s\right\|^2 - \sum_{s=1}^r (h_s^\top \bar{z}_s)^2 \le \frac{\delta}{2r}\sum_{s=1}^r \left\|h_s\right\|^2. \tag{3.8}$$

Similarly, we bound the sum of the cross terms. For any fixed s, k such that $s \neq k$, we have

$$\frac{\partial^2 f(\overline{Z})}{\partial \bar{z}_s \partial \bar{z}_k^\top} = \frac{1}{m} f(\overline{Z}) \sum_{i=1}^m 2A_i \bar{z}_s \bar{z}_k^\top A_i,$$
$$\mathbb{E}\left[\frac{\partial^2 f(\overline{Z})}{\partial \bar{z}_s \partial \bar{z}_k^\top}\right] = 2\bar{z}_s^\top \bar{z}_k I + 2\bar{z}_k \bar{z}_s^\top,$$

and consequently

$$-\frac{\delta}{r}\sum_{s<k}\left\|h_s\right\|\left\|h_k\right\| \le \frac{1}{m}\sum_{i=1}^m\sum_{s<k} 2(h_s^\top A_i \bar{z}_s)(h_k^\top A_i \bar{z}_k) - 2\sum_{s<k}\bar{z}_s^\top \bar{z}_k h_s^\top h_k - 2\sum_{s<k} h_s^\top \bar{z}_k \bar{z}_s^\top h_k \tag{3.9}$$
$$\le \frac{\delta}{r}\sum_{s<k}\left\|h_s\right\|\left\|h_k\right\|.$$

We combine equations (3.9) and (3.8) to get

$$-\frac{\delta}{2r}\sum_{sk}\|h_s\|\,\|h_k\| \le \frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(H^\top A_i\overline{Z})^2 - \sum_{sk}\bar{z}_s^\top \bar{z}_k h_s^\top h_k - \sum_{sk} h_s^\top \bar{z}_k \bar{z}_s^\top h_k \le \frac{\delta}{2r}\sum_{sk}\|h_s\|\,\|h_k\|\,. \tag{3.10}$$

Note that $\sum_{sk} h_s^\top \bar{z}_k \bar{z}_s^\top h_k = \mathrm{tr}(H^\top \overline{Z} H^\top \overline{Z})$, $\sum_{sk} \bar{z}_s^\top \bar{z}_k h_s^\top h_k = \left\|\overline{Z}H^\top\right\|_F^2$ and

$$\sum_{sk}\|h_s\|\,\|h_k\| = \left(\sum_{s=1}^{r}\|h_s\|\right)^2 \le r\sum_{s=1}^{r}\|h_s\|^2 = r\,\|H\|_F^2\,.$$

By Lemma 3.12, $\mathrm{tr}(H^\top \overline{Z} H^\top \overline{Z}) = \left\|H^\top\overline{Z}\right\|_F^2$. Replacing those terms in equation (3.10) gives us

$$-\frac{\delta}{2}\|H\|_F^2 + \left\|\overline{Z}H^\top\right\|_F^2 + \left\|H^\top\overline{Z}\right\|_F^2 \le \frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(H^\top A_i\overline{Z})^2 \le \frac{\delta}{2}\|H\|_F^2 + \left\|\overline{Z}H^\top\right\|_F^2 + \left\|H^\top\overline{Z}\right\|_F^2\,.$$

Finally, we obtain the claim by noticing that

$$\sqrt{\sigma_r}\,\|H\|_F \le \left\|\overline{Z}H^\top\right\|_F \le \sqrt{\sigma_1}\,\|H\|_F\,,$$

where $\sqrt{\sigma_1} = \sigma_{\max}(\overline{Z}) \ge \cdots \ge \sigma_{\min}(\overline{Z}) = \sqrt{\sigma_r}$ are the singular values of $\overline{Z}$. □

Lemma 3.12. *tr*$(H^\top \overline{Z} H^\top \overline{Z}) = \left\|H^\top\overline{Z}\right\|_F^2$.

Proof. Let $\overline{U} = \arg\min_{UU^\top = U^\top U = I}\|Z - Z^\star U\|_F^2 = \arg\max_{UU^\top = U^\top U = I}\langle U, Z^{\star\top} Z\rangle$. Note that $\langle A, B\rangle \le \|A\|_*\,\|B\|$ for any matrices A, B that are of the same size. The equality holds when $B = U_A V_A^\top$ where $A = U_A \Sigma_A V_A^\top$ is the SVD of A. Hence, $\overline{U} = \widetilde{U}\widetilde{V}^\top$ where $\widetilde{U}\widetilde{S}\widetilde{V}^\top$ is the SVD of $Z^{\star\top} Z$; $\overline{Z} = Z^\star \overline{U}$. Therefore, $Z^\top \overline{Z} = Z^\top Z^\star \overline{U} = \widetilde{V}\widetilde{S}\widetilde{V}^\top$ is symmetric and positive semidefinite. Thus, $H^\top \overline{Z} = Z^\top \overline{Z} - \overline{Z}^\top \overline{Z}$ is also symmetric. This implies that $\mathrm{tr}(H^\top \overline{Z} H^\top \overline{Z}) = \left\|H^\top\overline{Z}\right\|_F^2$. □

3.7.3 Linear Convergence

Proof of Theorem 3.2

Let $H^k = Z^k - \overline{Z}^k$. Then we have that

$$\begin{aligned}
\left\|Z^{k+1} - \overline{Z}^k\right\|_F^2 &= \left\|Z^k - \frac{\mu}{\|Z^\star\|_F^2}\nabla f(Z^k) - \overline{Z}^k\right\|_F^2 \\
&= \left\|H^k\right\|_F^2 + \frac{\mu^2}{\|Z^\star\|_F^4}\left\|\nabla f(Z^k)\right\|_F^2 - \frac{2\mu}{\|Z^\star\|_F^2}\langle \nabla f(Z^k), H^k\rangle \\
&\leq \left\|H^k\right\|_F^2 + \frac{\mu^2}{\|Z^\star\|_F^4}\left\|\nabla f(Z^k)\right\|_F^2 - \frac{2\mu}{\|Z^\star\|_F^2}\left(\frac{1}{\alpha}\sigma_r\left\|H^k\right\|_F^2 + \frac{1}{\beta\|Z^\star\|_F^2}\left\|\nabla f(Z^k)\right\|_F^2\right) \\
&= \left(1 - \frac{2\mu}{\alpha}\cdot\frac{\sigma_r}{\sum_{s=1}^r \sigma_s}\right)\left\|H^k\right\|_F^2 + \frac{\mu(\mu - 2/\beta)}{\|Z^\star\|_F^4}\left\|\nabla f(Z^k)\right\|_F^2 \\
&\leq \left(1 - \frac{2\mu}{\alpha}\cdot\frac{\sigma_r}{r\sigma_1}\right)\left\|H^k\right\|_F^2 \\
&= \left(1 - \frac{2\mu}{\alpha\kappa r}\right) d(Z^k, Z^\star)^2,
\end{aligned}$$

where we use the definition of $RC(\varepsilon, \alpha, \beta)$ in the third line, $\|Z^\star\|_F^2 = \|X^\star\|_* = \sum_{s=1}^r \sigma_s$ in the third to last line and $0 < \mu < \min\{\alpha/2, 2/\beta\}$ in the second to last line. Therefore,

$$d(Z^{k+1}, Z^\star) = \min_{\widetilde{Z}\in\mathcal{S}}\left\|Z^{k+1} - \widetilde{Z}\right\|_F^2 \leq \sqrt{1 - \frac{2\mu}{\alpha\kappa r}}\, d(Z^k, Z^\star).$$

3.7.4 Regularity Condition

As mentioned before, Nesterov [2004, Theorem 2.1.11] shows that the gradient scheme converges linearly under a condition similar to the regularity condition, which is satisfied if the function is strongly convex and has a Lipschitz continuous gradient (*strongly smooth*). In order to prove Lemma 3.3, we show that with high probability the function f satisfies the local curvature condition, which is analogous to strong convexity, and the local smoothness condition, which is analogous to strong smoothness.

C1 *Local Curvature Condition*

There exists a constant C_1 such that for any Z satisfying $d(Z, Z^\star) \le \sqrt{\frac{3}{16}\sigma_r}$,

$$\langle \nabla f(Z), Z - \overline{Z} \rangle \ge C_1 \left\| Z - \overline{Z} \right\|_F^2 + \left\| (Z - \overline{Z})^\top \overline{Z} \right\|_F^2 .$$

C2 *Local Smoothness Condition*

There exist constants C_2, C_3 such that for any Z satisfying $d(Z, Z^\star) \le \sqrt{\frac{3}{16}\sigma_r}$,

$$\|\nabla f(Z)\|_F^2 \le C_2 \left\| Z - \overline{Z} \right\|_F^2 + C_3 \left\| (Z - \overline{Z})^\top \overline{Z} \right\|_F^2 .$$

Proof of the Local Curvature Condition

$$\begin{aligned}
\langle \nabla f(Z), H \rangle = & \overbrace{\frac{2}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})^2}^{p^2} + \overbrace{\frac{1}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i H)^2}^{q^2} + \frac{3}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})\,\mathrm{tr}(H^\top A_i H) \\
\ge & \; p^2 + q^2 - \frac{3}{m}\sqrt{\sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})^2}\sqrt{\sum_{i=1}^m \mathrm{tr}(H^\top A_i H)^2} \\
= & \; p^2 + q^2 - \frac{3}{\sqrt{2}}\overbrace{\sqrt{\frac{2}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})^2}}^{p}\overbrace{\sqrt{\frac{1}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i H)^2}}^{q} \\
= & \left(p - \frac{3}{2\sqrt{2}}q\right)^2 - \frac{1}{8}q^2 \\
\ge & \left(\frac{p^2}{2} - \frac{9}{8}q^2\right) - \frac{1}{8}q^2 \\
= & \; \frac{p^2}{2} - \frac{5}{4}q^2 = \frac{1}{m}\sum_{i=1}^m \mathrm{tr}(H^\top A_i \overline{Z})^2 - \frac{5}{4}\frac{1}{m}\sum_i \mathrm{tr}(H^\top A_i H)^2 \\
\ge & \left(\sigma_r - \frac{\delta}{2}\right)\|H\|_F^2 + \left\|H^\top \overline{Z}\right\|_F^2 - \frac{5\delta}{4}\|H\|_F^2 - \frac{5}{2}\left\|HH^\top\right\|_F^2 \\
\ge & \left(\sigma_r - \frac{5}{2}\|H\|_F^2 - \frac{7}{4}\delta\right)\|H\|_F^2 + \left\|H^\top \overline{Z}\right\|_F^2 .
\end{aligned}$$

where we use Cauchy-Schwarz inequality in the 2nd line, the inequality $(a-b)^2 \geq \frac{a^2}{2} - b^2$ in the 5th line, Lemma 3.10 and 3.11 in the 7th line, and the fact that $\left\| HH^\top \right\|_F \leq \|H\|_F^2$ in the 8th line. Since $\|H\|_F \leq \sqrt{\frac{3}{16}\sigma_r}$ and $\delta \leq \frac{1}{16}\sigma_r$, we have

$$\langle \nabla f(Z), H \rangle \geq \frac{27}{64}\sigma_r \|H\|_F^2 + \left\| H^\top \overline{Z} \right\|_F^2. \tag{3.11}$$

Proof of the Local Smoothness Condition

We need to upper bound $\|\nabla f(Z)\|_F^2 = \max_{\|W\|_F=1} |\langle \nabla f(Z), W \rangle|^2$. It suffices to show that for any $W \in \mathbb{R}^{n\times R}$ of unit Frobenius norm, $|\langle \nabla f(Z), W \rangle|^2$ is upper bounded if $Z \in E\left(\sqrt{\frac{3}{16}\sigma_r}\right)$.

Since $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, we have

$$\begin{aligned}
|\langle \nabla f(Z), W \rangle|^2 = & \left(\frac{1}{m}\sum_{i=1}^m \Big(\operatorname{tr}(H^\top A_i H) + 2\operatorname{tr}(H^\top A_i \overline{Z}) \Big) \Big(\operatorname{tr}(W^\top A_i H) + \operatorname{tr}(W^\top A_i \overline{Z}) \Big) \right)^2 \\
= & \left(\frac{1}{m}\sum_{i=1}^m \operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i H) + 2\operatorname{tr}(H^\top A_i \overline{Z})\operatorname{tr}(W^\top A_i H) \right. \\
& \qquad \left. + \operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i \overline{Z}) + 2\operatorname{tr}(H^\top A_i \overline{Z})\operatorname{tr}(W^\top A_i \overline{Z}) \right)^2 \\
\leq & \ 4\left(\frac{1}{m}\sum_{i=1}^m \operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i H) \right)^2 \\
& \quad + 4\left(\frac{2}{m}\sum_{i=1}^m \operatorname{tr}(H^\top A_i \overline{Z})\operatorname{tr}(W^\top A_i H) \right)^2 \\
& \quad + 4\left(\frac{1}{m}\sum_{i=1}^m \operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i \overline{Z}) \right)^2 \\
& \quad + 4\left(\frac{2}{m}\sum_{i=1}^m \operatorname{tr}(H^\top A_i \overline{Z})\operatorname{tr}(W^\top A_i \overline{Z}) \right)^2.
\end{aligned}$$

The first term in the righthand side can be upper bounded as

$$\begin{aligned}
4\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i H)\right)^2 &\le 4\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i H)^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(W^\top A_i H)^2\right)\\
&\le 4\left(2\|H\|_F^4+\delta\|H\|_F^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\|W\|_F^2\|A_iH\|_F^2\right)\\
&= 4\left(2\|H\|_F^4+\delta\|H\|_F^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\|A_iH\|_F^2\right)\\
&\le 4\left(2\|H\|_F^4+\delta\|H\|_F^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\|A_i\|^2\|H\|_F^2\right)\\
&\le 36n\|H\|_F^2\left(2\|H\|_F^4+\delta\|H\|_F^2\right),
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in the first and second line, Lemma 3.10 and

$$\left\|HH^\top\right\|_F \le \left\|H\right\|_F^2$$

in the third line, and Corollary 3.1 in the last line.

The other three terms are bounded similarly. For the second term, we have

$$\begin{aligned}
4\left(\frac{2}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i\overline{Z})\operatorname{tr}(W^\top A_i H)\right)^2 &\le 16\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i\overline{Z})^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(W^\top A_i H)^2\right)\\
&\le 36n\|H\|_F^2\left((4\sigma_1+2\delta)\|H\|_F^2+4\left\|H^\top\overline{Z}\right\|_F^2\right),
\end{aligned}$$

where we use Lemma 3.11 and 3.1. The third term is bounded as

$$\begin{aligned}
4\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i H)\operatorname{tr}(W^\top A_i\overline{Z})\right)^2 &\le 4\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(H^\top A_i H)^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{tr}(W^\top A_i\overline{Z})^2\right)\\
&\le 36n\left\|\overline{Z}\right\|_F^2\left(2\|H\|_F^4+\delta\|H\|_F^2\right),
\end{aligned}$$

and the fourth term is bounded as

$$
\begin{aligned}
4\left(\frac{2}{m}\sum_{i=1}^{m}\mathrm{tr}(H^\top A_i\overline{Z})\,\mathrm{tr}(W^\top A_i\overline{Z})\right)^2 &\le 16\left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{tr}(H^\top A_i\overline{Z})^2\right)\left(\frac{1}{m}\sum_{i=1}^{m}(W^\top A_i\overline{Z})^2\right)\\
&\le 36n\left\|\overline{Z}\right\|_F^2\left((4\sigma_1+2\delta)\left\|H\right\|_F^2+4\left\|H^\top\overline{Z}\right\|_F^2\right).
\end{aligned}
$$

Putting these inequalities together, we have

$$
\left\|\nabla f(Z)\right\|_F^2 \le 36n\left(\left\|\overline{Z}\right\|_F^2+\left\|H\right\|_F^2\right)\left(2\left\|H\right\|_F^4+(4\sigma_1+3\delta)\left\|H\right\|_F^2+4\left\|H^\top\overline{Z}\right\|_F^2\right).
$$

Hence,

$$
\frac{\left\|\nabla f(Z)\right\|_F^2}{144n\left(\left\|\overline{Z}\right\|_F^2+\left\|H\right\|_F^2\right)} \le \left(\sigma_1+\frac{1}{2}\left\|H\right\|_F^2+\frac{3}{4}\delta\right)\left\|H\right\|_F^2+\left\|H^\top\overline{Z}\right\|_F^2.
$$

Since $\left\|H\right\|_F \le \sqrt{\frac{3}{16}\sigma_r}$ and $\delta \le \frac{1}{16}\sigma_r$, we have

$$
\frac{\left\|\nabla f(Z)\right\|^2}{144n\left(\left\|\overline{Z}\right\|_F^2+(3/16)\sigma_r\right)} \le \left(\sigma_1+\frac{9}{64}\sigma_r\right)\left\|H\right\|_F^2+\left\|H^\top\overline{Z}\right\|_F^2.
$$

Proof of the Regularity Condition

Now we combine the curvature and the smoothness conditions. For any $\gamma \in \left(0, \frac{\sigma_1}{\sigma_r}\right)$, it holds that

$$
\gamma\frac{\sigma_r}{\sigma_1}\cdot\frac{\left\|\nabla f(Z)\right\|_F^2}{144n\left(\left\|\overline{Z}\right\|_F^2+(3/16)\sigma_r\right)} \le \gamma\frac{\sigma_r}{\sigma_1}\cdot\left(\sigma_1+\frac{9}{64}\sigma_r\right)\left\|H\right\|_F^2+\left\|H^\top\overline{Z}\right\|_F^2. \tag{3.12}
$$

Combining equation (3.11) and (3.12), we obtain

$$
\begin{aligned}
\langle\nabla f(Z),H\rangle &\ge \left(\frac{27}{64}-\gamma-\gamma\frac{\sigma_r}{\sigma_1}\frac{9}{64}\right)\sigma_r\left\|H\right\|_F^2+\gamma\frac{\sigma_r}{\sigma_1}\cdot\frac{\left\|\nabla f(Z)\right\|_F^2}{144n(\left\|\overline{Z}\right\|_F^2+(3/16)\sigma_r)}\\
&\ge \left(\frac{27}{64}-\frac{73}{64}\gamma\right)\sigma_r\left\|H\right\|_F^2+\gamma\frac{\sigma_r}{\sigma_1}\cdot\frac{\left\|\nabla f(Z)\right\|_F^2}{144n(\left\|\overline{Z}\right\|_F^2+(3/16)\sigma_r)}.
\end{aligned}
$$

If we take $\gamma = \frac{1}{3}$, then

$$\begin{aligned}\langle \nabla f(Z), H\rangle \geq & \frac{1}{24}\sigma_r \|H\|_F^2 + \frac{\sigma_r}{\sigma_1}\cdot \frac{\|\nabla f(Z)\|_F^2}{3\cdot 144n\left(\left\|\overline{Z}\right\|_F^2 + (3/16)\sigma_r\right)} \\ \geq & \frac{1}{24}\sigma_r \|H\|_F^2 + \frac{\sigma_r/\sigma_1}{513n\|Z^\star\|_F^2}\|\nabla f(Z)\|_F^2,\end{aligned}$$

where we use $\left\|\overline{Z}\right\|_F^2 = \|Z^\star\|_F^2 = \|X^\star\|_* \geq \sigma_r$. Thus we have

$$\langle \nabla f(Z), H\rangle \geq \frac{1}{\alpha}\sigma_r \|H\|_F^2 + \frac{1}{\beta\|Z^\star\|_F^2}\|\nabla f(Z)\|_F^2$$

for $\alpha \geq 24$ and $\beta \geq \frac{\sigma_1}{\sigma_r}\cdot 513n$.

3.7.5 Initialization

Proof of Lemma 3.4

By assumption, we have

$$\left\|\frac{1}{m}\sum_{i=1}^m (z_s^{\star\top} A_i z_s^\star)A_i - 2z_s^\star z_s^{\star\top}\right\| \leq \frac{\delta}{r}, \quad s\in[r].$$

Hence,

$$\begin{aligned}\|M - 2X^\star\| = & \left\|\frac{1}{m}\sum_{i=1}^m\sum_{s=1}^r (z_s^{\star\top} A_i z_s^\star)A_i - 2\sum_{s=1}^r z_s^\star z_s^{\star T}\right\| \\ \leq & \sum_{s=1}^r \left\|\frac{1}{m}\sum_{i=1}^m (z_s^{\star\top} A_i z_s^\star)A_i - 2z_s^\star z_s^{\star\top}\right\| \\ \leq & \delta.\end{aligned} \tag{3.13}$$

Let $\lambda_1' \geq \cdots \geq \lambda_n'$ be the eigenvalues of M. By Weyl's theorem, we have

$$|\lambda_s' - 2\sigma_s| \leq \delta, \quad s\in[n].$$

Since $\delta < \sigma_r$, it is easy to see $\lambda_1' \geq \cdots \geq \lambda_r' > \delta$ and $|\lambda_s'| \leq \delta, s = r+1,\ldots,n$. Hence, $\lambda_s = \lambda_s'$,

$s \in [r]$, and $Z^0 {Z^0}^\top$ is the best rank r approximation of $\frac{1}{2}M$. Therefore,

$$\begin{aligned}
\left\| Z^0 {Z^0}^\top - Z^\star {Z^\star}^\top \right\|_F &\leq \sqrt{2r} \left\| Z^0 {Z^0}^\top - Z^\star {Z^\star}^\top \right\| \\
&= \sqrt{2r} \left\| Z^0 {Z^0}^\top - \frac{1}{2}M + \frac{1}{2}M - Z^\star {Z^\star}^\top \right\| \\
&\leq \sqrt{2r} \left(\left\| Z^0 {Z^0}^\top - \frac{1}{2}M \right\| + \left\| \frac{1}{2}M - Z^\star {Z^\star}^\top \right\| \right) \\
&\leq \sqrt{2r}\delta,
\end{aligned}$$

where we used $\|A\|_F \leq \sqrt{\text{rank}(A)}\,\|A\|$ in first line, the fact $\left\| Z^0 {Z^0}^\top - \frac{1}{2}M \right\| = \frac{1}{2}|\lambda_{r+1}| \leq \frac{1}{2}\delta$ and inequality (3.13) in the last line.

Let $H = Z^0 - \overline{Z}^0$. We want to bound $d(Z^0, Z^\star)^2 = \|H\|_F^2$. According to the discussion in Lemma 3.12, $H^\top \overline{Z}^0$ is symmetric and ${Z^0}^\top \overline{Z}^0$ is positive semidefinite.

The following step closely follows [Tu et al., 2016]. It holds that

$$\begin{aligned}
\left\| Z^0 {Z^0}^\top - Z^\star {Z^\star}^\top \right\|_F^2 &= \left\| Z^0 {Z^0}^\top - \overline{Z}^0 {\overline{Z}^0}^\top \right\|_F^2 \\
&= \left\| H{\overline{Z}^0}^\top + \overline{Z}^0 H^\top + HH^\top \right\|_F^2 \\
&= \text{tr}\Big(\overline{Z}^0 H^\top H {\overline{Z}^0}^\top + H{\overline{Z}^0}^\top H {\overline{Z}^0}^\top + HH^\top \overline{Z}^0 {}^\top \\
&\quad + \overline{Z}^0 H^\top \overline{Z}^0 H^\top + H{\overline{Z}^0}^\top \overline{Z}^0 H + HH^\top \overline{Z}^0 H^\top \\
&\quad + \overline{Z}^0 H^\top HH^\top + H{\overline{Z}^0}^\top HH^\top + HH^\top HH^\top \Big) \\
&= \text{tr}\Big((H^\top H)^2 + 2(H^\top \overline{Z}^0)^2 + 2(H^\top H)({\overline{Z}^0}^\top \overline{Z}^0) \\
&\quad + 4(H^\top H)(H^\top \overline{Z}^0) \Big) \\
&= \text{tr}\Big(\left(H^\top H + \sqrt{2} H^\top \overline{Z}^0 \right)^2 + (4 - 2\sqrt{2})(H^\top H)(H^\top \overline{Z}^0) \\
&\quad + 2(H^\top H)({\overline{Z}^0}^\top \overline{Z}^0) \Big) \\
&\geq \text{tr}\Big((4 - 2\sqrt{2})(H^\top H)(H^\top \overline{Z}^0) + 2(H^\top H)(\overline{Z}^\top \overline{Z}) \Big) \\
&= \text{tr}\Big((4 - 2\sqrt{2})(H^\top H)({Z^0}^\top \overline{Z}^0) \Big) + \text{tr}\Big((2\sqrt{2} - 2)(H^\top H)(\overline{Z}^\top \overline{Z}) \Big),
\end{aligned}$$

where in the fourth line we used the property that the trace is invariant under cyclic permutations and $H^\top \overline{Z}^0 = \overline{Z}^{0\top} H$.

Since $Z^{0\top}\overline{Z}^0$ is positive semidefinite, $\text{tr}((H^\top H)(Z^{0\top}\overline{Z}^0))$ is nonnegative. Hence,

$$\begin{aligned}
\left\| Z^0 Z^{0\top} - Z^\star Z^{\star\top} \right\|_F^2 &\geq (2\sqrt{2}-2)\,\text{tr}\left((H^\top H)(\overline{Z}^\top \overline{Z})\right) \\
&= (2\sqrt{2}-2)\left\| H\overline{Z}^\top \right\|_F^2 \\
&\geq (2\sqrt{2}-2)\,\|H\|_F^2\,\sigma_r \\
&= (2\sqrt{2}-2)\sigma_r d(Z^0, Z^\star)^2.
\end{aligned}$$

If $\delta \leq \frac{\sigma_r}{4\sqrt{r}}$, then

$$d(Z^0, Z^\star)^2 \leq \frac{\left\| Z^0 Z^0 - Z^\star Z^{\star\top} \right\|_F^2}{(2\sqrt{2}-2)\sigma_r} \leq \frac{2r\delta^2}{(2\sqrt{2}-2)\sigma_r} \leq \frac{3}{16}\sigma_r.$$

3.7.6 Sample Complexity

In this subsection, we verify that our assumptions hold with high probability if $m \geq cn\log n$, where c is a constant that depends on δ, r, and κ. Our proof relies on the following concentration inequality.

Theorem 3.3. *(Matrix Bernstein Inequality [Tropp, 2015]) Let $S_1, \ldots, S_m$ be independent random matrices with dimension $n \times n$. Assume that $\mathbb{E}(S_i) = 0$ and $\|S_i\| \leq L$, for all $i \in [m]$. Let $\nu^2 = \max\left\{ \left\| \sum_{i=1}^m \mathbb{E}(S_i S_i^\top) \right\|, \left\| \sum_{i=1}^m \mathbb{E}(S_i^\top S_i) \right\| \right\}$. Then for all $\delta \geq 0$,*

$$\mathbb{P}\left(\left\| \frac{1}{m}\sum_{i=1}^m S_i \right\| \geq \delta \right) \leq 2n \exp\left(\frac{-m^2\delta^2}{\nu^2 + Lm\delta/3} \right).$$

We first give a technical lemma that we will use later.

Lemma 3.13. *Let $A = (a_{ij})$ be a random matrix drawn from GOE. Let $S = a_{11}A - 2e_1 e_1^\top$. There*

exist absolute constants C, ρ such that with probability at least $1 - Ce^{-\rho n}$, we have

$$\|S\| \leq 18n.$$

Proof. Let $\widetilde{A} = A - a_{11}e_1e_1^\top$. $S = a_{11}\widetilde{A} + (a_{11}^2 - 2)e_1e_1^\top$. Note that a_{11} and $\widetilde{A}$ are independent, hence $\|S\| \leq |a_{11}|\|\widetilde{A}\| + |a_{11}^2 - 2|$. Besides, since $a_{11} \sim \mathcal{N}(0, 2)$, we can see that $a_{11}^2/2$ is χ^2 distributed.

First we bound the operator norm of $\widetilde{A}$. We rewrite $\|\widetilde{A}\|$ as

$$\|\widetilde{A}\| = \max_{\|u\|=1} |u^\top \widetilde{A}u| = \max_{\|u\|=1} |u^\top Du - du_1^2| \leq \|D\| + |d|,$$

where $D = \widetilde{A} + de_1e_1^\top$, $d \sim \mathcal{N}(0, 2)$. As D is GOE distributed, by Lemma 3.9,

$$\mathbb{P}\left(\|D\| > 3\sqrt{n}\right) \leq C'e^{-\rho' n}, \tag{3.14}$$

where C' and ρ' are absolute constants.

Using the Gaussian tail inequality, we have

$$\mathbb{P}\left(|d| > 2\sqrt{n}\right) \leq 2e^{-n}. \tag{3.15}$$

Combining inequalities (3.14) and (3.15), we have

$$\mathbb{P}\left(\|\widetilde{A}\| > 5\sqrt{n}\right) \leq \mathbb{P}\left(\|D\| > 3\sqrt{n} \vee |d| > 2\sqrt{n}\right) \leq C'e^{-\rho' n} + 2e^{-n}, \tag{3.16}$$

where the last inequality follows from the union bound.

Next we bound the deviation of the χ^2 term. By the corollary of Lemma 1 in Laurent and Massart [2000], we have

$$\mathbb{P}(|a_{11}^2 - 2| > 4(\sqrt{n} + n)) \leq 2e^{-n}. \tag{3.17}$$

Since a_{11} is identically distributed as d, inequality (3.15) holds for a_{11} as well. Namely,

$$\mathbb{P}\left(|a_{11}| > 2\sqrt{n}\right) \leq 2e^{-n}.$$

Combining this with inequalities (3.17), (3.16), we have

$$\mathbb{P}\left(\|S\| \leq 14n + 4\sqrt{n}\right) \geq 1 - 6e^{-n} - C'e^{-\rho' n}.$$

Finally, the statement is obtained by choosing proper C, ρ, and using $\sqrt{n} \leq n$. □

Proof of Lemma 3.5

Proof. It is equivalent to show that for any unit vector u, with high probability,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (u^\top A_i u) A_i - 2uu^\top \right\| \leq \frac{\delta}{r\sigma_1}.$$

If P is an orthonormal matrix, then

$$\begin{aligned}
\left\| \frac{1}{m} \sum_{i=1}^{m} \left((Pu)^\top A_i (Pu) \right) A_i - 2(Pu)(Pu)^\top \right\| &= \left\| \frac{1}{m} \sum_{i=1}^{m} \left(u^\top (P^\top A_i P) u A_i \right) - 2Puu^\top P^\top \right\| \\
&= \left\| \frac{1}{m} \sum_{i=1}^{m} u^\top (P^\top A_i P) u P^\top A_i P - 2uu^\top \right\| \\
&= \left\| \frac{1}{m} \sum_{i=1}^{m} u^\top \widetilde{A}_i u \widetilde{A}_i - 2uu^\top \right\|,
\end{aligned}$$

where in the second line we use unitary invariance of the operator norm, and in the last line we denote $P^\top A_i P$ by $\widetilde{A}_i$. Since the GOE is invariant under orthogonal conjugation, $\widetilde{A}_i$ and A_i are identically distributed. Hence, it suffices to prove the claim when $u = e_1$, i.e.

$$\left\| \frac{1}{m} \sum_{i=1}^{m} a_{11}^{(i)} A_i - 2e_1 e_1^\top \right\| \leq \delta_0,$$

where $a_{11}^{(i)}$ is the $(1,1)$ entry of A_i and $\delta_0 = \frac{\delta}{r\sigma_1}$.

To show this, we apply Theorem 3.3, where $S_i = a_{11}^{(i)} A_i - 2e_1 e_1^\top$. This requires that the operator norm of S_i is bounded, for each i. We address this by noticing that with high probability $\|S_i\| \leq 18n$, $\forall i$. To be precise, by Lemma 3.13 there exist constants C, ρ, such that

$$\mathbb{P}\left(\|S_i\| > 18n\right) \leq Ce^{-\rho n}, \quad i = 1, \ldots, m.$$

Taking the union bound over all the S_is leads to

$$\mathbb{P}\left(\max_i \|S_i\| > 18n\right) \leq mCe^{-\rho n}. \tag{3.18}$$

Next, we calculate $\nu^2 = \left\|\sum_{i=1}^m \mathbb{E}(S_i^2)\right\| = m\left\|\mathbb{E}(S_1^2)\right\|$. Let $A = (a_{ij})$ denote A_1, S denote S_1. We have $\mathbb{E}(S^2) = \mathbb{E}({a_{11}}^2 A^2) - 4e_1 e_1^\top$, and

$$\begin{aligned}
\left(a_{11}^2 A^2\right)_{11} &= a_{11}^4 + \sum_{k=2}^n a_{11}^2 a_{1k}^2, \\
\left(a_{11}^2 A^2\right)_{ii} &= a_{11}^2 \left(a_{ii}^2 + \sum_{k\neq i}^n a_{ik}^2\right), \quad \forall i \neq 1, \\
\left(a_{11}^2 A^2\right)_{ij} &= a_{11}^2 \sum_{k=1}^n a_{ik} a_{jk}, \quad \forall i \neq j.
\end{aligned}$$

It is easy to see that $\mathbb{E}(a_{11}^2 A^2) = \text{diag}(2n{+}10, 2n{+}2, \ldots, 2n{+}2)$. Consequently, $\nu^2 = (2n{+}6)m$.

By Theorem 3.3, if $m \geq \frac{42}{\min(\delta_0^2, \delta_0)} \cdot n \log n$, then

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m} S_i\right\| \geq \delta_0\right) &\leq 2n \exp\left(\frac{-m\delta_0^2}{2n(1+3\delta_0)+6}\right) \\
&\leq 2n \exp\left(\frac{-m\delta_0^2}{2n(4+3\delta_0)}\right) \\
&\leq 2n \exp\left(\frac{-m\delta_0^2}{14n \cdot \max(1, \delta_0)}\right) \\
&\leq \frac{2}{n^2}.
\end{aligned} \tag{3.19}$$

Combining inequalities (3.18) and (3.19), we conclude that

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m} a_{11}^{(i)} A_i - 2e_1 e_1^\top\right\| \leq \delta_0\right) \geq 1 - mCe^{-\rho n} - \frac{2}{n^2}.$$

□

Proof of Lemma 3.6

The formulation of the second order partial derivatives and their expectations is given in Appendix 3.7.2.

It is easy to see that for any $\overline{Z} \in \mathcal{S}$, $\max_{s\in[r]} \|\bar{z}_r\| \leq \sqrt{\sigma_1}$. Thus it is sufficient to prove that for any two unitary vector u and y with high probability it holds that

$$\left\|\frac{1}{m}\sum_{i=1}^{m} 2A_i u y^\top A_i - 2u^\top y I - 2y u^\top\right\| \leq \frac{\delta}{r\sigma_1}.$$

We can decompose y as $y = \beta u + \beta_\perp u_\perp$ for a certain unit vector $u_\perp$ that is orthogonal to u, where $\beta^2 + \beta_\perp^2 = 1$. Let $\delta_0 = \dfrac{\delta}{2r\sigma_1}$. It suffices to prove the following two claims.

(i) For any unitary vector u, with high probability

$$\left\| \frac{1}{m} \sum_{i=1}^{m} 2A_i uu^\top A_i - 2I - 2uu^\top \right\| \leq \delta_0.$$

(ii) For any two orthogonal unit vectors u and $u_\perp$, with high probability

$$\left\| \frac{1}{m} \sum_{i=1}^{m} 2A_i uu_\perp^\top A_i - 2u_\perp u^\top \right\| \leq \delta_0.$$

Proof of (i)

If P is an orthonormal matrix, then

$$\begin{aligned} \left\| \frac{1}{m} \sum_{i=1}^{m} 2A_i Puu^\top PA_i - 2I - 2Puu^\top P^\top \right\| &= \left\| \frac{1}{m} \sum_{i=1}^{m} 2P^\top A_i Puu^\top P^\top A_i P - 2I - 2uu^\top \right\| \\ &= \left\| \frac{1}{m} \sum_{i=1}^{m} 2\widetilde{A}_i uu^\top \widetilde{A}_i - 2I - 2uu^\top \right\|, \end{aligned}$$

where $\widetilde{A}_i$ and A_i have the same distribution. Hence we only need to prove the case where $u = e_1$:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} 2v^{(i)} {v^{(i)}}^\top - 2I - 2e_1 e_1^\top \right\| \leq \delta_0,$$

where $v^{(i)} = A_i e_1$ is the first column of A_i.

Let $S_i = 2(v^{(i)} {v^{(i)}}^\top - I - e_1 e_1^\top)$. To apply Theorem 3.3, we need to show that with high probability $\|S_i\|$ is bounded for each i and calculate $\nu^2 = \left\| \sum_{i=1}^{n} \mathbb{E}(S_i^2) \right\| = m \left\| \mathbb{E}(S_1^2) \right\|$.

Let S, v, A denote S_1, $v^{(1)}$, and $A^{(1)}$ respectively. It is easy to see that

$$\|S\| \leq 2 \|v\|^2 + 4 = 2(w + a_{11}^2) + 4,$$

where $w = \sum_{k=2}^{n} a_{1k}^2$. As $a_{11} \sim \mathcal{N}(0, 2)$, $a_{1k} \sim \mathcal{N}(0, 1)$ for $k \neq 1$, we can see that $a_{11}^2/2$ and w

are χ^2 distributed with degrees of freedom 1 and $n-1$, respectively. Using the χ^2 tail bound, we have

$$\begin{aligned}
&\mathbb{P}\left(a_{11}^2/2 > 2(\sqrt{n}+n)+1\right) \leq e^{-n}, \\
&\mathbb{P}\left(w > 5n-1\right) \leq e^{-n}, \;\; k = 2, \ldots, n.
\end{aligned}$$

It follows from the union bound that

$$\mathbb{P}\left(\|S\| > 26n+6\right) \leq 2e^{-n},$$

and consequently

$$\mathbb{P}\left(\max_i \|S_i\| > 26n+6\right) \leq 2me^{-n}. \tag{3.20}$$

To calculate ν^2, we expand $\mathbb{E}(S^2)$ as

$$\begin{aligned}
\mathbb{E}(S^2) &= 4\mathbb{E}\left((vv^\top)^2\right) - 4(I + e_1e_1^\top)^2 \\
&= 4\mathbb{E}\left(\|v\|^2 vv^\top\right) - 4(I + 3e_1e_1^\top).
\end{aligned}$$

Some simple calculations show that

$$\begin{aligned}
\left(\|v\|^2 vv^\top\right)_{11} &= {v_1}^4 + \sum_{k=2}^n {v_k}^2{v_1}^2, \\
\left(\|v\|^2 vv^\top\right)_{jj} &= {v_1}^2{v_j}^2 + {v_j}^4 + \sum_{k\neq 1,j} {v_k}^2{v_j}^2, \;\; j = 2, \ldots, n, \\
\left(\|v\|^2 vv^\top\right)_{jl} &= \sum_{k=1}^n {v_k}^2 v_j v_l, \;\; j < l.
\end{aligned}$$

As $v_1 \sim \mathcal{N}(0,2)$, $v_j \sim \mathcal{N}(0,1)$ for $j \neq 1$,

$$\begin{aligned}
&\mathbb{E}\left(\|v\|^2 vv^\top\right)_{11} = 2n+10,\\
&\mathbb{E}\left(\|v\|^2 vv^\top\right)_{jj} = n+3, \quad j=2,\ldots,n,\\
&\mathbb{E}\left(\|v\|^2 vv^\top\right)_{jl} = 0, \quad j<l.
\end{aligned}$$

Hence, $\mathbb{E}(S^2) = \text{diag}(8n+24, 4n+8, \ldots, 4n+8)$ and thus $\nu^2 = m(8n+24)$.

If $m \geq (128/\min(\delta_0^2, \delta_0))n\log n$, then by applying Theorem 3.3 we can see

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m 2v^{(i)}{v^{(i)}}^\top - 2I - 2e_1e_1^\top\right\| > \delta_0\right) &\leq 2n\exp\left(\frac{-m\delta_0^2}{8n+24+(\frac{26}{3}n+2)\delta_0}\right)\\
&\leq 2n\exp\left(\frac{-m\delta_0^2}{(128/3)n\max(1,\delta_0)}\right) \qquad (3.21)\\
&\leq \frac{2}{n^2}.
\end{aligned}$$

Combining inequalities (3.21) and (3.20) leads to

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m 2v^{(i)}{v^{(i)}}^\top - 2I - 2e_1e_1^\top\right\| \leq \delta_0\right) \geq 1 - 2me^{-n} - \frac{2}{n^2}.$$

Proof of (ii)

We only need to prove the case where $u = e_1$ and $u_\perp = e_2$ due to the same reason above. That is,

$$\left\|\frac{1}{m}\sum_{i=1}^m 2v^{(i)}{q^{(i)}}^\top - 2e_2e_1^\top\right\| \leq \delta_0,$$

where $v^{(i)}$ and $q^{(i)}$ are the first and second columns of A_i.

As before, let $S_i = 2(v^{(i)}{q^{(i)}}^\top - e_2e_1^\top)$ and let S, v, q, A denote S_1, $v^{(1)}$, $q^{(1)}$ and $A^{(1)}$ respectively. From the proof of (i), we can see that with probability at least $1-4e^{-n}$ both $\|v\|$ and

$\|q\|$ are no larger than $\sqrt{13n+1}$. Since $\|S\| \leq 2\|v\|\|q\| + 2$, we have

$$\mathbb{P}\left(\max_i \|S_i\| \geq 26n + 4\right) \leq 4me^{-n}.$$

Next, we calculate $\nu^2 = m \max\left\{\left\|\mathbb{E}(SS^\top)\right\|, \left\|\mathbb{E}(S^\top S)\right\|\right\}$.

$$\mathbb{E}(SS^\top) = 4\mathbb{E}(\|q\|^2)\mathbb{E}(vv^\top) + 4e_2e_2^\top.$$

$$\mathbb{E}(S^\top S) = 4\mathbb{E}(\|v\|^2)\mathbb{E}(qq^\top) + 4e_1e_1^\top.$$

Some simple calculation shows that $\mathbb{E}(\|v\|^2) = \mathbb{E}(\|q\|^2) = n+1$, $\mathbb{E}(vv^\top) = I + e_1e_1^\top$ and $\mathbb{E}(qq^\top) = I + e_2e_2^\top$. Hence,

$$\mathbb{E}(SS^\top) = 4(n+1)I + 4(n+1)e_1e_1^\top + 4e_2e_2^\top,$$

$$\mathbb{E}(S^\top S) = 4(n+1)I + 4(n+1)e_2e_2^\top + 4e_1e_1^\top,$$

and $\nu^2 = 8(n+1)m$. If $m \geq \frac{78}{\min(\delta_0^2, \delta_0)} n \log n$, then by applying Theorem 3.3 we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m 2v^{(i)}{q^{(i)}}^\top - 2e_1e_2^\top\right\| > \delta_0\right) &\leq 2n\exp\left(\frac{-m\delta_0^2}{8n+8+(\frac{26n+4}{3})\delta_0}\right) \\ &\leq 2n\exp\left(\frac{-m\delta_0^2}{26n\max(1,\delta_0)}\right) \\ &\leq \frac{2}{n^2}.\end{aligned} \tag{3.22}$$

This means,

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m 2v^{(i)}{q^{(i)}}^\top - 2e_1e_2^\top\right\| \leq \delta_0\right) \geq 1 - 4me^{-n} - \frac{2}{n^2}.$$

3.7.7 ADMM for Nuclear Norm Minimization

We reformulate the nuclear norm minimizing problem as

$$\min_{X\in\mathbb{R}^{n\times n}} \quad \frac{1}{2\lambda}\left\|\mathcal{A}(X)-b\right\|^2+\left\|X\right\|_*, \tag{3.23}$$

where $\lambda>0$ is the regularization parameter. $\lambda\to 0$ will enforce the minimizer X^*_{nuc} satisfying the affine constraint $\mathcal{A}(X^*_{\text{nuc}})=b$.

We apply ADMM to the dual problem of (3.23):

$$\begin{aligned}\min_{\alpha\in\mathbb{R}^m, V\in\mathbb{R}^{n\times n}} \quad & \frac{\lambda}{2}\left\|\alpha\right\|^2-\alpha^\top b\\ \text{subject to} \quad & \left\|V\right\|\le 1\\ & \mathcal{A}^\top(\alpha)=V,\end{aligned} \tag{3.24}$$

where we introduce an auxiliary variable V to make this problem equality constrained.

The augmented Lagrangian of problem (3.24) can be written as

$$L_\eta(\alpha,X)=\frac{\lambda}{2}\left\|\alpha\right\|^2-\alpha^\top b+\mathbf{1}_{\|\cdot\|\le 1}(V)+\langle X,\mathcal{A}^\top(\alpha)-V\rangle+\frac{\eta}{2}\left\|\mathcal{A}^\top(\alpha)-V\right\|_F^2,$$

where X is the multiplier, η is the penalty parameter, and $\mathbf{1}_{\|\cdot\|\le 1}$ is the indicator function of the unit spectral norm ball i.e. $\mathbf{1}_{\|\cdot\|\le 1}(V)$ equals 0 if $\|V\|\le 1$ and $+\infty$ otherwise.

Let $\text{vec}(\cdot)$ denote the vectorization of a matrix, whose inverse mapping is denoted by $\text{mat}(\cdot)$. We can rewrite the transformations as $\mathcal{A}(X)=\boldsymbol{A}\text{vec}(X)$ and $\mathcal{A}^\top(\alpha)=\text{mat}(\boldsymbol{A}^\top\alpha)=\sum_{i=1}^m\alpha_i A_i$, where $\boldsymbol{A}$ is a $m\times n^2$ matrix whose ith row is $\text{vec}(A_i)^\top$.

The ADMM starts from initialization (α^0,V^0,X^0) and updates the three variables alternately.

The updates can be computed in close forms:

$$\begin{aligned}
\alpha^{k+1} &= (\lambda I + \eta \boldsymbol{A}\boldsymbol{A}^\top)^{-1}\bigg(b + \boldsymbol{A}\text{vec}\big(\eta V^k - X^k\big)\bigg),\\
V^{k+1} &= \text{proj}\bigg(\sum_{i=1}^m \alpha_i^{k+1} A_i + X^k/\eta\bigg),\\
X^{k+1} &= X^k + \eta\bigg(\sum_{i=1}^m \alpha_i^{k+1} A_i - V^{k+1}\bigg),
\end{aligned}$$

where $\text{proj}(\cdot)$ is the projection onto the unit spectral norm ball. Let $X = U\Sigma V^\top$ be the singular value decomposition of X,

$$\text{proj}(X) = U\min(\Sigma, 1)V^\top.$$

In fact, the update of V can be combined with other steps without being computed explicitly. One only has to iterate the following two steps:

$$\begin{aligned}
\alpha^{k+1} &= (\lambda I + \eta \boldsymbol{A}\boldsymbol{A}^\top)^{-1}\bigg(b + \boldsymbol{A}\text{vec}\big(\eta \sum_{i=1} \alpha_i^{k} A_i + X^{k-1} - 2X^k\big)\bigg),\\
X^{k+1} &= \text{prox}_\eta\bigg(\eta\sum_{i=1}^m \alpha_i^{k+1} A_i + X^k\bigg),
\end{aligned}$$

where $\text{prox}_\eta(\cdot)$ is the singular value soft-thresholding operator defined as

$$\text{prox}_\eta(X) = U\max(\Sigma - \eta, 0)V^\top.$$

The sequence of multipliers $\left\{X^k\right\}$ converges to the primal solution of (3.23). To speed up the update of α, the Cholesky decomposition of $\lambda I + \eta \boldsymbol{A}\boldsymbol{A}^\top$ is precomputed in our implementation.

CHAPTER 4

RECTANGULAR MATRIX COMPLETION

We have seen that the Burer-Monteiro technique is remarkably effective for a family of low rank random SDPs in the previous chapter. In this Chapter, we enlarge the collection of problems to which the factored approach can be successfully applied, by analyzing the convergence properties of gradient descent applied to the problem of rectangular matrix completion from incomplete measurements. The standard matrix completion problem asks for the recovery of a low rank matrix $X^\star \in \mathbb{R}^{n_1 \times n_2}$ given only a small fraction of observed entries. Let Ω be the set of m indices of the observed entries. Fixing a target rank $r \ll \min(n_1, n_2)$, the natural, but nonconvex objective is

$$\begin{aligned} \min_{X \in \mathbb{R}^{n_1 \times n_2}} \quad & \text{rank}(X) \\ \text{subject to} \quad & X_{ij} = X^\star_{ij},\ (i,j) \in \Omega. \end{aligned} \tag{4.1}$$

In order for this problem to be well-posed, it is important to understand when $X^\star$ is identifiable and, in particular, the unique minimizer of (4.1). Moreover, because the problem is in general NP-hard, it is essential to identify tractable families of instances, together with efficient algorithms having global convergence guarantees.

In the current work, we apply the factorization technique by "lifting" the matrix $X^\star$ to a positive semidefinite matrix $Y^\star \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ in higher dimension. Lifting is an established method that recasts vector or matrix estimation problems in terms of positive semidefinite matrices with special structure. It has been applied to sparse eigenvector approximation [d'Aspremont et al., 2004] and phase retrieval [Candès et al., 2015a], where the lifted matrix is of rank one. As explained in detail below, we can construct $Y^\star$ to be of the same rank as $X^\star$, thus obtaining a factorization $Y^\star = Z^\star Z^{\star\top}$ for some $Z^\star \in \mathbb{R}^{(n_1+n_2)\times r}$, and transforming the original matrix completion problem into the problem of recovering the semidefinite factor $Z^\star$. We formulate this as minimizing a nonconvex objective $f(Z)$, to which we apply a gradient descent scheme, using a particular spectral initialization. Our analysis of this algorithm establishes a lower bound on the

number of matrix measurements that are sufficient to guarantee identifiability of the true matrix and geometric convergence of the gradient descent algorithm, with explicit bounds on the rate.

In the following section we give a full description of our approach. Our theoretical results are presented in Section 4.2, with detailed proofs contained in Section 4.6. Our analysis subsumes the case where $X^\star$ is positive semidefinite. In Section 4.3 we briefly review related work. The experimental results are presented in Section 4.4, and we conclude with a brief discussion of future work in Section 4.5.

4.1 Semidefinite Lifting, Factorization, and Gradient Descent

For any $(n_1 + n_2) \times r$ matrix Z, we will use $Z_{(i)}$ to denote its ith row, and Z_U and Z_V to denote the top n_1 and bottom n_2 rows. The operator, Frobenius and ℓ_∞ norm of matrices are denoted by $\|\cdot\|$, $\|\cdot\|_F$ and $\|\cdot\|_\infty$, respectively. We define $\|Z\|_{2,\infty} = \max_i \left\|Z_{(i)}\right\|_2$ as the largest ℓ_2 norm of its rows, and similarly $\|Z\|_{\infty,2} = \max\left\{\|Z\|_{2,\infty}, \left\|Z^\top\right\|_{2,\infty}\right\}$. Let $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ be the operator where

$$\mathcal{P}_\Omega(X)_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

In this paper, we focus on completing an incoherent or "non-spiky" matrix $X^\star$. With $U^\star \Sigma^\star V^\star$ denoting the rank-r SVD of $X^\star$, we assume $X^\star$ is μ-*incoherent*, as defined below.

Definition 4.1. *The matrix $X^\star$ is μ-incoherent with respect to the canonical basis if its singular vectors satisfy*

$$\|U^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}, \quad \|V^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}, \tag{4.3}$$

where μ is a constant.[1]

Our main interest is the uniform model where m entries of $X^\star$ are observed uniformly at random, though we shall analyze a Bernoulli sampling model, where each entry of $X^\star$ is observed

1. Note that $\mu \ge 1$, since $r = \|U^\star\|_F^2 = \sum_{i \in [n_1]} \left\|U^\star_{(i)}\right\|_2^2 \le \mu r$.

with probability $p = m/n_1 n_2$. One can transfer the results back to the uniform model, as the probability of failure under the uniform model is at most twice that under the Bernoulli model; see [Candès and Recht, 2009; Candès and Tao, 2010].

Using the rank-r SVD of $X^\star$, we can lift $X^\star$ to

$$Y^\star = \begin{bmatrix} U^\star \Sigma^\star U^{\star\top} & X^\star \\ X^{\star\top} & V^\star \Sigma^\star V^{\star\top} \end{bmatrix} = Z^\star Z^{\star\top}, \quad \text{where } Z^\star = \begin{bmatrix} U^\star \\ V^\star \end{bmatrix} \Sigma^{\star\frac{1}{2}}. \tag{4.4}$$

The symmetric decomposition of $Y^\star$ is not unique; our goal is to find a matrix in the set

$$\mathcal{S} = \left\{ \widetilde{Z} \in \mathbb{R}^{(n_1+n_2)\times r} \mid \widetilde{Z} = Z^\star R \text{ for some } R \text{ with } RR^\top = R^\top R = I \right\}, \tag{4.5}$$

since for any $\widetilde{Z} \in \mathcal{S}$ we have $X^\star = \widetilde{Z}_U \widetilde{Z}_V^\top$. Let $\underline{\Omega}$ denote the corresponding observed entries of $Y^\star$, and consider minimization of the squared error

$$\min_Z \frac{1}{2p} \sum_{(i,j)\in\underline{\Omega}} (ZZ_{ij}^\top - Y_{ij}^\star)^2 = \min_Z \frac{1}{2p} \left\| \mathcal{P}_{\underline{\Omega}}(ZZ^\top - Y^\star) \right\|_F^2. \tag{4.6}$$

Note that $Y^\star$ is not the unique minimizer of (4.6), nor is it the only possible positive semidefinite lifting of $X^\star$. For example, let P be an $r \times r$ nonsingular matrix, and form the matrices

$$Z' = \begin{bmatrix} U^\star \Sigma^{\star\frac{1}{2}} P \\ V^\star \Sigma^{\star\frac{1}{2}} P^{-1} \end{bmatrix} \qquad Y' = \begin{bmatrix} U^\star \Sigma^{\star\frac{1}{2}} P^2 \Sigma^{\star\frac{1}{2}} U^{\star\top} & X^\star \\ X^{\star\top} & V^\star \Sigma^{\star\frac{1}{2}} P^{-2} \Sigma^{\star\frac{1}{2}} V^{\star\top} \end{bmatrix}. \tag{4.7}$$

Since $\underline{\Omega}$ does not contain any entry in the top-left or bottom-right block, Y' is also a minimizer of (4.6). Thus, the solution set of the lifted problem is much larger than the set $\mathcal{S}$ of actual interest. For the sake of simple analysis, we shall focus on exact recovery of $Y^\star$ only, and thus impose an additional regularizer to align the column spaces of Z_U and Z_V, as in [Tu et al., 2016]. The

regularized loss is

$$f(Z) = \frac{1}{2p}\left\|\mathcal{P}_{\underline{\Omega}}(ZZ^\top - Y^\star)\right\|_F^2 + \frac{\lambda}{4}\left\|Z^\top DZ\right\|_F^2, \quad \text{where } D = \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix}. \tag{4.8}$$

While this apparently introduces an extra tuning parameter, our analysis establishes linear convergence of the projected gradient descent algorithm when $\lambda = \frac{1}{2}$, and thus one may treat λ as a fixed number.

It is discussed in [Chen and Wainwright, 2015] that one needs to ensure the iterates stay incoherent. Let $\mathcal{C}$ be the set of incoherent matrices

$$\mathcal{C} = \left\{Z : \|Z\|_{2,\infty} \le \sqrt{\frac{2\mu r}{n_1 \wedge n_2}}\left\|Z^0\right\|\right\} \tag{4.9}$$

where we assume μ is known and Z^0 will be determined.

Our algorithm is simply gradient descent on $f(Z)$, with projection onto $\mathcal{C}$.

Let $M = p^{-1}\mathcal{P}_\Omega(UV^\top - X^\star)$. Then the gradient of f is given by

$$\nabla f(Z) = \begin{bmatrix} 0 & M \\ M^\top & 0 \end{bmatrix} Z + \lambda DZZ^\top DZ. \tag{4.10}$$

The projection $\mathcal{P}_{\mathcal{C}}$ to the feasible set $\mathcal{C}$ has closed form solution, given by row-wise clipping:

$$\mathcal{P}_{\mathcal{C}}(Z)_{(i)} = \begin{cases} Z_{(i)} & \text{if } \left\|Z_{(i)}\right\| \le \sqrt{\frac{2\mu r}{n_1 \wedge n_2}}\left\|Z^0\right\|, \\ \frac{Z_{(i)}}{\left\|Z_{(i)}\right\|} \cdot \sqrt{\frac{2\mu r}{n_1 \wedge n_2}}\left\|Z^0\right\| & \text{otherwise.} \end{cases} \tag{4.11}$$

Note that $X^0 \equiv p^{-1}\mathcal{P}_\Omega(X^\star)$ is an unbiased estimator of $X^\star$ under the Bernoulli model. To initialize, we thus construct Z^0 from the top rank-r factors of X^0. This leads to the following algorithm.

Remarks. (i) The step size η is normalized by $\left\|Z^0\right\|^2$. Our analysis will establish linear con-

Algorithm 2: Projected gradient descent for matrix completion

input: Ω, $\left\{X^\star_{ij} : (i,j) \in \Omega\right\}$, m, n_1, n_2, r, λ, η

initialization

$p = m/n_1 n_2$

$U^0 \Sigma^0 V^{0\top}$ = rank-r SVD of $p^{-1}\mathcal{P}_\Omega(X^\star)$

$Z^0 = [U^0 \Sigma^{0\frac{1}{2}}; V^0 \Sigma^{0\frac{1}{2}}]$

$Z^1 = \mathcal{P}_{\mathcal{C}}(Z^0)$

$k \leftarrow 1$

repeat

$M^k = p^{-1}\mathcal{P}_\Omega(Z^k_U Z^{k\top}_V - X^\star)$

$$\nabla f(Z^k) = \begin{bmatrix} 0 & M^k \\ M^{k\top} & 0 \end{bmatrix} Z^k + \lambda D Z^k Z^{k\top} D Z^k.$$

$$Z^{k+1} = \mathcal{P}_{\mathcal{C}}\left(Z^k - \frac{\eta}{\left\|Z^0\right\|^2}\nabla f(Z^k)\right)$$

$k \leftarrow k+1$

until *convergence*;

output: $\widehat{Z} = Z^k, \widehat{X} = Z^k_U Z^{k\top}_V$.

vergence when taking step sizes of the form $\eta/\sigma_1^\star$, where η is a sufficiently small constant. We replace $\sigma_1^\star$ by $\left\|Z^0\right\|^2$ in the actual algorithm since it is unknown in practice. (ii) The feasible set (4.9) depends on $\left\|Z^0\right\|$ as well. Under the above spectral initialization, our analysis shows that when $p \geq O(\mu\kappa^2 r^2 \log n/n_1 \wedge n_2)$, the term $\sqrt{\frac{2\mu r}{n_1 \wedge n_2}}\left\|Z^0\right\|$ is an upper bound of $\|Z^\star\|_{2,\infty}$ with high probability (see Corollary 4.1 below). This means $\mathcal{S}$ is a subset of $\mathcal{C}$. Note that this does not change the global optimality of $Z^\star$ and its equivalent elements, since $f(Z^\star) = 0$. In practice, we find that the iterates of our algorithm remain incoherent, so that one may drop the projection step. (iii) The column space regularizer (4.8) is needed in our analysis. We also found that when $\lambda = 0$, our algorithm typically converges to another PSD lifted matrix of $X^\star$, with minor difference from $Y^\star$ in the top-left and bottom-right blocks.

In the following section we state and sketch a proof of our main convergence result for this algorithm.

4.2 Main Result: Convergence Analysis

Theorem 4.1. *Suppose that $X^\star$ is of rank r, with condition number $\kappa = \sigma_1^\star/\sigma_r^\star$, and μ-incoherent as defined in Definition 4.1. Suppose further that we observe m entries of $X^\star$ chosen uniformly at random. Let $Y^\star = Z^\star Z^{\star\top}$ be the lifted matrix as in* (4.4) *and write $n = \max(n_1, n_2)$. Then there exist universal constants c_0, c_1, c_2, c_3 such that if*

$$m \geq c_0 \mu r^2 \kappa^2 \max(\mu, \log n) n, \tag{4.12}$$

then with probability at least $1 - c_1 n^{-c_2}$ the iterates of Algorithm 2 converge to $Z^\star$ geometrically, when using regularization parameter $\lambda = 1/2$, correctly specified input rank r, and constant step size $\eta/\sigma_1^\star$ with $\eta \leq c_3/\mu^2 r^2 \kappa$.

We shall analyze the Bernoulli sampling model, as justified in Section 4.1.

Similar to Chapter 3, let us define the distance to $Z^\star$ in terms of the solution set $\mathcal{S}$.

Definition 4.2. *Define the distance between Z and $Z^\star$ as*

$$d(Z, Z^\star) = \min_{\widetilde{Z} \in \mathcal{S}} \left\| Z - \widetilde{Z} \right\|_F = \min_{RR^\top = R^\top R = I} \left\| Z - Z^\star R \right\|_F .$$

The next theorem establishes the global convergence of Algorithm 2, assuming that the input rank is correctly specified. The proof sketch is given in the next subsection.

Theorem 4.2. *There exist universal constants c_0, c_1, c_2 such that if $p \geq \dfrac{c_0 \mu r^2 \kappa^2 \log n}{n_1 \wedge n_2}$, with probability at least $1 - c_1 n^{-c_2}$, the initialization $Z^1 \in \mathcal{C}$ satisfies*

$$d(Z^1, Z^\star) \leq \frac{1}{4}\sqrt{\sigma_r^\star}. \tag{4.13}$$

Moreover, there exist universal constants c_3, c_4, c_5, c_6 such that if $p \geq \dfrac{c_3 \max(\mu^2 r^2 \kappa^2, \mu r \log n)}{n_1 \wedge n_2}$, when using constant step size $\eta/\sigma_1^\star$ with $\eta \leq \dfrac{c_4}{\mu^2 r^2 \kappa}$ and initial value $Z^1 \in \mathcal{C}$ obeying (4.13)*, the*

kth step of Algorithm 2 with $\lambda = 1/2$ *satisfies*

$$d(Z^k, Z^\star) \le \frac{1}{4}\left(1 - \frac{99}{256}\cdot\frac{\eta}{\kappa}\right)^{k/2}\sqrt{\sigma_r^\star}$$

with probability at least $1 - c_5 n^{-c_6}$.

Remarks.

(i) After each update, the distance of our iterates to $Z^\star$ is reduced by at least a factor of $1 - O(1/\mu^2 r^2 \kappa^2)$.

(ii) Hence, the output $\widehat{Z}$ satisfies $d(\widehat{Z}, Z^\star) \le \varepsilon$ after at most $\left\lceil 2\log^{-1}\left(1/(1-\frac{99}{256}\cdot\frac{\eta}{\kappa})\right)\log\left(\sqrt{\sigma_r^\star}/4\varepsilon\right)\right\rceil$ iterations.

4.2.1 Proof Sketch

Our proof idea is of the same nature as the analysis in Candès et al. [2015b]; Zheng and Lafferty [2015]. We show two appealing properties when sufficient entries are observed. First, our spectral initialization produces a starting point within the $O(\sigma_r^\star)$ neighborhood of the solution set.

Lemma 4.1. *There exist universal constants* c, c_1, c_2*, such that if* $p \ge \dfrac{c\mu r^2\kappa^2 \log n}{n_1 \wedge n_2}$ *then with probability at least* $1 - c_1 n^{-c_2}$*,*

$$d(Z^1, Z^\star) \le d(Z^0, Z^\star) \le \frac{1}{4}\sqrt{\sigma_r^\star}.$$

To demonstrate this, we exploit the concentration around the mean of $p^{-1}\mathcal{P}_\Omega(X^\star)$. See Section for the proof. Using this lemma, we can immediately show that $Z^\star$ and all other elements of $\mathcal{S}$ are contained in the feasible set (4.9).

Corollary 4.1. *With probability at least* $1 - c_1 n^{-c_2}$*,* $\|Z^\star\|_{2,\infty} \le \sqrt{\frac{2\mu r}{n_1 \wedge n_2}}\left\|Z^0\right\|$.

The second crucial property is that $f(Z)$ is well-behaved within the $O(\sqrt{\sigma_r^\star})$ neighborhood, so that the iterates move closer to the optima in every iteration. The key step is to set up a *local regularity condition* [Candès et al., 2015b] similar to Nesterov's conditions [Nesterov, 2004].

Definition 4.3. *Let $\overline{Z} = \arg\min_{\widetilde{Z}\in\mathcal{S}} \left\|Z - \widetilde{Z}\right\|_F$ denote the matrix closest to Z in the solution set. We say that f satisfies the regularity condition $RC(\varepsilon, \alpha, \beta)$ if there exist constants α, β such that for any $Z \in \mathcal{C}$ satisfying $d(Z, Z^\star) \le \varepsilon$, we have*

$$\langle \nabla f(Z), Z - \overline{Z} \rangle \ge \frac{1}{\alpha}\sigma_r^\star \left\|Z - \overline{Z}\right\|_F^2 + \frac{1}{\beta\sigma_1^\star}\left\|\nabla f(Z)\right\|_F^2.$$

Using this condition, one can show the iterates converge linearly to the optima if we start close enough to $Z^\star$.

Lemma 4.2. *Consider the update $Z^{k+1} = \mathcal{P}_{\mathcal{C}}\left(Z^k - \frac{\mu}{\sigma_1^\star}\nabla f(Z^k)\right)$. If f satisfies $RC(\varepsilon, \alpha, \beta)$, $d(Z^k, Z^\star) \le \varepsilon$ and $0 < \mu \le \min(\alpha/2, 2/\beta)$, then*

$$d(Z^{k+1}, Z^\star) \le \sqrt{1 - \frac{2\mu}{\alpha\kappa}}\, d(Z^k, Z^\star).$$

The following lemma illustrates the local regularity of $f(Z)$. Nesterov's criterion is established upon strong convexity and strong smoothness of the objective. Here we show analogous *curvature* and *smoothness* conditions holds for $f(Z)$ locally – within the $O(\sqrt{\sigma_r^\star})$ neighborhood – with high probability. Interestingly, we found that to show the local curvature condition holds, it suffices to set $\lambda = \frac{1}{2}$. The proof can be found in Section 4.6.3, for which we have generalized some technical lemmas of [Chen and Wainwright, 2015].

Lemma 4.3. *Let the regularization constant be set to $\lambda = \frac{1}{2}$. There exists universal constant c, c_1, c_2, such that if $p \ge \frac{c\max(\mu^2 r^2\kappa^2, \mu r\log n)}{n_1 \wedge n_2}$, then f satisfies $RC(\frac{\sqrt{\sigma_r^\star}}{4}, 512/99, 13196\mu^2 r^2\kappa)$, with probability at least $1 - c_1 n^{-c_2}$.*

4.3 Related Work

Matrix completion is one instance of the general low rank linear inverse problem

$$\text{find } X \text{ of minimum rank such that } \mathcal{A}(X) = b, \tag{4.14}$$

where $\mathcal{A}$ is an affine transformation and $b = \mathcal{A}(X^\star)$ is the measurement of the ground truth $X^\star$. Considerable progress has been made towards algorithms for recovering $X^\star$ including both convex and nonconvex approaches. One of the most popular methods is nuclear norm minimization, a convenient convex relaxation of rank minimization. It was first proposed in [Fazel, 2002; Recht et al., 2010], and analyzed under a certain *restricted isometry property* (RIP). Subsequent work clarified the conditions for reconstruction, and studied recovery guarantees for both exact and approximately low rank matrices, with or without noise [Candès and Recht, 2009; Candès and Tao, 2010; Negahban and Wainwright, 2012; Chen, 2015]. One significant advantage for this approach is its near-optimal sample complexity. Under the same incoherence assumption as ours, Chen [2015] establishes the currently best-known lower bound of $O(\mu r n \log^2 n)$ samples. Using a closely related notion of incoherence, Negahban and Wainwright [2012] show that if $X^\star$ is "α-nonspiky" with $\frac{\|X^\star\|_\infty}{\|X^\star\|_F} \leq \frac{\alpha}{\sqrt{n_1 n_2}}$, then $O(\alpha^2 r n \log n)$ samples are sufficient for exact recovery. However, convexity and low sample complexity aside, in practice the power of nuclear norm relaxation is limited due to high computational cost. The popular algorithms for nuclear norm minimization are proximal methods that perform iterative singular value thresholding [Cai et al., 2010; Tomioka et al., 2010]. However, such algorithms don't scale to large instances because the per-iteration SVD is expensive.

Another popular convex surrogate for the rank function is the max-norm [Srebro et al., 2004; Foygel and Srebro, 2011], given by $\|X\| = \min_{X=UV^\top} \|U\|_{2,\infty} \|V\|_{2,\infty}$. For certain types of problems, the max-norm offers better generalization error bounds than the nuclear norm [Srebro and Shraibman, 2005]. But practically solving large scale problems that incorporate the max-norm is also non-trivial. In 2010, Lee et al. [2010] rephrased the max-norm constrained problem as an

SDP, and applied Burer-Monteiro factorization. Although this ends up with an $\ell_{2,\infty}$ constraint similar to ours (4.9), we emphasize that the constraint plays a different role in our setting. While [Srebro et al., 2004; Lee et al., 2010] use it to promote low rank solutions, our purpose is to enforce incoherent solutions; and experimental results suggest that one can drop it. Moreover, the convergence of projected gradient descent for this problem was not previously understood.

In a parallel line of work, the problem of developing techniques that exactly solve nonconvex formulations has attracted significant recent research attention. In chronological order, Keshavan et al. [2010] proposed a manifold gradient method for matrix completion. They factorize $X^\star = U^\star \Sigma^\star V^{\star\top}$, where $U^\star \in \mathbb{R}^{n_1 \times r}, U^\top U = n_1 I_{n_1}$ and $V^\star \in \mathbb{R}^{n_2 \times r}, V^\top V = n_2 I_{n_2}$. Similar to our definition of $\mathcal{S}$, the equivalence classes of $U^\star$ and $V^\star$ are Grassmann manifolds of r dimensional subspaces. The authors then minimize the nonconvex objective $F(U,V) = \min_{S \in \mathbb{R}^{r\times r}} \left\| \mathcal{P}_\Omega(USV^\top - X^\star) \right\|_F^2$ over the manifolds. In each iteration, U and V are updated along their manifold gradients, followed by the update of the optimal scaling matrix S. This algorithm can exactly exactly reconstruct the matrix, though the convergence rate is unknown. However, its per-iteration update also has high computational complexity, see Section 4.4 for details. There are other manifold optimization methods for matrix completion including [Boumal and Absil, 2011; Mishra et al., 2013; Vandereycken, 2013].

In the same year, Jain et al. [2010] suggested minimizing the squared residual $\|\mathcal{A}(X) - b\|^2$ under a rank constraint $\text{rank}(X) \leq r$. While this constraint is nonconvex, projection onto the feasible set can be computed using low rank SVD. Under certain RIP assumption on $\mathcal{A}$, Jain et al. establish the global convergence of projected gradient descent for (4.14). This algorithm is named Singular Value Projection (`SVP`). Yet in the setting of completion, only experimental support for the effectiveness of `SVP` is provided. More importantly, `SVP` also suffers from expensive per-iteration SVD for large scale problems.

Keshavan [2012]; Jain et al. [2013] further analysed the alternating minimization procedure for (4.14). `AltMin` factorizes $X = UV^\top$ where $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$, and alternately solves $\left\| \mathcal{A}(UV^\top) - b \right\|_2^2$ over U and V, while fixing the other factor. The authors obtain sample

complexity bounds with $r\kappa^8$, $r^7\kappa^6$ dependency, respectively. In 2014, Hardt [2014] improved the bounds to $r^2\kappa^2$. Notably, all these works assume the use of *resampling*—independent sequences of samples $\Omega_k, k = 1, 2, \ldots$. In other words, in every iteration we can sample the true matrix under a certain Bernoulli model independently. However, in practice Ω is usually given and fixed. To get around the dependence on the sample sets, they partition Ω into a predefined number of subsets of equal size. However, sample sets obtained by partitioning are not independent, and partitioning, if used in practice, does not make the most efficient use of the data. Thus, Hardt and Wootters [2014] considered a new resampling scheme. They assume a known generative model of $\{\Omega_k\}$, where each Ω_k is obtained under a Bernoulli model with probability p_k, $p = \sum_k p_k$ and $\Omega = \cup_k \Omega_k$. While not practical, under this assumption the authors obtain a sample complexity that is logarithmic in κ.

Another theoretical disadvantage of the resampling scheme is that the sample complexity depends on the desired accuracy ε, as established by [Keshavan, 2012; Jain et al., 2013; Hardt, 2014; Hardt and Wootters, 2014]. As the accuracy goes to zero, the sample complexity increases. In contrast, our algorithm doesn't require resampling, and the sample complexity is independent of ε.

In 2014, Candès et al. [2015b] proposed *Wirtinger flow* for phase retrieval. Wirtinger flow is a fast first-order algorithm that minimizes a fourth order (nonconvex) objective, geometrically converging to the global optimum. While previous work [Candès et al., 2015a, 2013; Candès and Li, 2014] lifts the phase retrieval problem into an SDP where the solution is rank one, this work bridges SDP and first-order algorithms via the Burer-Monteiro technique. It has inspired further research on related topics; last year, the authors of [Zheng and Lafferty, 2015; Tu et al., 2016; Bhojanapalli et al., 2016a; Chen and Wainwright, 2015] considered factorizations for (4.14), assuming $X^\star$ is semidefinite, and proved global optimality of first-order algorithms under appropriate initializations. Tu et al. [2016] have extended this algorithm to handle rectangular matrix via asymmetric factorization, and have shown exact recovery of $X^\star$, assuming $\mathcal{A}$ satisfies a certain RIP. They use lifting implicitly, factorizing $X = Z_U Z_V^\top$ and applying gradient updates on both factors Z_U and

Z_V simultaneously, with the nonconvex objective function

$$g(Z_U, Z_V) = \frac{1}{2p} \left\| \mathcal{P}_\Omega(Z_U Z_V^\top - X^\star) \right\|_F^2 + \frac{\lambda'}{4} \left\| Z_U^\top Z_U - Z_V^\top Z_V \right\|_F^2 . \tag{4.15}$$

Their proof strategy also shows convergence of Z in the lifted space. For the specific case of matrix completion, Chen and Wainwright [2015] obtained guarantees when $X^\star$ is semidefinite. Our work generalizes the results obtained in [Tu et al., 2016; Chen and Wainwright, 2015], extending the recent literature on first-order algorithms for factorized models.

After completing this work we learned of independent research of Sun and Luo [2015], who also analysed a gradient algorithm for rectangular matrix completion. Their formulation is similar to ours, with additional Frobenius norm constraints on the factors. The authors established a sample complexity of $O(r^7\kappa^6)$ observations; in comparison our bound scales as $O(r^2\kappa^2)$. The authors also analyzed block coordinate descent type alternating minimization, which cyclically updates the rows of U and then the rows of V, showing exact recovery of this algorithm without resampling. Recent independent work of Yi et al. [2016] analyzes a gradient scheme for Robust PCA. Under the setting of partial observation without corruption, this is the standard matrix completion problem. In other related work, [Zhao et al., 2015; Wei et al., 2016] also study nonconvex optimization methods for matrix completion, using algorithms that still require low rank SVD in each iteration.

4.4 Experiments

We conduct experiments on synthetic datasets to support our analytical results. As the column space regularizer and incoherence constraint of our gradient method (`GD`) are merely for analytical purpose, we drop them in all the experiments; simply optimize the ℓ_2 loss $\frac{1}{2} \left\| \mathcal{P}_{\underline{\Omega}}(ZZ^\top - Y^\star) \right\|_F^2$. We compare `GD` with `SVP`, `OptSpace`, nuclear norm minimization (`nuclear`) and trust region methods on Riemannian manifolds (`trustRegion`). For `nuclear`, we rescale the standard objective to be

$$\min_X \quad \frac{1}{2\lambda} \|\mathcal{P}_\Omega(X - X^\star)\|_F^2 + \|X\|_* , \tag{4.16}$$

Method	Complexity
GD	$2mr + m + n^2r + 4nr$
SVP	$O(n^2r)$
OptSpace	$O(mr^3 + n^2r^2 + nr^4 + r^6)$
nuclear	$O(n^3)$
AltMin	$O(mr^2)$

Table 4.1: Matrix completion: per-iteration computational complexities of different methods.

where $\lambda = 0$ will enforce the minimizer fitting the observed values exactly. We use ADMM to solve (4.16). It is based on the algorithm for *the matrix approach* in [Tomioka et al., 2010], and can neatly handle the case $\lambda = 0$. We emphasize there is no computational difference between cases whether λ is zero or not. All methods are implemented in MATLAB. We use the toolbox `Manopt` for `trustRegion` [Boumal et al., 2014] and the implementation of `OptSpace` from the authors. For `AltMin`, we use the same sample sets in every iteration. The experiments were run on a Linux machine with a 3.4GHz Intel Core i7 processor and 8 GB memory.

4.4.1 Computational Complexity

Table 4.1 summarizes the per-iteration complexity of all the methods for completing a $n \times n$ matrix. Since M^k is a sparse matrix with m nonzero entries, and we have dropped the regularizer and constraint, our method `GD` only needs $2mr + m + n^2r$ operations to compute the gradient, and $4nr$ operations to update the iterate. The computation of `nuclear` is dominated by singular value thresholding and updating the objective value, which require the $O(n^3)$ cost full SVD. Similarly, `SVP` needs $O(n^2r)$ operations to compute the rank-r SVD for low rank projection. For `OptSpace`, $O(mr + n^2r + nr^2)$ operations are needed to compute the manifold gradient and line search. The most expensive part is to determine the optimal scaling matrix $S \in \mathbb{R}^{r\times r}$, which boils down to solving a r^2 by r^2 dense linear system. In total $O(mr^3 + n^2r^2 + nr^4 + r^6)$ operations are used to construct and solve this system. For `AltMin`, in every iteration we have to solve $(n_1 + n_2)$ linear systems of size $r \times r$. See [Sun, 2015] for the exact formulation. The time cost for each iteration is $O(mr^2)$. One can see that `GD` reduces the computation than the others. Though the

dominating terms for SVP and GD are in the same order, in practice the partial SVD are more expensive than the gradient update, especially on large instances.

4.4.2 Runtime Comparison

We randomly generated a true matrix $X^\star$ of size 4000×2000 and rank 3. It is constructed from the rank-3 SVD of a random 4000×2000 matrix with i.i.d normal entries. We sampled $m = 199057$ entries of $X^\star$ uniformly at random, where m is roughly equal to $2nr \log n$ with $n = 4000$ and $r = 3$. For simplicity, we feed SVP, OptSpace and GD with the true rank. For all these methods, we use the randomized algorithm of Halko et al. [2011] to compute the low rank SVD, which is approximately 15 times faster than MATLAB built-in SVD on instances of such size. We report relative error measured in the Frobenius norm, defined as $\|\widehat{X} - X^\star\|_F / \|X^\star\|_F$. For nuclear, we set $\lambda = 0$ to enforce exact fitting. The convergence speed of ADMM mildly depends on the choice of penalty parameter. We tested 5 values $0.1, 0.2, 0.5, 1, 1.5$ and selected 0.2, which leads to fastest convergence. Similarly, for SVP, we would like to choose the largest step size for which the algorithm is converging. We evaluated $15, 20, 30, 35, 40$ and selected 30. The step size is chosen for GD in the same way. Five values $20, 50, 70, 75, 80$ are tested for η and we picked 70. For OptSpace, we compared fixed step sizes $0.50.10.050.010.005$ with line search, and found the algorithm converged fastest under line search. Figure 4.1a shows the results. GD is slightly slower than trustRegion and faster than competing approaches.

To further illustrate how runtime scales as the dimension increases, we run larger instances of size 10000×5000 and 20000×5000, where the true rank is 40. The parameters are selected in the same manner, and we terminate the computation once the relative error is below $1e^{-9}$. We report the results of AltMin GD, SVP and trustRegion in Figure 4.2a; nuclear, OptSpace do not scale well to such sizes so that we didn't include them. The runtime of AltMin scales the slowest, while the runtimes of GD and trustRegion increase slower than SVP.

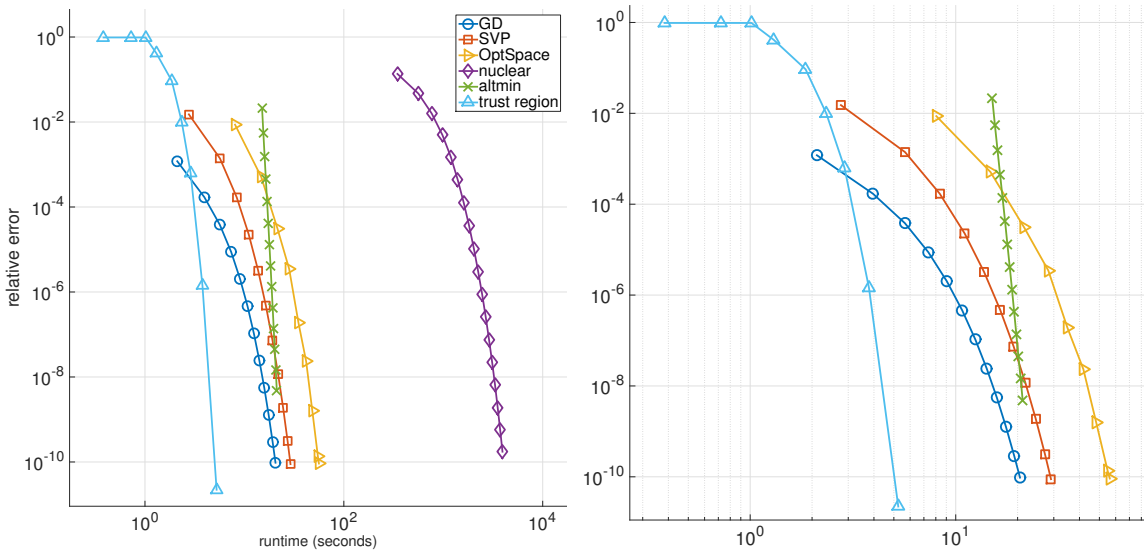


Figure 4.1: (a) Runtime comparison where $X^\star$ is 4000×2000 and of rank 3. 199057 entries are observed. (b) Magnified plots to compare other methods except `nuclear`.

4.4.3 Sample Complexity

We evaluate the number of observations required by `GD` for exact recovery. For simplicity, we consider square but asymmetric $X^\star$. We conducted experiments in 4 cases, where the randomly generated $X^\star$ is of size 500×500 or 1000×1000, and of rank 10 or 20. In each case, we compute the solutions of `GD` given m random observations, and a solution with relative error below $1e^{-6}$ is considered to be successful. We run 20 trials and compute the empirical probability of successful recovery. The results are shown in Figure 4.2b. For all four cases, the phase transitions occur around $m \approx 3.5nr$. This suggests that the actual sample complexity of `GD` may scale linearly with both the dimension n and the rank r.

4.5 Discussion

We propose a lifting procedure together with Burer-Monteiro factorization and a first-order algorithm to carry out rectangular matrix completion. While optimizing a nonconvex objective, we establish linear convergence of our method to the global optimum with $O(\mu r^2 \kappa^2 n \max(\mu, \log n))$

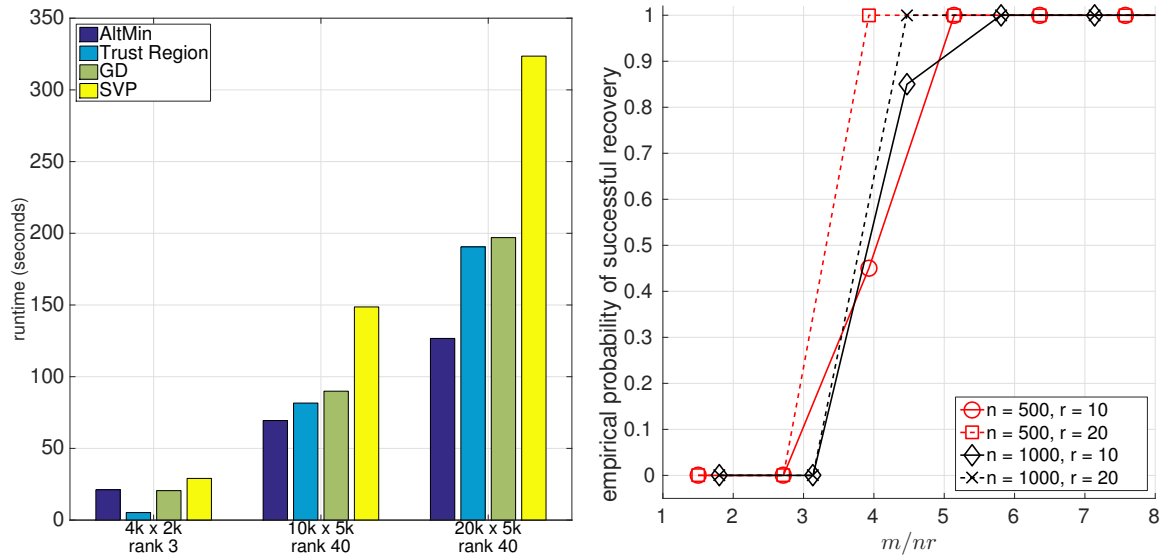


Figure 4.2: (a) Runtime growth of `AltMin`, `trustRegresion`, `GD` and `SVP`. (b) Sample complexity of gradient scheme.

random observations. We conjecture that $O(nr)$ observations are sufficient for exact recovery, and that the column space regularizer can be dropped. We provide empirical evidence showing this simple algorithm is fast and scalable, suggesting that lifting techniques may be promising for much more general classes of problems.

4.6 Proofs

4.6.1 Technical Lemmas

Another way of writing the objective function is

$$f(Z) = \frac{1}{2p}\sum_{l=1}^{2m}\left(\langle A_l, ZZ^\top\rangle - b_l\right)^2 + \frac{\lambda}{4}\left\|Z^\top DZ\right\|_F^2,$$

where l is an index of $\underline{\Omega}$, A_l is a matrix with 1 at the corresponding observed entry and 0 elsewhere. Let $H = Z - \overline{Z}$, the gradient can be written as

$$\begin{aligned}\nabla f(Z) &= \frac{1}{p}\sum_{l=1}^{2m}\left(\langle A_l, ZZ^\top\rangle - b_l\right)(A_l + A_l^\top)Z + \lambda \overbrace{DZ\left(Z^\top DZ\right)}^{\Gamma} \\ &= \frac{1}{p}\sum_{l=1}^{2m}\left(\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top + HH^\top\rangle\right)(A_l + A_l^\top)(\overline{Z} + H) + \lambda\Gamma.\end{aligned}$$

We will use the following facts throughout the proof:

$$\left\|\overline{Z}\right\|_{2,\infty} = \|Z^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1 \wedge n_2}\sigma_1^\star}, \tag{4.17}$$

$$\|H\|_{2,\infty} \le 3\sqrt{\frac{\mu r}{n_1 \wedge n_2}\sigma_1^\star}, \tag{4.18}$$

$$\langle (A_l + A_l^\top)B, C\rangle = \langle A_l, BC^\top + CB^\top\rangle, \tag{4.19}$$

$$Z^\top\overline{Z} \text{ is positive semidefinite, } H^\top\overline{Z} \text{ is symmetric.} \tag{4.20}$$

Inequality (4.17) is a direct result of Definition 4.1. To see (4.18), note that $\|H\|_{2,\infty} \le \|Z\|_{2,\infty} + \left\|\overline{Z}\right\|_{2,\infty} \le \sqrt{\frac{2\mu r}{n_1\wedge n_2}\sigma_1} + \sqrt{\frac{\mu r}{n_1\wedge n_2}\sigma_1^\star}$, and $|\sigma_1 - \sigma_1^\star| \le \frac{1}{16}\sigma_1^\star$ by the discussion of initialization in Section 4.6.2. For (4.20), it holds that

$$\operatorname*{arg\,min}_{RR^\top = R^\top R = I} \|Z - Z^\star R\|_F^2 = AB^\top,$$

where $A\Lambda B^\top$ is the SVD of ${Z^\star}^\top Z$. Clearly, $Z^\top\overline{Z}$ is positive semidefinite, and $H^\top\overline{Z} = Z^\top\overline{Z} - \overline{Z}^\top\overline{Z} = B\Lambda B^\top - \overline{Z}^\top\overline{Z}$ is symmetric.

Next, we list several technical lemmas that are utilized later. We will use c to denote a numerical constant, whose value may vary from line to line.

Lemma 4.4. *For any Z of the form $Z = \begin{bmatrix} Z_U \\ Z_V \end{bmatrix} = \begin{bmatrix} U\Sigma^{\frac{1}{2}}R \\ V\Sigma^{\frac{1}{2}}R \end{bmatrix}$, where U, V, R are unitary matrices*

and $\Sigma \succeq 0$ is a diagonal matrix, we have

$$\left\|ZZ^\top - Z^\star Z^{\star\top}\right\|_F \leq 2\left\|U\Sigma V^\top - U^\star\Sigma^\star V^{\star\top}\right\|_F.$$

Proof. Recall that

$$Z^\star = \begin{bmatrix} Z_U^\star \\ Z_V^\star \end{bmatrix} = \begin{bmatrix} U^\star\Sigma^{\star\frac{1}{2}} \\ V^\star\Sigma^{\star\frac{1}{2}} \end{bmatrix}$$

where $X^\star = U^\star\Sigma^\star V^{\star\top}$. We have

$$\begin{aligned} &\left\|ZZ^\top - Z^\star Z^{\star\top}\right\|_F^2 \\ &= \left\|U\Sigma U^\top - U^\star\Sigma^\star U^{\star\top}\right\|_F^2 + \left\|V\Sigma V^\top - V^\star\Sigma^\star V^{\star\top}\right\|_F^2 + 2\left\|U\Sigma V^\top - U^\star\Sigma^\star V^{\star\top}\right\|_F^2, \end{aligned} \tag{4.21}$$

and

$$\begin{aligned} &\left\|U\Sigma U^\top - U^\star\Sigma^\star U^{\star\top}\right\|_F^2 + \left\|V\Sigma V^\top - V^\star\Sigma^\star V^{\star\top}\right\|_F^2 \\ &= 2\left(\|\Sigma\|_F^2 + \|\Sigma^\star\|_F^2 - \langle\Sigma, U^\top U^{\star\top}\Sigma^\star U^{\star\top}U + V^\top V^{\star\top}\Sigma^\star V^{\star\top}V\rangle\right). \end{aligned} \tag{4.22}$$

We can obtain the lower bound

$$\begin{aligned} &\langle\Sigma,\, U^\top U^{\star\top}\Sigma^\star U^{\star\top}U + V^\top V^{\star\top}\Sigma^\star V^{\star\top}V\rangle \\ &\quad= \sum_{i=1}^r \sigma_i\left(U^\top U^{\star\top}\Sigma^\star U^{\star\top}U + V^\top V^{\star\top}\Sigma^\star V^{\star\top}V\right)_{ii} \\ &\quad= \sum_{i=1}^r \sigma_i \sum_{k=1}^r \sigma_k^\star\left((U^\top U^\star)_{ik}^2 + (V^\top V^\star)_{ik}^2\right) \\ &\quad\geq \sum_{i=1}^r \sigma_i \sum_{k=1}^r \sigma_k^\star \cdot 2\,(U^\top U^\star)_{ik}(V^\top V^\star)_{ik} \\ &\quad= 2\langle\Sigma,\, U^\top U^\star\Sigma^\star V^{\star\top}V\rangle. \end{aligned} \tag{4.23}$$

Combining (4.22) and (4.23), we obtain

$$\begin{aligned}
&\left\|U\Sigma U^\top - U^\star\Sigma^\star U^{\star\top}\right\|_F^2 + \left\|V\Sigma V^\top - V^\star\Sigma^\star V^{\star\top}\right\|_F^2 \\
&\quad\le\ 2\left(\|\Sigma\|_F^2 + \|\Sigma^\star\|_F^2 - 2\langle\Sigma, U^\top U^{\star\top}\Sigma^\star V^{\star\top}V\rangle\right) \\
&\quad=\ 2\left(\left\|U\Sigma V^\top\right\|_F^2 + \left\|U^\star\Sigma^\star V^{\star\top}\right\|_F^2 - 2\langle U\Sigma V^\top, U^{\star\top}\Sigma^\star V^{\star\top}\rangle\right) \\
&\quad=\ 2\left\|U\Sigma V^\top - U^\star\Sigma^\star V^{\star\top}\right\|_F^2.
\end{aligned} \tag{4.24}$$

Plugging (4.24) back into (4.21), we obtain the lemma. □

Recall that $n = \max(n_1, n_2)$. We will exploit the following two known concentration results.

Lemma 4.5 (Chen [2015], Lemma 2)**.** *For any fixed matrix $X^\star \in \mathbb{R}^{n_1\times n_2}$, there exist universal constants c, c_1, c_2 such that with probability at least $1 - c_1 n^{-c_2}$,*

$$\left\|p^{-1}\mathcal{P}_\Omega(X^\star) - X^\star\right\| \le c\left(\frac{\log n}{p}\|X^\star\|_\infty + \sqrt{\frac{\log n}{p}}\|X^\star\|_{\infty,2}\right).$$

Lemma 4.6 (Candès and Recht [2009], Theorem 4.1)**.** *Define subspace*

$$T = \left\{M \in \mathbb{R}^{n_1\times n_2} : M = U^\star X^\top + YV^{\star\top} \text{ for some } X \text{ and } Y\right\}. \tag{4.25}$$

Let $\mathcal{P}_T$ be the Euclidean projection onto T. There is a numerical constant c such that for any $\delta \in (0, 1]$, if $p \ge \frac{c}{\delta^2}\frac{\mu r\log n}{n_1 \wedge n_2}$, then with probability $1 - 3n^{-3}$, we have

$$p^{-1}\|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - p\mathcal{P}_T\| \le \delta.$$

Lemma 4.7 upper bounds the spectral norm of the adjacency matrix of a random Erdős-Rényi graph. It is a variant of Lemma 7.1 of Keshavan et al. [2010], which uses known results of Feige and Ofek [2005].

Lemma 4.7 (Chen and Wainwright [2015], Lemma 9)**.** *Suppose that $\bar{\Omega} \subset [d] \times [d]$ is the set*

of edges of a random Erdős-Rényi graph with n nodes, where any pair of nodes is connected with probability p. There exists two numerical constants c_1, c_2 such that, for any $\delta \in (0,1]$, if $p \geq \frac{c_1 \log d}{\delta^2 d}$, then with probability at least $1 - \frac{1}{2}d^{-4}$, uniformly for all $x, y \in \mathbb{R}^n$ it holds that

$$p^{-1} \sum_{(i,j)\in\bar{\Omega}} x_i y_j \leq (1+\delta) \|x\|_1 \|y\|_1 + c_2 \sqrt{\frac{d}{p}} \|x\|_2 \|y\|_2 . \tag{4.26}$$

We refer readers to [Keshavan et al., 2010] for a complete proof, in particular noticing that one can choose p large enough so that the constant factor in the first term in (4.26) is only $1+\delta$.

Lemma 4.8, 4.9 and 4.10 are direct generalizations of Lemma 4 and 5 of [Chen and Wainwright, 2015].

Lemma 4.8. *There exists a constant c such that, for any $\delta \in (0,1]$, if*

$$p \geq \frac{c}{\delta^2} \max\big(\frac{\log(n_1+n_2)}{n_1+n_2}, \frac{\mu^2 r^2 \kappa^2}{n_1 \wedge n_2}\big),$$

then with probability at least $1 - \frac{1}{2}(n_1+n_2)^{-4}$, uniformly for all H such that $\|H\|_{2,\infty} \leq 3\sqrt{\frac{\mu r}{n_1 \wedge n_2}\sigma_1^\star}$, we have

$$p^{-1} \left\| \mathcal{P}_{\underline{\Omega}}(HH^\top) \right\|_F^2 \leq (1+\delta) \|H\|_F^4 + \delta \sigma_r^\star \|H\|_F^2 .$$

Proof. It holds that

$$\begin{aligned} p^{-1} \left\| \mathcal{P}_{\underline{\Omega}}(HH^\top) \right\|_F^2 &= p^{-1} \sum_{(i,j)\in\underline{\Omega}} \langle H_{(i)}, H_{(j)} \rangle^2 \\ &\leq p^{-1} \sum_{(i,j)\in\underline{\Omega}} \left\| H_{(i)} \right\|_2^2 \left\| H_{(j)} \right\|_2^2 . \end{aligned} \tag{4.27}$$

Since $\underline{\Omega}$ is a reduced sampling of $Y \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ under a Bernoulli model, Lemma 4.7 is applicable here. Assume $p \geq \frac{c_1 \log(n_1+n_2)}{\delta^2(n_1+n_2)}$, we then have with probability at least $1 - \frac{1}{2}(n_1 +$

$n_2)^{-4}$, for all H such that $\|H\|_{2,\infty} \le 3\sqrt{\frac{\mu r}{n_1 \wedge n_2}\sigma_1^\star}$,

$$\begin{aligned} p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2 \le & \quad p^{-1}\sum_{(i,j)\in\underline{\Omega}}\left\|H_{(i)}\right\|_2^2\left\|H_{(j)}\right\|_2^2 \\ & \overset{(a)}{\le} (1+\delta)\Big(\sum_{i\in[n_1+n_2]}\left\|H_{(i)}\right\|_2^2\Big)^2 + c_2\sqrt{\frac{n_1+n_2}{p}}\sum_{i\in[n_1+n_2]}\left\|H_{(i)}\right\|_2^4 \\ & \le (1+\delta)\left\|H\right\|_F^4 + c_2\sqrt{\frac{n_1+n_2}{p}}\left\|H\right\|_F^2\left\|H\right\|_{2,\infty}^2 \\ & \overset{(b)}{\le} \|H\|_F^2\left((1+\delta)\left\|H\right\|_F^2 + \sqrt{\frac{81c_2^2\mu^2r^2\sigma_1^{\star 2}(n_1+n_2)}{p(n_1\wedge n_2)^2}}\right), \end{aligned} \tag{4.28}$$

where (a) follows from Lemma 4.7 and (b) follows from $\|H\|_{2,\infty} \le 3\sqrt{\frac{\mu r}{n_1\wedge n_2}\sigma_1^\star}$.

Let us further assume $p \ge \frac{162c_2^2\mu^2r^2\kappa^2\gamma}{\delta^2(n_1\wedge n_2)}$, where $\gamma = n/(n_1\wedge n_2)$ is a fixed constant, then we can bound

$$p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2 \le \quad \|H\|_F^2\left((1+\delta)\left\|H\right\|_F^2 + \delta\sigma_r^\star\right). \tag{4.29}$$

The final threshold we obtain is thus $p \ge \frac{c}{\delta^2}\max\left(\frac{\log(n_1+n_2)}{n_1+n_2}, \frac{\mu^2r^2\kappa^2}{n_1\wedge n_2}\right)$ for some constant c. □

Lemma 4.9. *There exists a constant c, if $p \ge \dfrac{c\log n}{n_1\wedge n_2}$, then with probability at least $1-2n_1^{-4}-2n_2^{-4}$, uniformly for all matrices A, B such that $AB^\top$ is of size $(n_1+n_2)\times(n_1+n_2)$,*

$$p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(AB^\top)\right\|_F^2 \le 2n\min\left\{\left\|A\right\|_F^2\left\|B\right\|_{2,\infty}^2, \left\|B\right\|_F^2\left\|A\right\|_{2,\infty}^2\right\}$$

Proof. Let $\Omega_{Y_i} = \{j : (i,j)\in\underline{\Omega}\}$ denote the set of entries sampled in the ith row of $AB^\top$. Note that because of the structure of $\underline{\Omega}$, at most n_2 entries are sampled at the frist n_1 rows, and at most n_1 entries are sampled at the rest n_2 rows.

Using a binomial tail bound, if $p \ge \dfrac{c\log n_2}{n_2}$ for sufficiently large c, the event $\max_{i\in[n_1]}|\Omega_{Y_i}| \le 2pn_2$ holds with probability at least $1-n_2^{-4}$. Similarly for the rest n_2 rows. Hence, if $p \ge \dfrac{c\log n}{n_1\wedge n_2}$ for some constant c, with probability at least $1-n_1^{-4}-n_2^{-4}$, we have $\max_{i\in[n_1+n_2]}|\Omega_{Y_i}| \le 2pn$.

Conditioning on this event, we then have for all A, B of proper size,

$$\begin{aligned} p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(AB^\top)\right\|_F^2 = & \; p^{-1}\sum_{i=1}^{n_1+n_2}\sum_{j\in\Omega_{Y_i}}\langle A_{(i)}, B_{(j)}\rangle^2 \\ \le & \; p^{-1}\sum_{i=1}^{n_1+n_2}\left\|A_{(i)}\right\|_2^2\sum_{j\in\Omega_{Y_i}}\left\|B_{(j)}\right\|_2^2 \\ \le & \; p^{-1}\sum_{i=1}^{n_1+n_2}\left\|A_{(i)}\right\|_2^2\max_{i\in[n_1+n_2]}|\Omega_{Y_i}|\,\|B\|_{2,\infty}^2 \\ \le & \; 2n\,\|A\|_F^2\,\|B\|_{2,\infty}^2\,. \end{aligned}$$

Similarly we can prove with probability at least $1 - n_1^{-4} - n_2^{-4}$,

$$p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(AB^\top)\right\|_F^2 \le 2n\,\|B\|_F^2\,\|A\|_{2,\infty}^2\,.$$

□

The following lemma establishes restricted strong convexity and smoothness of the observation operator for matrices in T.

Lemma 4.10. *Let T be the subspace defined in* (4.25)*. There exists a universal constant c such that, if $p \ge \dfrac{c}{\delta^2}\dfrac{\mu r\log n}{n_1\wedge n_2}$, with probability at least $1-3n^{-3}$, uniformly for all $A\in T$, we have*

$$p(1-\delta)\,\|A\|_F^2 \le \|\mathcal{P}_\Omega(A)\|_F^2 \le p(1+\delta)\,\|A\|_F^2\,. \tag{4.30}$$

Consequently, uniformly for all $A, B \in T$,

$$|p^{-1}\langle\mathcal{P}_\Omega(A),\mathcal{P}_\Omega(B)\rangle - \langle A,B\rangle| \le \delta\,\|A\|_F\,\|B\|_F\,. \tag{4.31}$$

Proof. By Lemma 4.6, with probability at least $1-3n^{-3}$, for any $X\in\mathbb{R}^{n_1\times n_2}$ it holds that

$$p(1-\delta)\,\|X\|_F \le \|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T(X)\|_F \le p(1+\delta)\,\|X\|_F\,. \tag{4.32}$$

Let A be a matrix in T. Rewriting $\|\mathcal{P}_\Omega(A)\|_F^2 = \langle \mathcal{P}_\Omega \mathcal{P}_T(A), \mathcal{P}_\Omega \mathcal{P}_T(A)\rangle = \langle A, \mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T(A)\rangle$, and using the Cauchy-Schwarz inequality and (4.32) we can bound

$$\|\mathcal{P}_\Omega(A)\|_F^2 \leq p(1+\delta)\,\|A\|_F^2\,. \tag{4.33}$$

In addition, we have

$$\begin{aligned}
\|\mathcal{P}_\Omega(A)\|_F^2 &= \langle A, \mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T(A)\rangle \\
&= \langle A, \mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T(A) - p\mathcal{P}_T(A) + p\mathcal{P}_T(A)\rangle \\
&\geq -\|A\|_F\,\|(\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - p\mathcal{P}_T)(A)\|_F + p\,\|A\|_F^2 \\
&\overset{(a)}{\geq} p(1-\delta)\,\|A\|_F^2\,,
\end{aligned} \tag{4.34}$$

where (a) follows from Lemma 4.6. Combining (4.33) and (4.34) proves (4.30). To show (4.31), let $A' = \frac{A}{\|A\|_F}$ and $B' = \frac{B}{\|B\|_F}$. Both $A' + B'$ and $A' - B'$ are in T. We have

$$\begin{aligned}
\langle \mathcal{P}_\Omega(A'), \mathcal{P}_\Omega(B')\rangle &= \frac{1}{4}\left\{ \overbrace{\left\|\mathcal{P}_\Omega(A'+B')\right\|_F^2}^{①} - \overbrace{\left\|\mathcal{P}_\Omega(A'-B')\right\|_F^2}^{②} \right\} \\
&\overset{(b)}{\leq} \frac{1}{4}\left\{ (1+\delta)p\left\|A'+B'\right\|_F^2 - (1-\delta)p\left\|A'-B'\right\|_F^2 \right\} \\
&= \frac{1}{4}\left\{ 2\delta p\left(\left\|A'\right\|_F^2 + \left\|B'\right\|_F^2\right) + 4p\langle A', B'\rangle \right\} \\
&= p\delta + p\langle A', B'\rangle,
\end{aligned} \tag{4.35}$$

where (b) follows from (4.30). Thus, we have

$$p^{-1}\langle \mathcal{P}_\Omega(A), \mathcal{P}_\Omega(B)\rangle = p^{-1}\,\|A\|_F\,\|B\|_F\,\langle \mathcal{P}_\Omega(A'), \mathcal{P}_\Omega(B')\rangle \leq \delta\,\|A\|_F\,\|B\|_F + \langle A, B\rangle. \tag{4.36}$$

Similarly, we can show

$$p^{-1}\langle \mathcal{P}_\Omega(A), \mathcal{P}_\Omega(B)\rangle \geq -\delta\,\|A\|_F\,\|B\|_F + \langle A, B\rangle. \tag{4.37}$$

□

Last, we want to show the projection onto feasible set $\mathcal{C}$ is a contraction.

Lemma 4.11. *Let $y \in \mathbb{R}^r$ be a vector such that $\|y\|_2 \leq \theta$, for any $x \in \mathbb{R}^r$. Then*

$$\left\|\mathcal{P}_{\|\cdot\|_2 \leq \theta}(x) - y\right\|_2^2 \leq \|x - y\|_2^2 .$$

Proof. If $\|x\|_2 \leq \theta$, then $\mathcal{P}_{\|\cdot\|_2 \leq \theta}(x) = x$. Otherwise $\mathcal{P}_{\|\cdot\|_2 \leq \theta}(x) = \theta\bar{x}$, where $\bar{x} = \frac{x}{\|x\|_2}$. Write $y = (y^\top \bar{x})\bar{x} + \mathcal{P}_x^\perp(y)$, we have

$$\|\theta\bar{x} - y\|_2^2 = \left\|\theta\bar{x} - (y^\top \bar{x})\bar{x}\right\|_2^2 + \left\|\mathcal{P}_x^\perp(y)\right\|_2^2 = (\theta - y^\top \bar{x})^2 + \left\|\mathcal{P}_x^\perp(y)\right\|_2^2. \tag{4.38}$$

It suffices to show

$$(\theta - y^\top \bar{x})^2 \leq (\|x\| - y^\top \bar{x})^2. \tag{4.39}$$

If $y^\top \bar{x} \leq 0$, then (4.39) holds because $\|x\| > \theta$. If $y^\top \bar{x} > 0$, (4.39) still holds since $\|x\| > \theta \geq \|y\| \geq y^\top \bar{x}$. □

4.6.2 Initialization

Proof of Lemma 4.1

Let δ denote the upper bound of $\left\|p^{-1}\mathcal{P}_\Omega(X^\star) - X^\star\right\|$ as in Lemma 4.5, and let $\sigma_1 \geq \ldots \geq \sigma_n$ denote the singular values of $p^{-1}\mathcal{P}_\Omega(X^\star)$. By Weyl's theorem, we have

$$|\sigma_i - \sigma_i^\star| \leq \delta, \quad i \in [n]. \tag{4.40}$$

Note this implies $\sigma_{r+1} \leq \delta$, as $\sigma_{r+1}^\star = 0$.

By definition, $Z^0 = [U; V]\Sigma^{\frac{1}{2}}$, where $U\Sigma V^\top$ is the rank-r SVD of $p^{-1}\mathcal{P}_\Omega(X^\star)$. According

to Lemma 4.4, one has

$$\begin{aligned}\left\|Z^0 Z^{0\top} - Z^\star Z^{\star\top}\right\|_F &\leq 2\left\|U\Sigma V^\top - X^\star\right\|_F \\ &\overset{(a)}{\leq} 2\sqrt{2r}\left\|U\Sigma V^\top - X^\star\right\| \\ &\leq 2\sqrt{2r}\left(\left\|U\Sigma V^\top - p^{-1}\mathcal{P}_\Omega(X^\star)\right\| + \left\|p^{-1}\mathcal{P}_\Omega(X^\star) - X^\star\right\|\right) \\ &\overset{(b)}{\leq} 2\sqrt{2r}\,(\delta+\delta) \\ &= 4\sqrt{2r}\delta,\end{aligned} \tag{4.41}$$

where (a) holds because $\text{rank}(U\Sigma V^\top - X^\star) \leq 2r$, (b) holds since $\left\|U\Sigma V^\top - p^{-1}\mathcal{P}_\Omega(X^\star)\right\| = \sigma_{r+1} \leq \delta$.

Let $H = Z^0 - \overline{Z}^0$. We want to bound $d(Z^0, Z^\star)^2 = \|H\|_F^2$. According to (4.20), $H^\top \overline{Z}^0$ is symmetric and ${Z^0}^\top \overline{Z}^0$ is positive semidefinite. Hence we can write

$$\begin{aligned}&\left\|Z^0 Z^{0\top} - Z^\star Z^{\star\top}\right\|_F^2 \\ &\quad= \left\|H\overline{Z}^{0\top} + \overline{Z}^0 H^\top + HH^\top\right\|_F^2 \\ &\quad= \text{tr}\left((H^\top H)^2 + 2(H^\top \overline{Z}^0)^2 + 2(H^\top H)({\overline{Z}^0}^\top \overline{Z}^0) + 4(H^\top H)(H^\top \overline{Z}^0)\right) \\ &\quad= \text{tr}\left(\left(H^\top H + \sqrt{2}H^\top \overline{Z}^0\right)^2 + (4-2\sqrt{2})(H^\top H)(H^\top \overline{Z}^0) + 2(H^\top H)({\overline{Z}^0}^\top \overline{Z}^0)\right) \\ &\quad\geq \text{tr}\left((4-2\sqrt{2})(H^\top H)(H^\top \overline{Z}^0) + 2(H^\top H)({\overline{Z}^0}^\top \overline{Z}^0)\right) \\ &\quad= (4-2\sqrt{2})\,\text{tr}\left((H^\top H)({Z^0}^\top \overline{Z}^0)\right) + (2\sqrt{2}-2)\left\|H\overline{Z}^{0T}\right\|_F^2,\end{aligned} \tag{4.42}$$

where in the second line we used that $H^\top \overline{Z}^0$ is symmetric. Besides, as ${Z^0}^\top \overline{Z}^0$ is positive semidefinite, $(4-\sqrt{2})\,\text{tr}((H^\top H)({Z^0}^\top \overline{Z}^0))$ is nonnegative. Therefore,

$$\left\|Z^0 Z^{0\top} - Z^\star Z^{\star\top}\right\|_F^2 \geq (2\sqrt{2}-2)\left\|H\overline{Z}^{0\top}\right\|_F^2 \geq 4(\sqrt{2}-1)\sigma_r^\star \|H\|_F^2\,. \tag{4.43}$$

Combining (4.41) and (4.43), it follows that

$$d(Z^0, Z^\star)^2 \le \frac{\left\| Z^0 Z^0 - Z^\star Z^{\star\top} \right\|_F^2}{4(\sqrt{2}-1)\sigma_r^\star} \le \frac{8r}{(\sqrt{2}-1)\sigma_r^\star}\delta^2. \tag{4.44}$$

Therefore, it suffices to show

$$\begin{aligned} d(Z^0, Z^\star)^2 &\le \frac{8r}{(\sqrt{2}-1)\sigma_r^\star}\delta^2 \\ &\overset{(a)}{=} c\frac{r}{\sigma_r^\star}\left(\frac{\log n}{p}\left\|X^\star\right\|_\infty + \sqrt{\frac{\log n}{p}}\left\|X^\star\right\|_{\infty,2}\right)^2 \\ &\overset{(b)}{\le} c\, r\frac{\sigma_1^{\star 2}}{\sigma_r^\star}\left(\frac{\mu r \log n}{p(n_1 \wedge n_2)} + \sqrt{\frac{\mu r \log n}{p(n_1 \wedge n_2)}}\right)^2 \\ &\le \frac{1}{16}\sigma_r^\star, \end{aligned} \tag{4.45}$$

where in (a) we replaced δ using Lemma 4.5, and (b) holds since by our incoherence assumption (4.3) we have

$$\left\|X^\star\right\|_\infty = \left\|U^\star \Sigma^\star V^{\star\top}\right\|_\infty \le \sigma_1^\star \max_{i,j}\left\|U_{(i)}^\star\right\|\left\|V_{(j)}^\star\right\| \le \sigma_1^\star \left\|U^\star\right\|_{2,\infty}\left\|V^\star\right\|_{2,\infty} \le \sigma_1^\star \frac{\mu r}{n_1 \wedge n_2}, \tag{4.46}$$

$$\left\|X^\star\right\|_{\infty,2} = \left\|U^\star \Sigma^\star V^{\star\top}\right\|_{\infty,2} \le \sigma_1^\star \left\|U^\star V^{\star\top}\right\|_{\infty,2} \overset{(c)}{\le} \sigma_1^\star \sqrt{\frac{\mu r}{n_1 \wedge n_2}}. \tag{4.47}$$

Note that for (c) we used $\left\|AB^\top\right\|_{2,\infty} \le \|A\|_{2,\infty}\|B\|$.

Hence, to obtain $d(Z^0, Z^\star)^2 \le \frac{1}{16}\sigma_r^\star$, it suffices to have

$$p \ge \max\left\{\frac{c\mu r^{3/2}\kappa \log n}{n_1 \wedge n_2}, \frac{c\mu r^2 \kappa^2 \log n}{n_1 \wedge n_2}\right\} = \frac{c\mu r^2 \kappa^2 \log n}{n_1 \wedge n_2}. \tag{4.48}$$

Since $\mathcal{P}_{\mathcal{C}}$ is just row-wise clipping, by Lemma 4.11 we have

$$d(Z^1, Z^\star)^2 \le \left\|\mathcal{P}_{\mathcal{C}}(Z^0) - Z^\star\right\|_F^2 \le \left\|Z^0 - Z^\star\right\|_F^2.$$

Proof of Corollary 4.1

By the incoherence assumption, we have $\|Z^\star\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_1 \wedge n_2}\sigma_1^\star}$, see (4.17). It suffices to show $2\sigma_1 \geq \sigma_1^\star$. From the above discussion, we can see that

$$\frac{8r}{(\sqrt{2}-1)\sigma_r^\star}\delta^2 \leq \frac{1}{16}\sigma_r^\star \Rightarrow \delta \leq \frac{1}{16}\sigma_r^\star.$$

By Wely's theorem, we have $|\sigma_1 - \sigma_1^\star| \leq \frac{1}{16}\sigma_r^\star$. As a result, $2\sigma_1 \geq \sigma_1^\star$.

4.6.3 Regularity Condition

Analogous to the restricted strong convexity (RSC) and restricted strong smoothness (RSS), we show that with high probability our objective function f satisfies the local curvature and local smoothness conditions defined below.

- *Local Curvature Condition*

  There exists constant c_1, c_2 such that for any $Z \in \mathcal{C}$ satisfying $d(Z, Z^\star) \leq \frac{1}{4}\sqrt{\sigma_r^\star}$,

$$\langle \nabla f(Z), H \rangle \geq c_1 \|H\|_F^2 + c_2 \left\| H^\top D\overline{Z} \right\|_F^2.$$

- *Local Smoothness Condition*

  There exist constants c_3, c_4 such that for any $Z \in \mathcal{C}$ satisfying $d(Z, Z^\star) \leq \frac{1}{4}\sqrt{\sigma_r^\star}$,

$$\|\nabla f(Z)\|_F^2 \leq c_3 \|H\|_F^2 + c_4 \left\| H^\top D\overline{Z} \right\|_F^2.$$

Proof of the Local Curvature Condition

$$
\begin{aligned}
&\langle \nabla f(Z), H\rangle \\
&= \frac{1}{p}\left(\sum_{l=1}^{2m}\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top + HH^\top\rangle \cdot \langle (A_l + A_l^\top)(\overline{Z}+H), H\rangle\right) + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&\overset{(i)}{=} \frac{1}{p}\left(\sum_{l=1}^{2m}\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top + HH^\top\rangle \cdot \langle A_l, H\overline{Z}^\top + \overline{Z}H^\top + 2HH^\top\rangle\right) + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&= \frac{1}{p}\Bigg\{\overbrace{\sum_{l=1}^{2m}\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top\rangle^2}^{a^2} + \overbrace{\sum_{l=1}^{2m} 2\langle A_l, HH^\top\rangle^2}^{b^2} + \sum_{l=1}^{2m} 3\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top\rangle\langle A_l, HH^\top\rangle\Bigg\} \\
&\qquad + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&\overset{(ii)}{\geq} \frac{1}{p}\Bigg\{a^2 + b^2 - \frac{3}{\sqrt{2}}\overbrace{\sqrt{\sum_{l=1}^{2m}\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top\rangle^2}}^{a}\overbrace{\sqrt{\sum_{l=1}^{2m} 2\langle A_l, HH^\top\rangle^2}}^{b}\Bigg\} + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&= \frac{1}{p}\left\{\left(a - \frac{3}{2\sqrt{2}}b\right)^2 - \frac{1}{8}b^2\right\} + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&\overset{(iii)}{\geq} \frac{1}{p}\left(\frac{a^2}{2} - \frac{5}{4}b^2\right) + \lambda\,\mathrm{tr}(H^\top\Gamma) \\
&= \frac{1}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top)\right\|_F^2 - \frac{5}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2 + \lambda\,\mathrm{tr}(H^\top\Gamma).
\end{aligned}
\tag{4.49}
$$

where we used equation (4.19) for (i), the Cauchy-Schwarz inequality for (ii), inequality $(a-b)^2 \geq \frac{a^2}{2} - b^2$ for (iii). Finally, in the last line we used $\sum_{l=1}^{2m}\langle A_l, M\rangle^2 = \left\|\mathcal{P}_{\underline{\Omega}}(M)\right\|_F^2$.

We first lower bound $\frac{1}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top)\right\|_F^2$. By the symmetry of $\underline{\Omega}$, it is equal to $p^{-1}\left\|\mathcal{P}_{\Omega}(H_U\overline{Z}_V^\top + \overline{Z}_U H_V^\top)\right\|_F^2$, which expands to

$$
p^{-1}\left\|\mathcal{P}_{\Omega}(H_U\overline{Z}_V^\top)\right\|_F^2 + p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(\overline{Z}_U H_V^\top)\right\|_F^2 + 2p^{-1}\langle \mathcal{P}_{\Omega}(H_U\overline{Z}_V^\top), \mathcal{P}_{\Omega}(\overline{Z}_U H_V^\top)\rangle. \tag{4.50}
$$

As both $H_U\overline{Z}_V^\top$ and $\overline{Z}_U H_V^\top$ belong to T, we use Lemma 4.10 to lower bound above three terms, respectively. This gives us

$$
\begin{aligned}
&\frac{1}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top+\overline{Z}H^\top)\right\|_F^2 \\
&\geq (1-\delta)\left(\left\|H_U\overline{Z}_V^\top\right\|_F^2+\left\|\overline{Z}_U H_V^\top\right\|_F^2\right)+2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle-2\delta\left\|H_U\overline{Z}_V^\top\right\|_F\left\|\overline{Z}_U H_V^\top\right\|_F \\
&\geq (1-\delta)\left(\left\|H_U\overline{Z}_V^\top\right\|_F^2+\left\|\overline{Z}_U H_V^\top\right\|_F^2\right)+2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle \\
&\quad -\delta\left(\left\|H_U\overline{Z}_V^\top\right\|_F^2+\left\|\overline{Z}_U H_V^\top\right\|_F^2\right) \\
&\overset{(iv)}{\geq}(1-2\delta)\sigma_r^\star\left(\|H_U\|_F^2+\|H_V\|_F^2\right)+2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle \\
&=(1-2\delta)\sigma_r^\star\|H\|_F^2+2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle.
\end{aligned}
\tag{4.51}
$$

where we used $\left\|H_U\overline{Z}_V^\top\right\|_F^2\geq\sigma_r^\star\|H_U\|_F^2$ and $\left\|\overline{Z}_U H_V^\top\right\|_F^2\geq\sigma_r^\star\|H_V\|_F^2$ for (iv).

Until now, we obtain

$$
\langle\nabla f(Z),H\rangle\geq(1-2\delta)\sigma_r^\star\|H\|_F^2+2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle+\lambda\,\mathrm{tr}(H^\top\Gamma)-\frac{5}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2. \tag{4.52}
$$

Next, we lower bound $2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle+\lambda\,\mathrm{tr}(H^\top\Gamma)$ together. Rewriting

$$
\begin{aligned}
&2\langle H_U\overline{Z}_V^\top,\overline{Z}_U H_V^\top\rangle=\langle H,\begin{bmatrix}0 & \overline{Z}_U H_V^\top\\ \overline{Z}_V H_U^\top & 0\end{bmatrix}\overline{Z}\rangle=\langle H,\frac{1}{2}(\overline{Z}H^\top-D\overline{Z}H^\top D)\overline{Z}\rangle, \\
&ZZ^\top-\overline{Z}\overline{Z}^\top=HH^\top+\overline{Z}H^\top+H\overline{Z}^\top,
\end{aligned}
\tag{4.53}
$$

and plugging in $\Gamma = DZZ^\top DZ$, we then have

$$
\begin{aligned}
&2\langle H_U\overline{Z}_V^\top, \overline{Z}_U H_V^\top\rangle + \lambda \operatorname{tr}(H^\top\Gamma) \\
&= \langle H, \frac{1}{2}(\overline{Z}H^\top - D\overline{Z}H^\top D)\overline{Z}\rangle + \lambda\langle H, D(ZZ^\top - \overline{Z}\overline{Z}^\top)DZ\rangle + \lambda\langle H, D(\overline{Z}\overline{Z}^\top)D\overline{Z}\rangle \\
&\quad + \lambda\langle H, D(\overline{Z}\overline{Z}^\top)DH\rangle \\
&\overset{(a)}{=} \langle H, \frac{1}{2}(\overline{Z}H^\top - D\overline{Z}H^\top D)\overline{Z}\rangle + \lambda\langle H, D(ZZ^\top - \overline{Z}\overline{Z}^\top)DZ\rangle + \lambda\left\|\overline{Z}^\top DH\right\|_F^2 \\
&\overset{(b)}{=} \lambda\left\|\overline{Z}^\top DH\right\|_F^2 + \langle H, \frac{1}{2}(\overline{Z}H^\top - D\overline{Z}H^\top D)\overline{Z} + \lambda D(HH^\top + \overline{Z}H^\top + H\overline{Z}^\top)D(\overline{Z}+H)\rangle \\
&\overset{(c)}{=} \lambda\left\|\overline{Z}^\top DH\right\|_F^2 + \frac{1}{2}\left\|H^\top\overline{Z}\right\|_F^2 + \lambda\left\|H^\top DH\right\|_F^2 + 3\lambda\operatorname{tr}(H^\top DHH^\top D\overline{Z}) \\
&\quad + \left(\lambda - \frac{1}{2}\right)\operatorname{tr}(H^\top D\overline{Z}H^\top D\overline{Z}) \\
&= \frac{\lambda}{2}\left\|\overline{Z}^\top DH\right\|_F^2 + \frac{\lambda}{2}\left\|\overline{Z}^\top DH + 3H^\top DH\right\|_F^2 - \frac{7}{2}\lambda\left\|H^\top DH\right\|_F^2 \\
&\quad + \frac{1}{2}\left\|H^\top Z\right\|_F^2 + \left(\lambda - \frac{1}{2}\right)\operatorname{tr}(H^\top D\overline{Z}H^\top D\overline{Z}) \\
&\geq \frac{\lambda}{2}\left\|\overline{Z}^\top DH\right\|_F^2 - \frac{7}{2}\lambda\left\|H\right\|_F^4 + \left(\lambda - \frac{1}{2}\right)\operatorname{tr}(H^\top D\overline{Z}H^\top D\overline{Z})
\end{aligned}
\tag{4.54}
$$

Equality (a) holds because $\overline{Z}^\top D\overline{Z} = 0$. We plug in (4.53) in (b). For (c), we use $\overline{Z}^\top D\overline{Z} = 0$ and that $H^\top\overline{Z}$ is symmetric. Finally, we take $\lambda = \frac{1}{2}$ and use Lemma 4.8 to upper bound $p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2$:

$$
\begin{aligned}
\langle\nabla f(Z), H\rangle \geq\ & (1-2\delta)\sigma_r^\star\left\|H\right\|_F^2 + \frac{1}{4}\left\|\overline{Z}^\top DH\right\|_F^2 - \frac{7}{4}\left\|H\right\|_F^4 - \frac{5}{2}(1+\delta)\left\|H\right\|_F^4 - \frac{5}{2}\delta\sigma_r^\star\left\|H\right\|_F^2 \\
=\ & \left((1-2\delta)\sigma_r^\star - \frac{5}{2}(\frac{17}{10}+\delta)\left\|H\right\|_F^2 - \frac{5}{2}\delta\sigma_r^\star\right)\left\|H\right\|_F^2 + \frac{1}{4}\left\|\overline{Z}^\top DH\right\|_F^2.
\end{aligned}
\tag{4.55}
$$

For simplicity, we take $\delta = \frac{1}{16}$. We also have $\left\|H\right\|_F^2 \leq \frac{1}{16}\sigma_r^\star$. This leads to

$$
\langle\nabla f(Z), H\rangle \geq \frac{227}{512}\sigma_r^\star\left\|H\right\|_F^2 + \frac{1}{4}\left\|\overline{Z}^\top DH\right\|_F^2. \tag{4.56}
$$

Note that this lower bound holds with high probability uniformly for all $Z \in \mathcal{C}$ such that $d(Z, Z^\star) \le \frac{1}{4}\sqrt{\sigma_r^\star}$, since Lemma 4.8 and 4.10 hold uniformly.

When the ground truth $X^\star$ is positive semidefinite, we don't need to do lifitng nor impose the regularizer. Using Lemma 4.10, we can lower bound $\frac{1}{2}p^{-1}\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top)\right\|_F^2 \gtrsim (1-\delta)\sigma_r^\star \|H\|_F^2$ directly. Taking proper constants, we can obtain the standard restricted strong convexity condition:

$$\langle \nabla f(Z), H\rangle \gtrsim \sigma_r^\star \|H\|_F^2 .$$

Proof of the Local Smoothness Condition

To upper bound $\|\nabla f(Z)\|_F^2 = \max_{\|W\|_F=1} |\langle \nabla f(Z), W\rangle|^2$, it suffices to show that for any $n \times r$ W of unit Frobenius norm, $|\langle \nabla f(Z), W\rangle|^2$ is upper bounded. We first write

$$\begin{aligned}
&\langle \nabla f(Z), W\rangle \\
&= \frac{1}{p}\sum_{l=1}^{2m}\Big(\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top\rangle + \langle A_l, HH^\top\rangle\Big)\cdot\langle (A_l + A_l^\top)(\overline{Z}+H), W\rangle + \lambda \operatorname{tr}(W^\top\Gamma) \\
&\overset{(i)}{=} \frac{1}{p}\sum_{l=1}^{2m}\Big(\langle A_l, H\overline{Z}^\top + \overline{Z}H^\top\rangle + \langle A_l, HH^\top\rangle\Big)\Big(\langle A_l, W\overline{Z}^\top + \overline{Z}W^\top\rangle + \langle A_l, WH^\top + HW^\top\rangle\Big) \\
&\quad + \lambda \operatorname{tr}(W^\top\Gamma) \\
&= \frac{1}{p}\Big\{\langle \mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top), \mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top + \overline{Z}W^\top)\rangle + \langle \mathcal{P}_{\underline{\Omega}}(HH^\top), \mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top + \overline{Z}W^\top)\rangle \\
&\quad + \langle \mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top), \mathcal{P}_{\underline{\Omega}}(WH^\top + HW^\top)\rangle + \langle \mathcal{P}_{\underline{\Omega}}(HH^\top), \mathcal{P}_{\underline{\Omega}}(WH^\top + HW^\top)\rangle\Big\} \\
&\quad + \lambda \operatorname{tr}(W^\top\Gamma),
\end{aligned} \tag{4.57}$$

where we used (4.19) for (i). Since $(a+b+c+d+e)^2 \le 5(a^2+b^2+c^2+d^2+e^2)$, we have

$$
\begin{aligned}
&|\langle \nabla f(Z), W\rangle|^2 \\
&\le \frac{5}{p^2}\Big\{\langle \mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top), \mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top + \overline{Z}W^\top)\rangle^2 + \langle \mathcal{P}_{\underline{\Omega}}(HH^\top), \mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top + \overline{Z}W^\top)\rangle^2 \\
&\quad + \langle \mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top), \mathcal{P}_{\underline{\Omega}}(WH^\top + HW^\top)\rangle^2 + \langle \mathcal{P}_{\underline{\Omega}}(HH^\top), \mathcal{P}_{\underline{\Omega}}(WH^\top + HW^\top)\rangle^2\Big\} \\
&\quad + 5\lambda^2 \operatorname{tr}(W^\top\Gamma)^2 \\
&\overset{(ii)}{\le} \frac{5}{p^2}\left(\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top + \overline{Z}H^\top)\right\|_F^2 + \left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2\right) \\
&\qquad \cdot \left(\left\|\mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top + \overline{Z}W^\top)\right\|_F^2 + \left\|\mathcal{P}_{\underline{\Omega}}(WH^\top + HW^\top)\right\|_F^2\right) + 5\lambda^2 \|\Gamma\|_F^2 \overbrace{\|W\|_F^2}^{=1} \\
&\overset{(iii)}{\le} \frac{5}{p}\Bigg(2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(H\overline{Z}^\top)\right\|_F^2}^{①} + 2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(\overline{Z}H^\top)\right\|_F^2}^{②} + \overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(HH^\top)\right\|_F^2}^{③}\Bigg) \\
&\qquad \cdot \frac{1}{p}\Bigg(2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(W\overline{Z}^\top)\right\|_F^2}^{④} + 2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(\overline{Z}W^\top)\right\|_F^2}^{⑤} + 2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(WH^\top)\right\|_F^2}^{⑥} + 2\overbrace{\left\|\mathcal{P}_{\underline{\Omega}}(HW^\top)\right\|_F^2}^{⑦}\Bigg) \\
&\qquad + 5\lambda^2 \|\Gamma\|_F^2 ,
\end{aligned}
\tag{4.58}
$$

where we used the Cauchy-Schwarz inequality for (ii), and $(a+b)^2 \le 2(a^2+b^2)$ for (iii). We then use Lemma 4.9 to upper bound ①, ②, ④, ⑤, ⑥, ⑦, and Lemma 4.8 for ③. Also since $\|W\|_F = 1$, one has

$$
\begin{aligned}
&|\langle \nabla f(Z), W\rangle|^2 \\
&\le \; 5\left(8n\|H\|_F^2\left\|\overline{Z}\right\|_{2,\infty}^2 + (1+\delta)\|H\|_F^4 + \delta\sigma_r^\star\|H\|_F^2\right)\cdot\left(8n\left\|\overline{Z}\right\|_{2,\infty}^2 + 8n\|H\|_{2,\infty}^2\right) \\
&\qquad + 5\lambda^2\|\Gamma\|_F^2 \\
&= \; 40n\left(8n\left\|\overline{Z}\right\|_{2,\infty}^2 + (1+\delta)\|H\|_F^2 + \delta\sigma_r^\star\right)\|H\|_F^2\cdot\left(\left\|\overline{Z}\right\|_{2,\infty}^2 + \|H\|_{2,\infty}^2\right) + 5\lambda^2\|\Gamma\|_F^2 \\
&\le \; 400\mu r\sigma_1^\star\left(8\mu r\sigma_1^\star + (1+\delta)\|H\|_F^2 + \delta\sigma_r^\star\right)\|H\|_F^2 + 5\lambda^2\|\Gamma\|_F^2 ,
\end{aligned}
\tag{4.59}
$$

where in the last line we plugged in $\left\|\overline{Z}\right\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n}\sigma_1^\star}$ and $\|H\|_{2,\infty} \leq 3\sqrt{\frac{\mu r}{n}\sigma_1^\star}$, i.e. (4.17) and (4.18).

Next, we bound

$$
\begin{aligned}
\|\Gamma\|_F^2 &= \left\|D(ZZ^\top - \overline{Z}\overline{Z}^\top)DZ + D\overline{Z}\overline{Z}^\top DZ\right\|_F^2 \\
&\leq 2\left\|D(ZZ^\top - \overline{Z}\overline{Z}^\top)DZ\right\|_F^2 + 2\left\|D\overline{Z}\overline{Z}^\top DZ\right\|_F^2 \\
&\overset{(a)}{\leq} 2\left\|ZZ^\top - \overline{Z}\overline{Z}^\top\right\|_F^2 \|Z\|^2 + 2\left\|\overline{Z}\right\|^2 \left\|\overline{Z}^\top DZ\right\|_F^2 \\
&\overset{(b)}{=} 2\left\|HH^\top + \overline{Z}H^\top + H\overline{Z}^\top\right\|_F^2 \|Z\|^2 + 2\left\|\overline{Z}\right\|^2 \left\|\overline{Z}^\top DH\right\|_F^2 \\
&\leq 6\left(\left\|HH^\top\right\|_F^2 + \left\|\overline{Z}H^\top\right\|_F^2 + \left\|H\overline{Z}^\top\right\|_F^2\right)\|Z\|^2 + 2\left\|\overline{Z}\right\|^2 \left\|\overline{Z}^\top DH\right\|_F^2 \\
&\overset{(c)}{\leq} 6\left(\|H\|_F^2 + 2\left\|\overline{Z}\right\|^2\right)\|H\|_F^2 \|Z\|^2 + 2\left\|\overline{Z}\right\|^2 \left\|\overline{Z}^\top DH\right\|_F^2 \\
&\overset{(d)}{=} 6\left(\|H\|_F^2 + 4\sigma_1^\star\right)\|H\|_F^2 \|Z\|^2 + 4\sigma_1^\star \left\|\overline{Z}^\top DH\right\|_F^2 .
\end{aligned}
\tag{4.60}
$$

Inequality (a) holds because $\|AB\|_F \leq \|A\|\,\|B\|_F$ and $\|D\| = 1$. To get (b), for the first term in the 3rd line we expand $ZZ^\top - \overline{Z}\overline{Z}^\top$, for the second term we expand $Z = \overline{Z} + H$ and use $\overline{Z}^\top D\overline{Z} = 0$. For (c), we use $\|AB\|_F \leq \|A\|\,\|B\|_F \leq \|A\|_F\,\|B\|_F$. Last, (d) holds because $\left\|\overline{Z}\right\|^2 = 2\sigma_1^\star$.

Finally, we combine (4.59) and (4.60). As before, take $\lambda = \frac{1}{2}$, $\delta = \frac{1}{16}$, and $\|H\|_F^2 \leq \frac{1}{16}\sigma_r^\star$, we obtain

$$
\begin{aligned}
&\|\nabla f(Z)\|_F^2 \\
&\leq \left\{400\mu r\sigma_1^\star \left(8\mu r\sigma_1^\star + (1+\delta)\|H\|_F^2 + \delta\sigma_r^\star\right) + 30\lambda^2\left(\|H\|_F^2 + 4\sigma_1^\star\right)\|Z\|^2\right\}\|H\|_F^2 \\
&\quad + 20\lambda^2\sigma_1^\star\left\|\overline{Z}^\top DH\right\|_F^2 \\
&\overset{(a)}{\leq} \left\{400\mu r\sigma_1^\star \left(8\mu r\sigma_1^\star + (1+\delta)\|H\|_F^2 + \delta\sigma_r^\star\right) + \frac{735}{8}\sigma_1^\star\lambda^2\left(\|H\|_F^2 + 2\sigma_1^\star\right)\right\}\|H\|_F^2 \\
&\quad + 20\lambda^2\sigma_1^\star\left\|\overline{Z}^\top DH\right\|_F^2 \\
&\overset{(b)}{\leq} \left\{400\left(8 + \frac{17}{256} + \frac{1}{16}\right) + \frac{735}{32}\left(\frac{1}{16} + 2\right)\right\}\mu^2 r^2\sigma_1^{\star 2}\|H\|_F^2 + 5\sigma_1^\star\left\|\overline{Z}^\top DH\right\|_F^2 \\
&\leq 3299\mu^2 r^2\sigma_1^{\star 2}\|H\|_F^2 + 5\sigma_1^\star\left\|\overline{Z}^\top DH\right\|_F^2,
\end{aligned}
\tag{4.61}
$$

where for (a) we used $\|Z\| \leq \|H\| + \left\|\overline{Z}\right\| \leq \frac{1}{4}\sqrt{\sigma_r^\star} + \sqrt{2\sigma_1^\star} \leq \frac{7}{4}\sqrt{\sigma_1^\star}$, for (b) we used $\mu, r \geq 1$.

As before, this condition holds uniformly for all Z such that $d(Z, Z^\star) \leq \frac{1}{4}\sqrt{\sigma_r^\star}$ and satisfying the incoherence condition.

For the case $X^\star$ is positive semidefinite, as we don't need to impose the regularizer, standard restricted strong smoothness condition follows:

$$
\|\nabla f(Z)\|_F^2 \lesssim \sigma_1^\star\|H\|_F^2.
$$

Proof of Lemma 4.3

Rearranging the terms in the smoothness condition (4.61), we can further bound

$$
\begin{aligned}
\frac{1}{4}\left\|\overline{Z}^\top DH\right\|_F^2 &\geq \frac{\|\nabla f(Z)\|_F^2}{20\mu^2 r^2\kappa\sigma_1^\star} - \frac{3299}{20}\sigma_r^\star\|H\|_F^2 \\
&\geq \frac{\|\nabla f(Z)\|_F^2}{13196\mu^2 r^2\kappa\sigma_1^\star} - \frac{128}{512}\sigma_r^\star\|H\|_F^2.
\end{aligned}
\tag{4.62}
$$

Combining equation (4.56) and (4.62), it follows that

$$\langle \nabla f(Z), H \rangle \geq \frac{99}{512}\sigma_r^\star \|H\|_F^2 + \frac{1}{13196\,\mu^2 r^2 \kappa \sigma_1^\star} \|\nabla f(Z)\|_F^2 . \tag{4.63}$$

Finally, by upper bounding the probability that Lemma 4.8, 4.9, or 4.10 fails, and the sample probability p these lemmas require, we conclude that once

$$p \geq c \max\left(\frac{\mu r \log n}{n_1 \wedge n_2}, \frac{\mu^2 r^2 \kappa^2}{n_1 \wedge n_2}\right), \tag{4.64}$$

regularity condition (4.63) holds with probability at least $1 - c_1 n^{-c_2}$, where c, c_1, c_2 are constants.

4.6.4 Linear Convergence

Proof of Lemma 4.2

Let $H^k = Z^k - \overline{Z}^k$. Our iterate is $Z^{k+1} = \mathcal{P}_{\mathcal{C}}(Z^k - \eta \nabla f(Z^k))$. Since $\mathcal{P}_{\mathcal{C}}$ is just row-wise clipping, by Lemma 4.11 we have

$$\left\| \mathcal{P}_{\mathcal{C}}\left(Z^k - \frac{\eta}{\sigma_1^\star}\nabla f(Z^k)\right) - \overline{Z}^k \right\|_F^2 \leq \left\| Z^k - \frac{\eta}{\sigma_1^\star}\nabla f(Z^k) - \overline{Z}^k \right\|_F^2 . \tag{4.65}$$

It follows that

$$
\begin{aligned}
&\left\|Z^{k+1}-\overline{Z}^{k}\right\|_F^2 \\
&\quad\leq \left\|Z^{k}-\frac{\eta}{\sigma_1^\star}\nabla f(Z^k)-\overline{Z}^{k}\right\|_F^2 \\
&\quad= \left\|H^k\right\|_F^2+\frac{\eta^2}{\sigma_1^{\star 2}}\left\|\nabla f(Z^k)\right\|_F^2-\frac{2\eta}{\sigma_1^\star}\langle\nabla f(Z^k),H^k\rangle \\
&\quad\overset{(a)}{\leq} \left\|H^k\right\|_F^2+\frac{\eta^2}{\sigma^{\star 2}_1}\left\|\nabla f(Z^k)\right\|_F^2-\frac{2\eta}{\sigma_1^\star}\left(\frac{1}{\alpha}\sigma_r^\star\left\|H^k\right\|_F^2+\frac{1}{\beta\sigma_1^\star}\left\|\nabla f(Z^k)\right\|_F^2\right) \\
&\quad= \left(1-\frac{2\eta}{\alpha\kappa}\right)\left\|H^k\right\|_F^2+\frac{\eta(\eta-2/\beta)}{\sigma^{\star 2}_1}\left\|\nabla f(Z^k)\right\|_F^2 \\
&\quad\overset{(b)}{\leq} \left(1-\frac{2\eta}{\alpha\kappa}\right)\left\|H^k\right\|_F^2,
\end{aligned}
\tag{4.66}
$$

where we use the definition of $RC(\varepsilon,\alpha,\beta)$ for (a) and $0<\eta\leq\min\{\alpha/2,2/\beta\}$ for (b). Therefore,

$$
d(Z^{k+1},Z^\star)=\min_{\widetilde{Z}\in\mathcal{S}}\left\|Z^{k+1}-\widetilde{Z}\right\|_F^2\leq\sqrt{1-\frac{2\eta}{\alpha\kappa}}d(Z^k,Z^\star). \tag{4.67}
$$

Part II

Sparse Graphs

CHAPTER 5

FASTEST MIXING MARKOV CHAIN

The recurring theme of this thesis is to reformulate a problem so that the number of parameters is the same as its intrinsic degrees of freedom, even if the resulting problem becomes nonconvex. In the previous two chapters, we have studied the situation where the target matrices are of low rank. However, in many other contexts, the matrices of interest might not be low-rank. Is there any other structure we could exploit and apply similar methods? In this chapter, we focus on the *graph Laplacian* matrix — a fundamental positive semidefinite matrix that connects graph theory, linear algebra, numerical computation and many other related fields [Boyd, 2006; Spielman, 2010]. For a connected graph with n nodes, the Laplacian matrix is of rank $n - 1$. However, when applying *Cholesky factorization* — a special symmetric decomposition where the factor is lower-triangular — to the graph Laplacian, the factors we obtained are often sparse. We consider the *fastest mixing Markov chain* (FMMC) problem [Boyd et al., 2004; Boyd, 2006], where one needs to find the best graph Laplacian matrix under certain constraints. We propose a nonconvex formulation for FMMC based on the Cholesky factorization, and study a first order method where the sparsity of the factor is utilized.

5.1 Graph Laplacian and Cholesky Factorization

The graph Laplacian is a core matrix representation of graph that naturally arises in many problems. It has various appealing algebraic properties and has received tremendous research interests. For example, solving large scale linear systems is ubiquitous in scientific computing, and efficient algorithms for solving systems in the Laplacian — more broadly, symmetric and diagonally dominant matrices — have been emerging as a primitive for other numerical methods [Spielman, 2010]. Extensive studies have shown that those systems can be solved in nearly-linear time [Spielman and Teng, 2004; Koutis et al., 2010; Cohen et al., 2014]. Our work is also inspired by the technical advances in this field. Let us first define the graph Laplacian matrix.

Definition 5.1. *Let $G = (V, E)$ be a simple, undirected, weighted graph, where V is the set of n nodes, $E \subseteq V \times V$ is the set of m edges excluding self loops. We assume that $i < j$ for any $(i, j) \in E$.*

Let W be the weight matrix whether

$$W_{ij} = \begin{cases} w_l & \textit{if edge } l = (i,j) \in E \textit{ or } l = (j,i) \in E, \\ 0 & \textit{otherwise.} \end{cases}$$

The graph Laplacian matrix is defined by

$$L = D - W, \tag{5.1}$$

where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Let $w \in \mathbb{R}^m$ be the vector of edge weights, C be a $n \times m$ matrix such that the lth column have all zero entries except

$$C_{il} = 1, \quad C_{jl} = -1, \quad (i,j) \in E, \tag{5.2}$$

i.e. nodes i and j are connected by edge l. The Laplacian matrix can also be written as

$$L = C\, \textit{diag}(w) C^\top, \tag{5.3}$$

where diag(w) is a $m \times m$ diagonal matrix formed from w.

An important property of L is that it is positive semidefinite, where the smallest eigenvalue is $\lambda_n = 0$ [1] . The algebraic multiplicity of the eigenvalue 0 is equal to the number of connected components of the graph. Throughout this chapter, we will assume that the graph is connected, which implies

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{n-1} > \lambda_n = 0.$$

1. We assume larger eigenvalues have smaller subscripts.

The inspiration of our work comes from the line of research for fast solving linear system in the graph Laplacian. There are two major approaches [Spielman, 2010]: *variants of Gaussian elimination* and *iterative methods*. For Gaussian elimination methods, one first uses the *Cholesky decomposition* to factorize L into the form

$$L = ZZ^\top,$$

where Z is a lower triangular matrix and has nonnegative diagonal entries.[2] In particular, one can often permute the rows and columns of L so that the factor Z is very sparse and can be computed in nearly linear time. Popular permutation methods include minimum degree ordering [Tinney and Walker, 1967], Cuthill-McKee reordering [Cuthill and McKee, 1969], approximate minimum degree ordering [Amestoy et al., 1996], etc. Without an appropriate permutation, one might observe the *fill-in* phenomenon: some nonzero entries that are not in L appear in Z. In the sequel, we will assume that such good permutation is known and there is no *fill-in*.

In this chapter, we consider problems of choosing edge weights so that some function of L is minimized. In particular, we look at the problem of constructing the fastest mixing Markov chain for a given graph. We are interested in the following questions:

1. Given the Cholesky factor of a valid Laplacian, does the factor of the optimal Laplacian have the same sparsity pattern, under the same permutation?

2. Can we build a nonconvex reformulation of FMMC, where the variable is the sparse Cholesky factor? Can first order methods successfully find a global minimizer?

3. Will the resulting algorithm have low computational cost?

To answer these questions, the rest of this chapter is organized as follows. Section 5.2 briefly describes the FMMC problem; more details can be found in Boyd et al. [2004]. In Section 5.3,

2. Cholesky decomposition is uniquely defined for positive definite matrix, where the diagonal entries of factors are positive. We can extend it to positive semidefinite matrix by allowing zeros on the diagonal line, but such decomposition might not be unique. For graph Laplacian matrix, one way to obtain a valid Z is to stop the decomposition algorithm when the remaining matrix has dimension 2. In this case, we will have $Z_{nn} = 0$.

we introduce a variant of the ADMM algorithm and analyze its computational cost. Section 5.4 discusses related work. Empirical results are presented in Section 5.5. Conclusions and future research directions are included in Section 5.6.

5.2 Problem Statement

We consider the discrete time Markov chain for sampling the nodes for a given undirected connected graph G. Each edge $l \in E$ is associated with a special edge weight w_l, which is the transition probability between these two nodes. Our weight matrix is the transition matrix $P \in \mathbb{R}^{n \times n}$ where $P_{ij} = P_{ji}$ is the probability of transits between node i and node j. The equilibrium distribution is the uniform distribution $1/n\mathbf{1}$ since P is symmetric.

To ensure the matrix P describes a valid Markov chain defined on the graph, it has to satisfy

(a) the nonnegative constraint $P \geq 0$,

(b) the doubly stochastic constraint $P\mathbf{1} = \mathbf{1}$, $P = P^\top$,

(c) the graph structure constraint $P_{ij} = 0$ if $(i, j) \notin E$.

Let $\pi(t)$ denote the probability distribution of the state at time t. The rate at which $\pi(t)$ converges to the uniform distribution is determined by the *second largest eigenvalue magnitude* (SLEM) of P:

$$\mu(P) = \max_{i>1} |\lambda_i(P)| = \max \{\lambda_2(P),\ -\lambda_n(P)\}. \tag{5.4}$$

The smaller the SLEM, the faster the mixing rate.

We are interested in finding the edge probabilities that give the fastest mixing chain. Boyd et al. [2004] show that the fastest chain and optimal SLEM can be exactly computed by the following

program:

$$\begin{array}{ll} \min\limits_{P} & \mu(P) \\ \text{subject to} & P \geq 0 \\ & P\mathbf{1} = \mathbf{1}, \quad P = P^\top \\ & P_{ij} = 0, \quad \forall i \neq j \text{ such that } (i,j) \notin E \end{array} \tag{5.5}$$

This problem can be formulated as a semidefinite program. One easy way to see this is to note that $\mu(P) = \left\| P - (1/n)\mathbf{1}\mathbf{1}^\top \right\|_2$, so that the minimization of $\mu(P)$ is equivalent to minimizing a constant γ satisfying the constraint $-\gamma I \preceq P - (1/n)\mathbf{1}\mathbf{1}^\top \preceq \gamma I$. As seen in previous chapters, standard interior point solvers for SDPs are not feasible for modern large scale problems. Boyd et al. [2004] give a projected subgradient method that can scale to much larger instances. We shall refer to this algorithm as `subgrad` and mainly compare our approach to it.

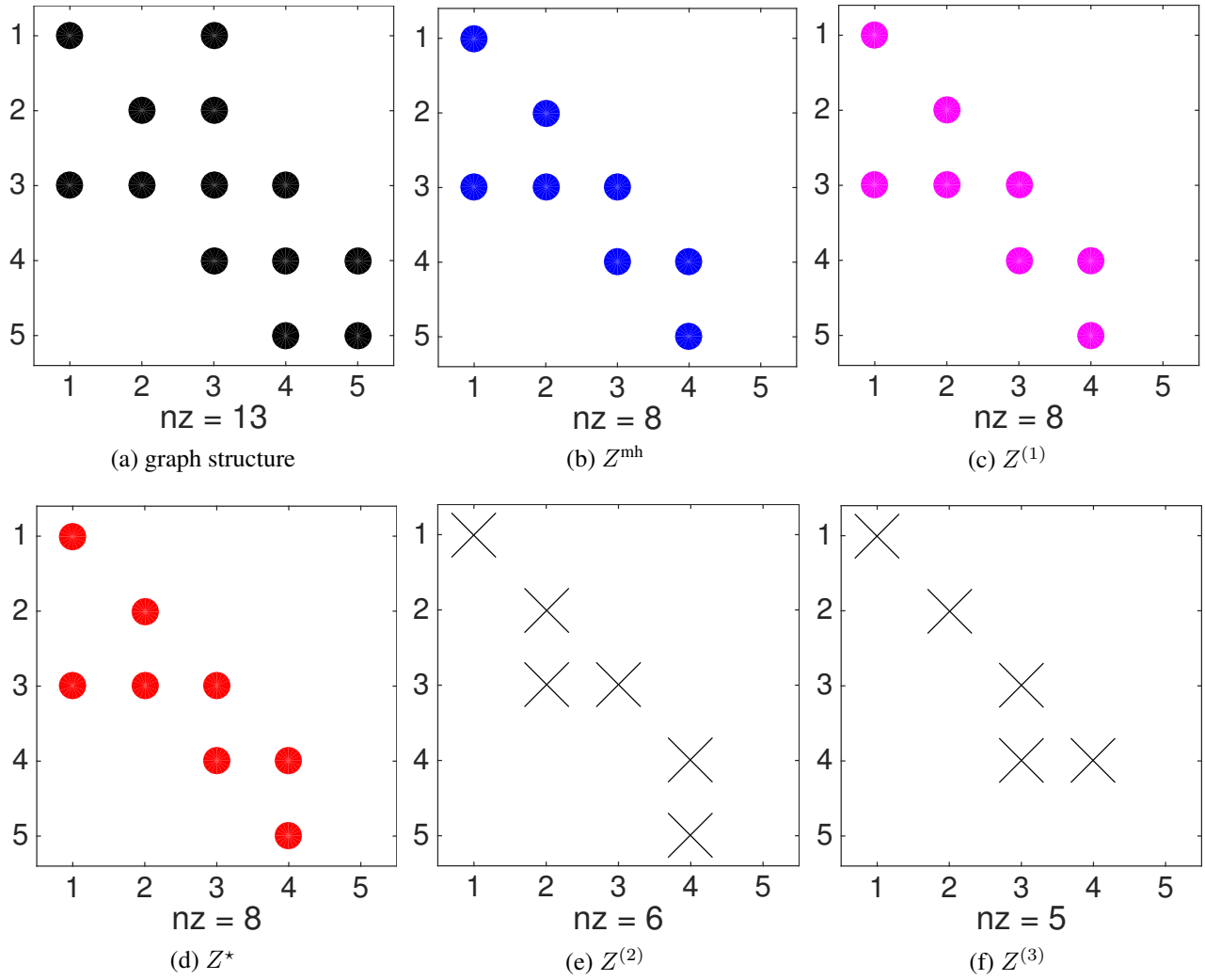
5.3 Nonconvex Formulation and First Order Method

This section presents our approach to FMMC. We first discuss one key observation about the sparsity pattern of Cholesky factor of Laplacian matrices. Based on it, we develop our algorithm and discuss the computational cost.

5.3.1 Sparsity Pattern of Cholesky Factor

Through a sequence of experiments, we have found an interesting phenomenon: given a valid graph Laplacian L of graph G where all the edges in E have positive weights, we compute the Cholesky factor Z of L. Let $\mathcal{S}$ be the set of nonzero entries of Z. Then for any other valid graph Laplacian of the same graph, the nonzero entries of its Cholesky factor are contained in $\mathcal{S}$.

Figure 5.1 illustrates this phenomenon. We randomly generated an Erdős–Rényi graph of 5 nodes, the edge structure (including self-loops) is plotted in (a). Let Z^{mh} be the Cholesky factor of the Laplacian obtained by the *Metropolis-Hastings* algorithm [Metropolis et al., 1953; Hastings, 1970]. The Metropolis-Hastings algorithm is a popular sampling technique. In our context, it gives



(a) graph structure (b) Z^{mh} (c) $Z^{(1)}$

(d) $Z^{\star}$ (e) $Z^{(2)}$ (f) $Z^{(3)}$

Figure 5.1: For a randomly generated Erdős–Rényi graph, the sparsity patterns of adjacency matrix and Cholesky factors of a few graph Laplacian matrices.

all-positive edge weights. Details of this method are explained in Section 5.4. $Z^{(1)}$ is associated with a randomly generated all-positive weighting. $Z^{\star}, Z^{(2)}$ and $Z^{(3)}$ correspond to the fastest chain, and two random weightings where some edges have zero weights, respectively. We can see that Z^{mh} and $Z^{(1)}$ have the same sparsity pattern, which contains all the nonzero entries of $Z^{\star}$, $Z^{(2)}$ and $Z^{(3)}$. Especially, the sparsity pattern of $Z^{\star}$ is the same as Z^{mh}. We have checked the edge weights of the optimal chain and confirmed they are all positive.

This phenomenon is made precise in Theorem 5.1, see Section 5.7 for the proof. Motivated by this observation, we restrict the update to this fixed sparsity pattern. The procedure is described in

the following subsection.

Theorem 5.1. *Let $L = C\,diag(w)C^\top$ be a valid Laplacian matrix of G, where $w > 0$. Let Z be its Cholesky factor and $\mathcal{S}$ be the set of its nonzero entries. Then for any other valid Laplacian of the same graph $L' = C\,diag(w')C = Z'Z'^\top$, we have*

$$supp(Z') \subseteq \mathcal{S}.$$

5.3.2 Algorithm: A Variant of ADMM

We first rewrite Problem 5.5 so that it is parameterized by L. By definition, the Laplacian matrix L is equal to $I - P$. The eigenvalues of P and L have the following relationship:

$$\lambda_i(P) = 1 - \lambda_{n+1-i}, \quad i = 1, \ldots, n,$$

where $\lambda_i(P)$ and λ_i are the ith largest eigenvalues of P and L, respectively. The SLEM of P can also be translated as

$$\bar{\mu}(L) = \mu(P) = \max\left\{1 - \lambda_{n-1},\, \lambda_1 - 1\right\}.$$

The nonnegative, doubly stochastic and graph structure constraints of P are also equivalent to the conditions

- $w \geq 0$,
- $L_{ii} \leq 1, \quad i = 1, \ldots, n.$

The second condition ensures that the probability of staying at the same node is nonnegative. It is equivalent to the expression $Bw \leq 1$, where B is a $n \times m$ matrix such that $B_{il} = 1$ if edge l is incident to node i otherwise 0.

Identifying those relationships, Problem 5.5 can be written as

$$\begin{aligned} \min_{L,\, w} \quad & \bar{\mu}(L) \\ \text{subject to} \quad & L = C\,\text{diag}(w)C^\top \\ & w \geq 0, \quad Bw \leq 1. \end{aligned} \tag{5.6}$$

Next, we apply the symmetric factorization ideas, and leading to

$$\begin{aligned} \min_{w,\, Z} \quad & f(Z) \\ \text{subject to} \quad & ZZ^\top = C\,\text{diag}(w)C^T \\ & w \geq 0,\, Bw \leq 1. \end{aligned} \tag{5.7}$$

where

$$f(Z) = \max\left\{\sigma_1^2 - 1,\ \ 1 - \sigma_{n-1}^2\right\} \tag{5.8}$$

and $\sigma_1 \geq \ldots \geq \sigma_n$ are the singular values of Z. One can see this is still a convex function of Z, but Problem 5.7 is nonconvex because of the quadratic equality constraint.

The crux of solving either Problem 5.6 or Problem 5.7 is how to couple with the equality constraints. That is, how we can ensure that L or $ZZ^\top$ is a valid graph Laplacian. Projected gradient descent methods may not be suitable since direct projection onto the constrained set is difficult. Instead, we use a variant of the *alternating direction method of multipliers* (ADMM). ADMM is an algorithm that originates from Gabay, Mercier, Glowinski and Marrocco in 1970s. It is a combination of *augmented Lagrangian methods* for constrained optimization and *dual decomposition*, and is often well suited for large scale problems [Boyd et al., 2011].

The augmented Lagrangian for Problem 5.7 is

$$\begin{aligned} \mathcal{L}_\rho(Z, w, M) = \quad & f(Z) + \mathbf{1}_{w\geq 0,\, Bw\leq 1}(w) + \langle M, ZZ^\top - C\,\text{diag}(w)C^T\rangle \\ & + \frac{\rho}{2}\left\|ZZ^\top - C\,\text{diag}(w)C^T\right\|_Z^2, \end{aligned} \tag{5.9}$$

where M is the multiplier, and $\mathbf{1}_{\mathcal{C}}(w)$ is an indicator function that equals zero if w satisfies $\mathcal{C}$, and otherwise is infinity. $\mathcal{L}_\rho$ is nonconvex in F but convex in both M and w. Standard ADMM repeats the following updates until convergence:

$$Z^{t+1} = \arg\min_{Z} \mathcal{L}_\rho(Z, w^t, M^t), \tag{5.10}$$

$$w^{t+1} = \arg\min_{w} \mathcal{L}_\rho(Z^{t+1}, w, M^t), \tag{5.11}$$

$$M^{t+1} = M^t + \rho\Big(Z^{t+1}{Z^{t+1}}^\top - C\,\text{diag}(w^{t+1})C^\top\Big), \tag{5.12}$$

$$t \leftarrow t + 1. \tag{5.13}$$

Our algorithm is a variant of this. Instead of solving the nonconvex inner step (5.10), we take one gradient update for Z, and this update only applies to a fixed subset of entries. Details of updating Z and w are explained below.

- **Update** Z

  The function $f(Z)$ is a convex function. According to the derivation in Boyd et al. [2004] and the chain rule [3], if $f(Z) = \sigma_1^2 - 1$, then one subgradient of f at Z is

  $$2uu^\top Z = 2\sigma_1 uv^\top,$$

  where u, v are singular vectors associated with σ_1. Analogously, if $f(Z) = 1 - \sigma_{n-1}^2$ and u, v are singular vectors associated with σ_{n-1}, the matrix

  $$-2uu^\top Z = -2\sigma_{n-1} uv^\top$$

3. https://people.eecs.berkeley.edu/ elghaoui/Teaching/EE227BT/Lectures/lect2_handout.pdf

will be one subgradient. The subdifferential of f at Z is a convex hull of these subgradients:

$$\begin{aligned}\partial f(Z) = \text{conv}\Bigg(&\left\{2\sigma_1 uv^\top \mid u^\top Z v = \sigma_1,\ f(Z) = \sigma_1^2 - 1\right\} \\ &\cup \left\{-2\sigma_{n-1} uv^\top \mid u^\top Z v = -\sigma_{n-1},\ f(Z) = 1 - \sigma_{n-1}^2\right\}\Bigg).\end{aligned} \tag{5.14}$$

Therefore, the subgradient of $\mathcal{L}_\rho(Z, w^t, M^t)$ at Z^t will be

$$g^t + 2\rho\left(Z^t Z^{t\top} - C\,\text{diag}(w^t)C^\top + 1/\rho M^t\right), \quad g^t \in \partial f(Z^t).$$

Hence, in every iteration, to obtain a subgradient, we will compute the top and bottom few singular vectors of Z, or equivalently some eigenvectors of $ZZ^\top$.

We restrict the update to a fixed sparsity pattern $\mathcal{S}$. This will keep the iterates of Z sparse, so that its eigenvalues and eigenvectors can be computed quickly. Empirically, we find that initializing by the Metropolis-Hastings algorithm is effective and robust. Let L^{mh} and Z^{mh} be the Laplacian matrix and its factor corresponding to the output of the Metropolis-Hastings algorithm. We choose $\mathcal{S}$ to be the subset of nonzero entries of Z^{mh}. We shall briefly review this algorithm in Section 5.4.

- **Update** w

  We need to solve

$$\underset{\substack{w\geq 0\\ Bw\leq 1}}{\arg\min}\ h(w) = \frac{1}{2}\left\| \overbrace{ZZ^\top + 1/\rho M}^{\Phi} - C\,\text{diag}(w)C^\top \right\|_F^2 \tag{5.15}$$

  where Z and M are fixed. Let $\Phi = ZZ^\top + 1/\rho M$, we can rewrite

$$\begin{aligned} h(w) &= 2\sum_l (\Phi_{i_{(l)}j_{(l)}} - w_l)^2 + \sum_{i=1}^n \left(\Phi_{ii} - \sum_{l=(i,\cdot)} w_l\right)^2 \\ &= 2\left\|\phi - w\right\|^2 + \left\|\text{diag}(\Phi) - Bw\right\|^2, \end{aligned} \tag{5.16}$$

Algorithm 3: Approximate projection of w

input: w, B

$w \leftarrow \max\{w, 0\}$

for *node* $i = 1, \ldots, n$ **do**

  $\mathcal{I}(i) = \{l \mid l = (i, \cdot)\}$

  while $\sum_{l \in \mathcal{I}(i)} w_l > 1$ **do**

    $\mathcal{I}(i) = \{l \mid l = (i, \cdot), w_l > 0\}$

    $\delta = \min\left\{\min_{l \in \mathcal{I}(i)} w_l,\ \left(\sum_{l \in \mathcal{I}(i)} w_l - 1\right) / |\mathcal{I}(i)|\right\}$

    $w_l = w_l - \delta, \quad l \in \mathcal{I}(i)$

  end

end

where $\left(i_{(l)}, j_{(l)}\right)$ is the subscript of one entry of Φ that corresponds to edge l, and we use ϕ to denote the vector that consists these entries. Note that ϕ can be simply read off from the lower triangular part of Φ. $\sum_{l=(i,\cdot)} w_l$ is the sum of the weights of all edges incident to node i, which is equal to the $\langle B_{(i)}, w\rangle$.

Problem 5.16 can be solved by projected gradient descent, possibly with early termination. The gradient of h at w is

$$\nabla h(w) = 4(w - \phi) + 2(B^\top B w - B^\top \operatorname{diag}(\Phi)). \tag{5.17}$$

We use the approximate method proposed by Boyd et al. [2004] to project w onto the feasible set $\{w | w \geq 0, Bw \leq 1\}$, see Algorithm 3. It first projects w onto the nonnegative orthant, then compute one w satisfying the inequality constraint $Bw \leq 1$ by thresholding.

The whole algorithm is presented in Algorithm 4.

5.3.3 Computational Complexity

It is important to understand the per-iteration computational cost of our algorithm.

Algorithm 4: Nonconvex ADMM variant for FMMC

initialization

$w^0 \leftarrow$ output of Metropolis-Hastings

$Z^0 = \text{Chol}(C\,\text{diag}(w^0)C^\top)$, $M^0 = 0$

$\mathcal{S} \leftarrow$ nonzero entries of F^0

$t \leftarrow 0$, res $\leftarrow 0$

repeat

`// update` F

$\nabla \mathcal{L}_Z \leftarrow g^t + 2\rho\left(\text{res} + M^t/\rho\right) Z^t$, where $g^t \in \partial f(Z^t)$

$Z^{t+1}_{\mathcal{S}} = Z^t_{\mathcal{S}} - \eta^t(\nabla \mathcal{L}_Z)_{\mathcal{S}}$

`// update` w `(Problem 5.16)`

$\Phi \leftarrow Z^{t+1}{Z^{t+1}}^\top + 1/\rho M^t$

$\phi \leftarrow$ (edge weight) entries read off from Φ

$w^{t+1} = w^t$

repeat

$\nabla h = 4(w^{t+1} + \phi) + 2(B^\top B w - B^\top \text{diag}(\Phi))$

$w^{t+1} \leftarrow$ `ApproxProj`$(w^{t+1} - \gamma \nabla h)$

until *convergence*;

`// update` M

res $\leftarrow Z^{t+1}{Z^{t+1}}^\top - C\,\text{diag}(w^{t+1})C^T$

$M^{t+1} = M^t + \rho\,\text{res}$

$t \leftarrow t + 1$

until *convergence*;

First of all, with an appropriate ordering, we would expect L and its Cholesky factor Z to have the same order of number of nonzero entries. For our particular choice of initialization, L^0 is exactly $(m+n)$-sparse, so that the iterates $\left\{Z^t\right\}$ will have $O(m)$ nonzero entries.

We found that the projected gradient descent algorithm for updating w usually converges very quickly in a few iterations. The dominating computational cost is the update of Z^t, for which we need to compute the largest and smallest two singular values and associated left singular vectors of Z^t. One efficient way to obtain them is to compute the left eigenvectors of $Z^t{Z^t}^\top$ using the *Lanczos* method [Lanczos, 1950].

To compute the eigenvalue decomposition of a $n \times n$ matrix, the Lanczos method first rotates the $Z^t {Z^t}^\top$ into a tridiagonal matrix and computes its eigenpairs, then rotates the eigenvectors back. The rotation is performed by generating and applying n Lanczos vectors one by one, called the *Lanczos iterations*. To compute only the top eigenpair accurately, it suffices to run a constant number of Lanczos iterations. Let k denote this number. In $O(mk)$ flops, we obtain a $k \times k$ tridiagonal matrix, as the dominating computation in every iteration is to multiply $Z^t {Z^t}^\top$ with the Lanczos vector. Afterwards, one can use QR decomposition or the MRRR algorithm [Dhillon et al., 2006] to get k eigenvalues and an orthonormal matrix $Q \in \mathbb{R}^{k \times k}$ in $O(k^2)$ flops . The eigenvectors of L are then obtained by multiplying the Lanczos vectors with Q.

The smallest two eigenvectors can be computed in the shift mode: apply the Lanczos method to compute the top two eigenvectors of $(\lambda_1 + \varepsilon)I - Z^t {Z^t}^\top$, where λ_1 is the largest eigenvalue of $Z^t {Z^t}^\top$ we obtained.

Overall, the per-iteration runtime of the nonconvex ADMM is $O(m)$. It should be noted that there is no computational improvement compared to either `subgrad` or ADMM for solving Problem 5.6, since we need to compute the eigenvectors of $Z^t {Z^t}^\top$, which is comparable to computing the eigenvectors of L^t for the other two methods.

5.4 Related Work

Our thinking about the using nonconvex methods to optimize graph Laplacian is inspired by the recent advances in solving linear systems in the Laplacian matrices [Lee et al., 2015; Kyng et al., 2016]. The authors have shown that for every Laplacian matrix L there exists a constant factor approximation $\bar{L}$, whose Cholesky factor $\overline{Z}$ only has $O(n)$ nonzero entries. The sparsified Cholesky factor $\overline{Z}$ can be computed in polynomial time via *spectral vertex sparsification*, a procedure that recursively approximates the Schur complement of subsets of nodes without constructing the Schur complement explicitly. While this is not fully developed in this thesis, our initial thought was to

use $O(n)$ sparse $\overline{Z}$ in the algorithm, either fixing or varying the sparsity pattern in every iteration.[4] Using an $O(n)$ sparse factor will reduce the per-iteration cost for computing the eigenvectors to $O(n)$, if such sparsification could be computed quickly. However, in practice we found that finding the $O(n)$ approximation is hard since the vertex sparsification algorithm is complex. There is a much simpler sparsifier proposed by Kyng and Sachdeva [2016], which is based on purely randomly sampling the edges. Nonetheless it only guarantees that the output Cholesky factor has $O(m \log^3 n)$ nonzero entries. Another crude way to achieve computational efficiency is to randomly sample $O(n)$ edges of the given graph, and use it as a replacement of the original graph. For Erdős–Rényi graphs, the subsampled graph might still be connected, yet this might not be feasible for graphs with more practical structures such as clusters, hubs, etc.

Another research field that led us to the FMMC problem is Markov Chain Monte Carlo sampling (MCMC). MCMC is widely used in many scientific fields to randomly sample from a high dimensional probability distribution. While asymptotically converging to the equilibrium distribution, determining when the chain is close to equilibrium is a challenging open question. As pointed out in Boyd et al. [2004], although the problem of FMMC focuses on finding edge weights rather than sampling itself — we hope that understanding the edge weights of the fastest chain can give insights into how to improve the efficiency of practical MCMC simulations. For solving FMMC itself, our benchmark is the projected gradient descent algorithm `subgrad` proposed by Boyd et al. [2004]. Regarding practical MCMC algorithms, the Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970] is a commonly used random walk sampling technique that works as follows. At current node x, one first samples a node y from a easily-computed proposal distribution $q(\cdot|x)$, then jumps to the candidate y with probability $\min\left\{\pi_y q(x|y)/\pi_x q(y|x), 1\right\}$; otherwise the walk remains at node x. For our problem, $\pi_y/\pi_x = 1$ as the equilibrium distribution is uniform. If we choose the natural proposal distribution $q(\cdot|x) = 1/d_i$, where d_i is the degree of node i, we will have that the transition probability between two connected (different) nodes i and j is equal to $\min\left\{1/d_i, 1/d_j\right\}$. This means the Metropolis-Hastings chain only depends on local information –

4. If such pattern is fixed, we might expect suboptimality of the algorithm.

one can sample the nodes while exploring the graph, without knowing the whole graph structure in advance. Therefore, if the graph structure is known, the Metropolis-Hastings chain can be easily computed in exact $(m+n)$ time. We thus feed the solution of it to our algorithm as an initialization.

5.5 Experiments

This section presents empirical experiments to study the effectiveness of Algorithm 4. We are interested to see

- whether the nonconvex ADMM converges to the global optimum, and
- whether ADMM or nonconvexification can provide a faster convergence rate than the projected subgradient method [Boyd et al., 2004].

Hence, we also considered the convex ADMM variant for solving Problem 5.6. The algorithm is roughly the same as Algorithm 4, except the gradient update of Z is replaced by a gradient update of L. All the methods are implemented in MATLAB and the experiments are conducted on a Macbook Pro with 2.4G HZ CPU and 16GB memory.

5.5.1 Initialization of Nonconvex ADMM

We first inspect the convergence of Algorithm 4. Nonconvex functions might have saddle points and local minima, so that a good initialization is important. Even for strict saddle problems like phrase retrieval and positive semidefinite matrix sensing [Sun et al., 2016; Bhojanapalli et al., 2016b], bad initialization could result in slow convergence. We have found that starting by the following two edge weights are effective and robust:

- the output of Metropolis-Hastings w^{mh}, or
- $w^{\text{unif}} = \texttt{ApproxProj}_{w\geq 0,\, Bw\leq 1}(\mathbf{1})$.

We randomly generated 10 Erdős–Rényi graphs, where the probabilities of connecting every two nodes vary from 0.05 to 0.5. For each graph, we compute the fastest mixing chain, and run

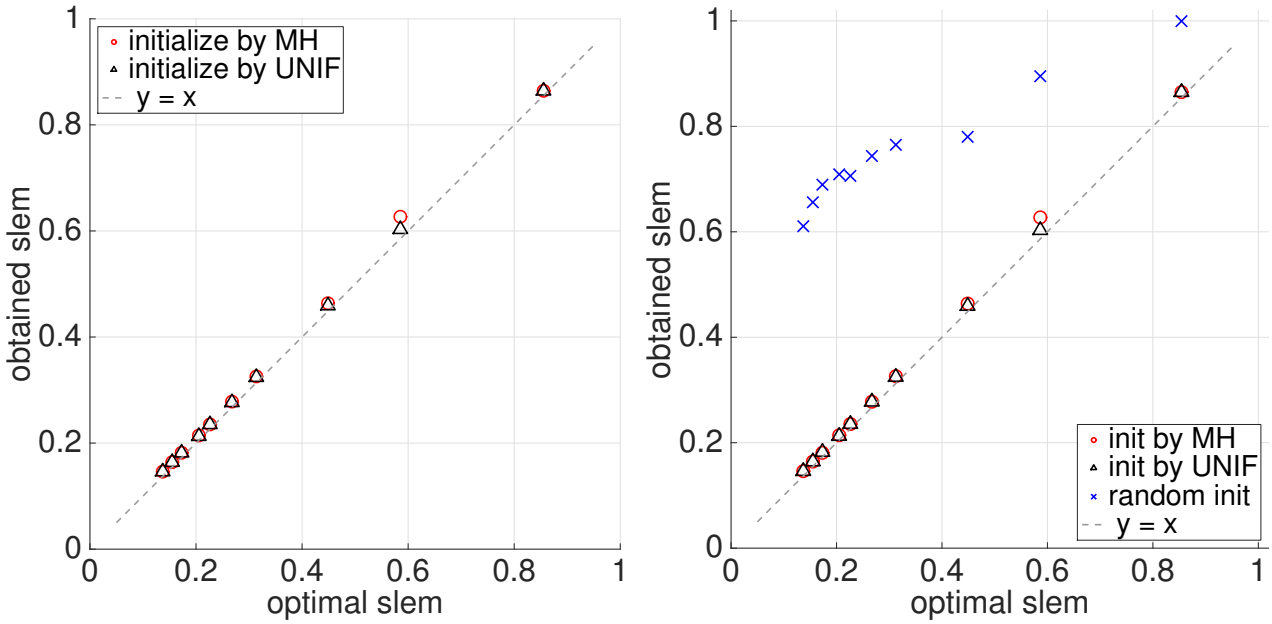
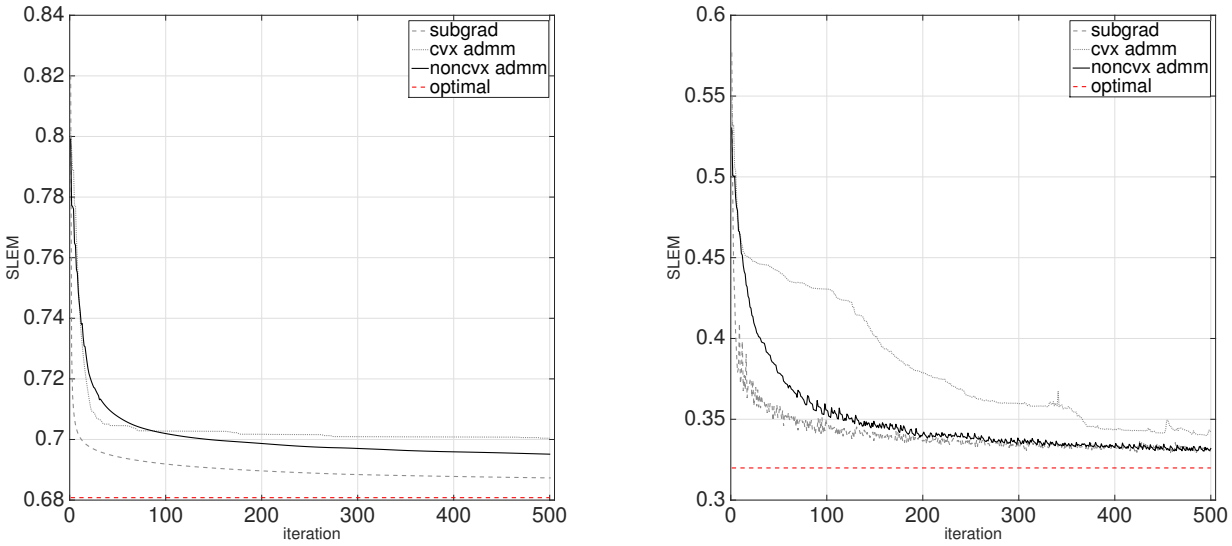


Figure 5.2: (a) Optimal SLEM and the output of our approach using heuristic initialization. (b) Comparing with random initialization.

our approach using the above two initializations. We use diminishing step size $\eta^t = 0.2/\sqrt{(t)}$ and stop the computation after 500 iterations. The penalty parameter ρ for ADMM is fixed to be 1. Figure 5.2(a) reports the SLEM of fastest chain and the SLEMs we obtained. We can see that for both heuristic initializations, the algorithm is converging towards the global optimum.

A natural question to ask is whether random initialization works. If the algorithm under random initialization also converges to the global optimum, it implies that the nonconvex objective might have favorable geometry. For example, the nonconvex objective for positive semidefinite matrix sensing and completion does not have spurious local minima. Hence, we also checked the performance of our approach when it is initialize randomly: we generate a random Gaussian vector $\beta \sim \mathcal{N}(0, I)$ and start from $w^0 = \texttt{ApproxProj}(\beta)$. For this type of initialization, w^0 often contains many zero entries, and the initial Cholesky factor Z^0 is very sparse. If we use the sparsity pattern of Z^0, it is hard to observe convergence. Hence, we feed the algorithm with the sparsity pattern obtained by Metropolis-Hastings, but randomly generated edge weights. We update the estimate for 1000 iterations, and the step size is determined by line search. The result is reported in



(a) $p = 0.1$ (b) $p = 0.2$

Figure 5.3: Convergence comparison for a random Erdős–Rényi graph.

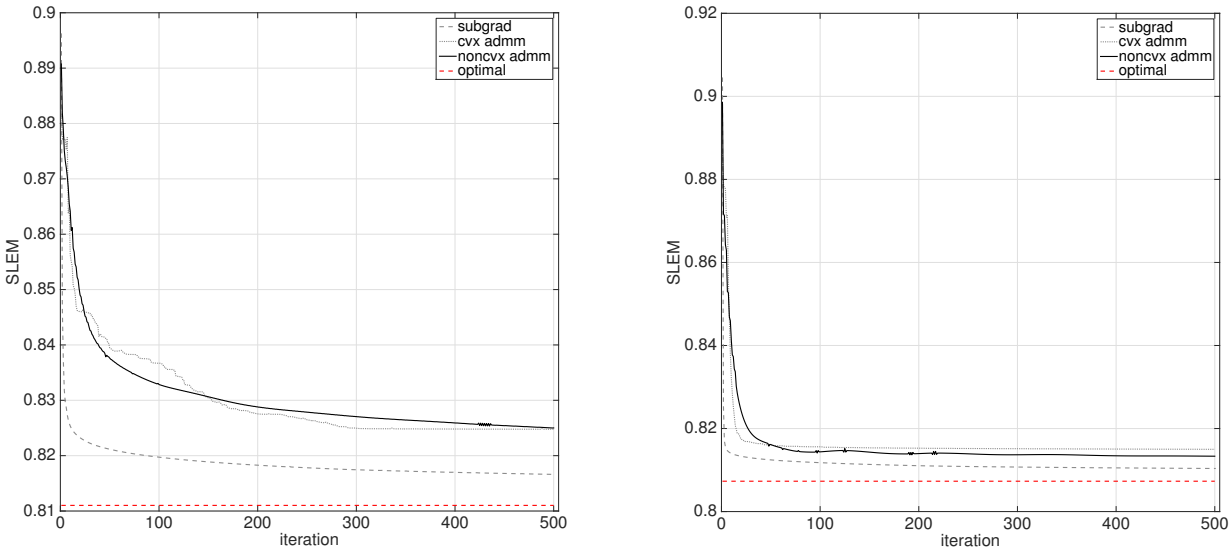
Figure 5.2(b). We have found random initialization performs worse than the other two heuristics.

5.5.2 Comparison with Other Methods

Figure 5.3 shows the convergence speed for different algorithms for two randomly generated Erdős–Rényi graph with 100 nodes. For the first graph, every pair of nodes is connected with probability 0.1. In the second graph, this probability is increased to be 0.2.

Figure 5.4 plots the results for graphs generated from stochastic block model. In the first case, there are two clusters in the graph, each has 50 nodes. The nodes are are randomly connected with probability 0.1 if they are in the same cluster, otherwise with probability 0.02. The between-cluster connection probability has been increased to 0.05 in the second case.

All the methods are initialized by the solution of Metropolis-Hastings, and use the same step size $\eta^t = 0.2/\sqrt{t}$. In all these four cases, `subgrad` converges first. The rate of both convex and nonconvex ADMM are comparable.



(a) $p_{\text{in}} = 0.1$, $p_{\text{out}} = 0.02$ (b) $p_{\text{in}} = 0.1$, $p_{\text{out}} = 0.05$

Figure 5.4: Convergence comparison for a random graph in stochastic block model.

5.6 Discussion

We propose a nonconvex formulation based on Cholesky factorization to compute the fastest mixing Markov chain on a given graph. We develop a variant of the ADMM algorithm for optimizing the objective. Empirical results suggest that this algorithm may converge to a global minimizer, with comparable rate and computational cost to `subgrad` and convex ADMM.

Unlike the results in previous chapters, the nonconvexified objective did not lead to faster convergence or reduced computational Complexity. The computational cost is not reduced since we did not avoid computing the eigenvectors. Further study needs to be carried out to provide comprehensive understanding of both nonconvex ADMM and the FMMC problem.

For example, an alternative way to write SLEM as a function of L is $\max_{\|u\|=1, u'\mathbf{1}=0} |u^\top L u - 1|$. This leads us to another nonconvex min-max formulation:

$$\min_{w \geq 0, Bw \leq 1} \max_{\|u\|=1, u'\mathbf{1}=0} (u^\top L u - 1)^2, \tag{5.18}$$

where we square the absolute value to make the loss function differentiable. This loss function is fourth order in terms of u and quadratic in terms of w. Simple and efficient algorithms that can

solve this nonconvex problem, are worth exploring.

5.7 Proofs

5.7.1 Proof of Theorem 5.1

The proof has two steps. We first explicitly identify the columns of the Cholesky factor, then discuss its sparsity pattern.

Step 1.

The operations to compute the Cholesky factorization of L can be considered as sequential elimination of the nodes of G. Let us define a chain of Schur complements as in Kyng and Sachdeva [2016].

Let $L = S^{(0)}$ be the Laplacian matrix of graph $G = (V, E)$. Let M_i denote the i column of a matrix M. In the k-th iteration, we eliminate node k and define

$$S^{(k)} = S^{(k-1)} - \frac{1}{S_{kk}^{(k-1)}} S_k^{(k-1)} S_k^{(k-1)^T}, \tag{5.19}$$

which is the Schur complement of $S^{(k-1)}$ respect to column k. We can see that the first k columns and rows of $S^{(k)}$ are all zeros. Mathematically, if $S_{kk}^{(k-1)} = 0$, we set $S^{(k)} = S^{(k-1)}$. Finally, we will end this sequence with a zero matrix $S^{(n)}$.

Now let us define $d_k = S_{kk}^{(k-1)}$ and $f_k = 1/d_k \cdot S_k^{(k-1)}$. Again, if $S_{kk}^{(k-1)} = 0$, we shall set them zero as well. It holds that

$$S^{(k)} = S^{(k-1)} - d_k f_k f_k^\top.$$

Observing $S^{(n)} = 0$ and $L = S^{(0)}$, we have

$$L = \sum_{k=1}^{n} S^{(k-1)} - S^{(k)} = \sum_{k=1}^{n} d_k f_k f_k^\top = F \,\mathrm{diag}(d) F^\top, \tag{5.20}$$

where f_k is the kth column of F. F is a lower-triangular matrix, since f_k inherits its sparsity pattern from $S_k^{(k-1)}$, and the first $k-1$ rows and columns of $S^{(k-1)}$ are all zero. Consequently, $F\,\mathrm{diag}(d)^{1/2}$ is the Cholesky factor of L.

Step 2.

It remains to show that the sparsity pattern of f_k is contained in a certain fixed set, when the initial edge weights are all positive. As the sparsity patterns of f_k and the kth column of $S^{(k-1)}$ are the same, it suffices to show that every Schur complement matrix $S^{(k)}$ (except the last one) in this chain is a Laplacian matrix of a fixed graph $G^{(k)}$, and all the edge weights are positive.

We prove this by induction. The base case is $f_1 = S_1^{(0)}$. This is trivial by our assumption.

Suppose that $S^{(k-1)}$ is the Laplacian of graph $G^{(k-1)}$ with positive weights:

$$(S^{(k-1)})_k = \sum_{l \text{ in } G^{(k-1)}} w_l C_l C_l^\top, \quad w > 0. \tag{5.21}$$

Define

$$(S^{(k-1)})_k = \sum_{l \text{ incident to } k \text{ in } G^{(k-1)}} w_l C_l C_l^\top, \tag{5.22}$$

where C is the incidence matrix defined in Definition 5.1. We can rewrite

$$S^{(k)} = \overbrace{S^{(k-1)} - (S^{(k-1)})_k}^{S_{-k}^{(k-1)}} + \overbrace{(S^{(k-1)})_k - \frac{1}{S_{kk}^{(k-1)}} S_k^{(k-1)} S_k^{(k-1)^T}}^{C^{(k)}}. \tag{5.23}$$

Clearly,

$$S_{-k}^{(k-1)} = \sum_{l \text{ not incident to } k \text{ in } G^{(k-1)}} w_l C_l C_l^\top,$$

this is a Laplacian matrix of a new graph formed by eliminating node k and associated edges from $G^{(k-1)}$, where the remaining edges still have unchanged positive weights. A well known fact is that $C^{(k)}$ is a Laplacian of the clique that contains all the neighbors of node k in $G^{(k-1)}$, see Lemma 5.1. In particular, from Equation (5.27), we can see all edges in this clique have positive weights.

Hence, $S^{(k)}$ is the Laplacian for graph $G^{(k)}$, which is formed by combining these two parts. It contains nodes $k, \ldots, n$; the edges are the union of the remaining edges and the clique, where overlapping edges have weights added up. Hence, the associated weights are all positive.

Lemma 5.1. *$C^{(k)}$ is a Laplacian of a clique formed by the neighbors of node k in $G^{(k-1)}$.*

Proof. Without lose of generality, let us assume $k = 1$ and denote the first row of $S^{(0)}$ by $[d, -a^\top]$. The vector a consists of the weights of edges incident to node 1, and $d = a^\top \mathbf{1}$.

We then have

$$\begin{aligned} C^{(1)} &= \sum_{l \text{ incident to node 1}} w_l c_l c_l^\top - \begin{bmatrix} d & -a^\top \\ -a & (1/d)aa^\top \end{bmatrix} \\ &= \begin{bmatrix} d & -a^\top \\ -a & \text{diag}(a) \end{bmatrix} - \begin{bmatrix} d & -a^\top \\ -a & (1/d)aa^\top \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0^\top \\ 0 & \text{diag}(a) - (1/d)aa^\top \end{bmatrix}. \end{aligned} \tag{5.24}$$

Note that

$$\begin{bmatrix} 0 & 0^\top \\ 0 & aa^\top \end{bmatrix} = \sum_{i \text{ incident to 1}} \sum_{j \text{ incident to 1}} w_{(1,i)} w_{(1,j)} c_{(i,j)} c_{(i,j)}^\top, \tag{5.25}$$

where $c_{(i,j)}$ is a vector whose ith entry is 1 and jth entry is -1. Besides, we also have $d = \sum_{i \text{ incident to 1}} w_{(1,i)}$, hence the diagonal entries are

$$C_{ii}^{(1)} = w_{(1,i)} - \frac{w_{(1,i)}^2}{\sum_j w_{(1,j)}} = \frac{\sum_{j \neq i} w_{(1,i)} w_{(1,j)}}{\sum_j w_{(1,j)}}, \quad i = 1, \ldots n. \tag{5.26}$$

To see the off-diagonal entries, we have

$$C^{(1)} - \text{diag}(C^{(1)}) = \sum_{i \text{ incident to 1}} \sum_{\substack{j \text{ incident to 1} \\ j \neq i}} -\frac{w_{(1,i)} w_{(1,j)}}{\sum_j w_{(1,j)}} c_{(i,j)} c_{(i,j)}^\top. \tag{5.27}$$

It follows immediately that $C^{(1)}$ is the Laplacian for the clique of neighbors of node 1 in $G^{(0)}$, where the edge weight for (i, j) is $w_{(1,i)}w_{(1,j)}/\sum_j w_{(1,j)}$. $\square$

Part III

Conclusion, Extensions and Future Work

CHAPTER 6

CONCLUSION

6.1 Summary

This thesis studies a new framework for optimizing semidefinite variables. We decompose the target semidefinite variables into symmetric factors, and reformulate the problem so as to optimize the factor. Depending on the nature of the problem, the structure of the factor, such as the low-rank property or sparsity, are utilized to reduce the number of parameters and computational cost.

The first direct application of our technique is semidefinite programming. While SDP is usually used as surrogate relaxations of difficult nonconvex problems, the newly proposed methods in this thesis approach SDPs via nonconvexifying. When the factor is of low rank, same factorization idea was proposed by Burer and Monteiro [2003]. We have shown that these simple methods are remarkably effective for several problems of practice interests, with analytical convergence guarantees and strong empirical performance. These algorithms are also fast, scalable, and easy to implement, and hence are well suited for very large scale problems. We emphasize that our technique is not limited to convex problems; the reformulation could be helpful for nonconvex problems too.

6.2 Future Work

A contribution of this work is to indicate that the road between nonconvex and convex approaches is in fact bidirectional. There might be other classes of problems for which nonconvex recasting could be helpful. While the power and limits of convex approximation are both analyzed and demonstrated for many problems, theoretical understanding of many nonconvex problems is nascent. For example, though empirical results show that the nonconvex ADMM in Chapter 5 may converge to a global minimizer reliably, it still lacks theoretical support. The current sample complexity obtained for both matrix sensing (Chapter 3) and completion (Chapter 4) are still sub-

optimal. In contrast, the nuclear norm relaxation achieves the information theoretically optimal bound. The transformation between two types of methods, and the potential trade-off between statistical and computational properties, deserves further study.

The convergence result we obtained in this thesis reveals two distinctive features for the nonconvex objectives we consider in Chapter 3 and 4. First, for a local region near the global optimum, the functions are essentially convex. Second, spectral initialization leads to a very good starting point located in that well-behaved area. However, our work is not the final word on this subject. Many recent advances have been made in understanding the geometry of nonconvex objectives. Recent studies have shown that many nonconvex functions do not have spurious local minima, and around each saddle point or local maximizer, these functions always have a negative direction of curvature. Therefore, popular optimization algorithms such as stochastic gradient descent and trust region methods can provably converge to a global minimizer with arbitrary initialization in polynomial time [Sun et al., 2015]. Examples includes phase retrieval [Sun et al., 2016], dictionary leanring [Sun et al., 2017], low rank positive semidefinite matrix sensing [Bhojanapalli et al., 2016b] and completion [Ge et al., 2016], and orthogonal fourth order tensor decomposition [Ge et al., 2015].

For standard semidefinite programming with m constraints, Boumal et al. [2016] studies the geometry of the rank r Burer-Monteiro reformulation. The authors show that when the search space of SDP is compact, and the search space of the reformulated problem is a smooth manifold, and one takes r large enough so that $r(r+1) > 2m$, then for almost all cost matrices, spurious local minima do not exist.

On the other hand, another line of interesting research considers measuring the suboptimality of spurious local minima. Montanari [2016] and Mei et al. [2017] have studied the rank-constrained version of SDPs arising in MaxCut and in synchronization problems. The authors have established Grothendieck-type inequalities that prove all the local maxima and dangerous saddle points are within a small multiplicative gap from the global maximum.

To conclude, it would be a fruitful to understand the geometry of more nonconvex objectives

of interest, and more broadly, the landscape of nonconvex optimization.

REFERENCES

Amestoy, P. R., Davis, T. A., and Duff, I. S. (1996). An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications*, 17(4):886–905.

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5):2877–2921.

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*.

Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. (2016a). Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016b). Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881.

Boumal, N. and Absil, P.-a. (2011). Rtrmc: A riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pages 406–414.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. (2014). Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(Apr):1455–1459.

Boumal, N., Voroninski, V., and Bandeira, A. (2016). The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765.

Boyd, S. (2006). Convex optimization of graph laplacian eigenvalues. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1311–1319.

Boyd, S., Diaconis, P., and Xiao, L. (2004). Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Burer, S. and Monteiro, R. D. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357.

Burer, S. and Monteiro, R. D. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Candès, E., Strohmer, T., and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274.

Candès, E. J. (2006). Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain.

Candès, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2015a). Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251.

Candès, E. J. and Li, X. (2014). Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval via Wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849.

Chen, Y. (2015). Incoherence-optimal matrix completion. *Information Theory, IEEE Transactions on*, 61(5):2909–2923.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv:1509.03025.

Cohen, M. B., Kyng, R., Miller, G. L., Pachocki, J. W., Peng, R., Rao, A. B., and Xu, S. C. (2014). Solving sdd linear systems in nearly m log 1/2 n time. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 343–352. ACM.

Cuthill, E. and McKee, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. (2004). A direct formulation for sparse PCA using semidefinite programming. In *S. Thrun, L. Saul, and B. Schoelkopf (Eds.), Advances in Neural Information Processing Systems (NIPS)*.

Dhillon, I. S., Parlett, B. N., and Vömel, C. (2006). The design and implementation of the mrrr algorithm. *ACM Transactions on Mathematical Software (TOMS)*, 32(4):533–560.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Fazel, M. (2002). Matrix rank minimization with applications. Technical report, Elec. Eng. Dept., Stanford University. PhD thesis.

Feige, U. and Ofek, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275.

Foygel, R. and Srebro, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. arXiv:1102.3923.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842.

Ge, R., Lee, J. D., and Ma, T. (2016). Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981.

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *FOCS 2014*. IEEE.

Hardt, M. and Wootters, M. (2014). Fast matrix completion without the condition number. In *COLT 2014*, pages 638–678.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hoffman, M., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14.

Jain, P., Meka, R., and Dhillon, I. S. (2010). Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM.

Keshavan, R. H. (2012). *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998.

Koutis, I., Miller, G. L., and Peng, R. (2010). Approaching optimality for solving sdd linear systems. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 235–244. IEEE.

Kyng, R., Lee, Y. T., Peng, R., Sachdeva, S., and Spielman, D. A. (2016). Sparsified cholesky and multigrid solvers for connection laplacians. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 842–850. ACM.

Kyng, R. and Sachdeva, S. (2016). Approximate gaussian elimination for laplacians-fast, sparse, and simple. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 573–582. IEEE.

Lanczos, C. (1950). *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Ledoux, M. and Rider, B. (2010). Small deviations for beta ensembles. *Electron. J. Probab.*, 15:no. 41, 1319–1343.

Lee, J. D., Recht, B., Srebro, N., Tropp, J., and Salakhutdinov, R. R. (2010). Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305.

Lee, Y. T., Peng, R., and Spielman, D. A. (2015). Sparsified cholesky solvers for sdd linear systems. *arXiv preprint arXiv:1506.08204*.

Mei, S., Misiakiewicz, T., Montanari, A., and Oliveira, R. I. (2017). Solving sdps for synchronization and maxcut problems via the grothendieck inequality. *arXiv preprint arXiv:1703.08729*.

Meka, R., Jain, P., Caramanis, C., and Dhillon, I. S. (2008). Rank minimization via online learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 656–663. ACM.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Mishra, B., Meyer, G., Bach, F., and Sepulchre, R. (2013). Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149.

Montanari, A. (2016). A grothendieck-type inequality for local maxima. *arXiv preprint arXiv:1603.04064*.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697.

Nesterov, Y. (2004). *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media.

Netrapalli, P., Jain, P., and Sanghavi, S. (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Spielman, D. A. (2010). Algorithms, graph theory, and linear equations in laplacian matrices. In *Proceedings of the international congress of mathematicians*, volume 4, pages 2698–2722.

Spielman, D. A. and Teng, S.-H. (2004). Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90. ACM.

Srebro, N., Rennie, J., and Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336.

Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory*, pages 545–560. Springer.

Sun, J., Qu, Q., and Wright, J. (2015). When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*.

Sun, J., Qu, Q., and Wright, J. (2016). A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE.

Sun, J., Qu, Q., and Wright, J. (2017). Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884.

Sun, R. (2015). *Matrix Completion via Nonconvex Factorization: Algorithms and Theory*. PhD thesis, University of Minnesota Twin Cities.

Sun, R. and Luo, Z.-Q. (2015). Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science, IEEE 56th Annual Symposium on*, pages 270–289.

Tinney, W. F. and Walker, J. W. (1967). Direct solutions of sparse network equations by optimally ordered triangular factorization. *Proceedings of the IEEE*, 55(11):1801–1809.

Tomioka, R., Hayashi, K., and Kashima, H. (2010). Estimation of low-rank tensors via convex optimization. arXiv:1010.0789.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016). Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*.

Vandereycken, B. (2013). Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236.

Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. (2016). Guarantees of riemannian optimization for low rank matrix completion. arXiv:1603.06610.

Yi, X., Park, D., Chen, Y., and Caramanis, C. (2016). Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*.

Zhao, T., Wang, Z., and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567.

Zheng, Q. and Lafferty, J. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117.

Zheng, Q. and Lafferty, J. (2016). Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*.