Matrix Completion via Nonconvex Factorization: Algorithms and Theory

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Ruoyu Sun

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor OF Philosophy

Advisor: Zhi-Quan Luo

May, 2015

Acknowledgements

First and foremost, I would like to gratefully and sincerely thank my advisor Prof. Zhi-Quan (Tom) Luo for his continued support and guidance. He has spent hundreds of hours discussing with me how to formulate and solve various research problems. He can always ask insightful and fundamental questions that force me to think more deeply about the problem and inspire me to find solutions to the problem. He has also spent a tremendous amount of time in improving my presentation and writing skills. He has taught me, either directly or indirectly, many more things than I could have expected from an advisor. I aspire to one day be as good an academic advisor as him. Special thanks to Yanling Huang, wife of Prof. Luo, for being an amazing host during the Thanksgiving/Spring Festival dinners and the biking-BBQ days.

I would like to thank Prof. Mostafa Kaveh, Prof. Nikos Sidiropoulos, Prof. Shuzhong Zhang and Prof. Jarvis Haupt for serving on my doctoral committee. Special thanks to Prof. Nikos Sidiropoulos for providing novel perspectives on the interference alignment problem I have been working on for years. Special thanks to Prof. Shuzhong Zhang for teaching me the beauty of conic optimization and many other topics of optimization. I would like to thank all the professors who taught me so much throughout my graduate career, especially Prof. Ahmed Tewfik and Prof. Nihar Jindal who taught me various courses related to signal processing and wireless communications. I would like to thank Prof. Arindam Banerjee for organizing a great seminar on high dimensional machine learning and allowing me to attend his group meetings and other discussions.

Special thanks are due to Prof. Yinyu Ye for giving me the opportunity to visit his research group at Stanford University; his vision and enthusiasm has inspired me a lot. I would like to thank Prof. Percy Liang for the reading group and ideas dinners, both of which I have enjoyed so much. Thanks are also due to Hadi Baligh for the guidance and

Philippe Sartori, Weimin Xiao and Bingyu Qu for many helps during my internships at Huawei Technologies Chicago office. I would also like to thank Prof. Hongwei Liu and Prof. Bo Chen for interesting discussions on real-world problems and their support during my visits at Xidian University.

I would like to thank all my colleagues in OPSAC group. Specifically, I would like to thank Mingyi Hong, Enbing Song, Bo Jiang, Meisam Razaviyayn, Andy Tseng, Maziar Sanjabi, Wei-Cheng Liao, Shu-Hsien Zhu, Mojtaba Kadkhodaie, Yafeng Liu, Qingjiang Shi, Xiangfeng Wang, Quanbo Ge, Jingran Lin, Hongbing Qiu, Yongchao Wang, Zhibin Zhu, Haibin Zhang, Yanjun Wang, Qiang Li, Shu Cai, Zi Xu, Zhenhua Xu, Lei Jiao, Jin Fu, Wei Liu, Qian Zhang, Yuan Yuan for their company and help throughout my stay. Special thanks are due to Mingyi Hong for being a great collaborator and friend, and Andy (Hung-Wei) Tseng for lots of hours of inspiring discussions on various research subjects. I would also like to thank my collegues in Electrical and Computer Engineering Department and the Digital Technology Center (DTC) who have enriched my graduate life and helped me in many ways: Hao Zhu, Jimeng Zheng, Wentao Shi, Yu Zhang, Vasilis Kekatos, Seung-Jun Kim, Xiao Fu, Kejun Huang, Aritra Konar, Balasubramanian Gopalakrishnan, Pingqiang Zhou, Kaishi Zhang, Xiaofan Wu, Yinglong Feng, Feilong Liu and many others. Special thanks are due to my soccer team members and many other great friends; they make my stay most enjoyable.

Most importantly, I would like to thank my parents Bingzhao Sun and Yinlan Deng for taking pride in me and unconditional love. None of this would have been possible without their support.

Abstract

Learning low-rank structure of the data matrix is a powerful method to deal with "big data". However, in many modern applications such as recommendation systems and sensor localization, it is impossible or too costly to obtain all data, resulting in a data matrix with most entries missing. A problem of great interest is then to infer the missing data based on very few observations, usually under the assumption that the true data matrix is low rank, which is widely known as the matrix completion problem. The most popular solution for large-scale matrix completion is the matrix factorization (MF) formulation, which has achieved great empirical success. However, due to the non-convexity caused by the factorization model, little is known about whether and when the algorithms for the MF formulation will generate a good solution. In this thesis, we will study the non-convex MF formulation for matrix completion, both algorithmically and theoretically.

First, we empirically analyze several standard algorithms for the MF formulation. We present a novel finding that the recovery ability of an algorithm mainly depends on whether it can control the row-norms (or incoherence constants) of the iterates. Motivated by this finding, we propose a new formulation that either adds constraints or penalty-function-type regularizers to directly control the row-norms. Simulation results show that the algorithms for the new formulation can recover the matrix in the regime very close to the fundamental limit, outperforming the standard algorithms.

We then establish a theoretical guarantee for the new MF formulation to correctly recover the underlying low-rank matrix. In particular, we show that under similar conditions to those in previous works, many standard optimization algorithms converge to the global optima of the new MF formulation, and recover the true low-rank matrix. This result is rather surprising since we prove convergence to global optima for a nonconvex problem. To the best of our knowledge, our result is the first one that provides exact recovery guarantee for many standard algorithms. A major difference of our work from the existing results is that we do not need resampling (i.e., using independent samples at each iteration) in either the algorithm or its analysis. Technically, we develop a novel perturbation analysis for matrix factorization which may be of independent interest.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The low-rank matrix completion problem, in which one aims to recover an unknown low rank matrix from only a subset of its entries, has attracted a significant amount of attention in recent years. At the most obvious level, its popularity can be largely attributed to its application in recommendation systems (e.g. Netflix competition) [1] and its connection to compressive sensing [2–5]. Nevertheless, there are many more reasons for its importance.

First, matrix completion is a common problem in many areas because collecting all data is usually impossible or too expensive in modern data applications, resulting in a data matrix with most entries missing. Classical applications of matrix completion include collaborative filtering in recommender systems [1], global positioning in sensor networks [6–9], phase retrieval [10], system identification in control [11], structure-from-motion in computer vision [12,13], multi-class learning in machine learning [14,15], DOA estimation in radar signal processing [16]. In recent few years, many new applications of matrix completions are emerging, such as index coding in information theory [17], state covariance estimation in control [18–20], genomic data integration [21] and robust spectral compressive sensing [22]. More discussions on the applications of matrix completion can be found at [23, 24].

Second, the low-rank model is ubiquitous in modern data applications since it extracts useful information from "big data". Analyzing the low-rank matrix completion

problem, a prototype example of the low-rank model, may help us understand many other low-rank models, including PCA (Principle Component Analysis), robust PCA [25], NMF (Non-negative matrix factorization) [26] and tensor completion [27, 28], to name a few.

Third, matrix completion has a rich theory: it is not only a natural extension of compressive sensing to the matrix domain, but also related to many theoretical fields including linear algebra, optimization, information theory, probability, graph theory and algebraic geometry.

Despite the extensive research, there are still many important open questions for matrix completion, among which one of the biggest puzzles is the large gap between practice and theory in algorithm design. In particular, the classical theory [2–5] is based on the nuclear norm based formulation, while in big data applications a different formulation based on matrix factorization (MF) has been dominant [1]. Rather surprisingly, algorithms for the non-convex MF formulation can exactly recover the unknown low-rank matrix in many numerical experiments, but we have little understanding of this phenomenon. From an optimization point of view, it is a fascinating question why a non-convex optimization problem can be solved to global optima. In this thesis, we will try to answer this question.

1.2 Relation with Compressive Sensing

We first discuss the connection of low-rank matrix recovery (a more general problem than matrix completion) and compressive sensing. In compressive sensing, one aims to recover a sparse vector $x \in \mathbb{R}^{n\times 1}$ based on linear observations $y = Ax$, where $A \in \mathbb{R}^{K\times n}$ is a matrix. In low-rank matrix recovery, one aims to recover a low-rank matrix M based on linear observations $\mathcal{A}(M)$, where $\mathcal{A} : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^K$ is a linear map (can be viewed as a 3-way tensor in $\mathbb{R}^{K\times m\times n}$). It is shown in [29] [30] that if A (resp. $\mathcal{A}$) satisfies RIP (Restricted Isometry Property), then the groundtruth x (resp. M) can be recovered by convex optimization. If the entries of A (resp. $\mathcal{A}$, viewed as a tensor) are drawn independently from several common random distributions, then A (resp. $\mathcal{A}$) does satisfy RIP.

It seems that matrix completion can be viewed as a special case of the low-rank

matrix recovery problem where the linear map $\mathcal{A}$ corresponds to the sampling operator $\mathcal{P}_\Omega$. However, this tensor $\mathcal{A} = (\mathcal{A}_{kij}) \in \mathbb{R}^{K\times m\times n}$ is rather unusual: it has only one non-zero entry 1 in each "slice" (the k-th slice is a $m\times n$ matrix $\mathcal{A}_{k,:,:}$). It seems very difficult to translate a random model of Ω into a tractable random model of $\mathcal{A}$. Consider two popular random models of Ω: the Bernolli model where where each element of $[m]\times[n]$ is included in Ω with probability p, and the uniform model where Ω is generated uniformly at random from the collection of all subsets of $[m] \times [n]$ with a fixed size K (here $[n]$ represents the set $\{1, \ldots, n\}$). In the Bernolli model, the size of Ω is not even fixed, thus it cannot be translated to any random model of $\mathcal{A}$ with fixed size. The uniform model of Ω can be translated to a uniform model of $\mathcal{A}$ where $\mathcal{A}$ is drawn from the collection of $\mathcal{A}$ with only one non-zero entry 1 in each "slice" and the positions of 1's in each slice are distinct. In this random model, the entries of $\mathcal{A}$ are not independent, making it highly challenging to analyze (we remark that even for compressive sensing, it is quite difficult to analyze the case that A has dependent random entries; see, e.g. [31]). Since a random sampling operator can hardly be translated to a tractable random tensor $\mathcal{A}$, the results of [30] that prove RIP of a random tensor $\mathcal{A}$ (more precisely, the paper considered a matrix form) is barely useful for the matrix completion problem. Therefore, we should treat matrix completion and low-rank matrix recovery discussed in [30] (can be called matrix sensing) as two different problems.

We have argued that RIP for the tensor-form of the linear map $\mathcal{A}$ in matrix completion is difficult to prove. One may ask whether it is possible to prove the RIP-type condition for the random sampling operator itself, i.e.

$$C_1\|Z\|_F^2 \le \frac{1}{p}\|\mathcal{P}_\Omega(Z)\|_F^2 \le C_2\|Z\|_F^2, \forall\ Z \text{ satisfying } \operatorname{rank}(Z) \le r', \tag{1.1}$$

where $r' \ge r$ is a certain integer. Unfortunately, this RIP-type condition does not hold for randomly generated Ω. Note that the order of Z and Ω is crucial: first randomly generate Ω and fix this Ω, then pick any Z to check whether (1.1) holds. If there exists one Z that it fails, then this RIP condition does not hold. Such a counterexample to (1.1) is given in [23]: for fixed Ω, pick $(i_0, j_0) \notin \Omega$ and define a matrix Z as $Z_{i_0,j_0} = 1$ and $Z_{i,j} = 0, \forall (i, j) \ne (i_0, j_0)$, then $\|\mathcal{P}_\Omega(Z)\|_F = 0, \|Z\|_F = 1$, which contradicts the first inequality of (1.1). That being said, variants of (1.1), not for the set of all low-rank matrices but for different sets, play an important role in the theory of matrix

completion. Proving (1.1) (especially the first inequality) is one of the main difficulties for guaranteed matrix completion in different scenarios [32–34] (note [34] contains part of this thesis).

1.3 Identifiability

Before discussing the mathematical formulation, the very first question is whether it is even possible to recover a rank-r matrix M from partial observations via any method. This is the identifiability (or recoverability) question: can a low-rank M be uniquely determined by its submatrix $\mathcal{P}_\Omega(M)$? Or equivalently, does the following relation

$$\mathcal{P}_\Omega(M') = \mathcal{P}_\Omega(M) \text{ and } \operatorname{rank}(M) \leq r, \operatorname{rank}(M') \leq r \Longrightarrow M' = M \tag{1.2}$$

hold? Here $\mathcal{P}_\Omega(M)$ denotes the matrix Z such that $Z_{ij} = M_{ij}, \forall (i,j) \in \Omega$ and $Z_{ij} = 0$ otherwise. If there are two different matrices M and M' such that $M_{ij} = M'_{ij}, \forall (i,j) \in \Omega$, it seems impossible to recover M: even if we find a rank-r matrix that matches the observation, this matrix might be M', not the original one M. The strongest version of the identifiability question is:

$$\text{What is the condition on } M \text{ and } \Omega \text{ so that (1.2) holds?} \tag{1.3}$$

This question seems too difficult to answer, and attempts have been made to answer its weaker versions.

A series of papers [2–5] proved that when M and Ω are both "generic" (M is incoherent, a notion we will define later, and Ω is randomly generated) and $|\Omega|$ is large enough, then M can be exactly recovered with high probability. Note that the recovery results of [2–5] do not imply the identifiability of a matrix from partial observations; in contrast, they only implies the identifiability of an *incoherent* matrix from partial observations. In other words, the results in [2–5] do not exclude the possibility that there is another rank-r matrix $M' \neq M$ which satisfies $\mathcal{P}_\Omega(M') = \mathcal{P}_\Omega(M)$ but is not incoherent. The lack of "uniqueness" is not an issue under the assumption that the true matrix is indeed incoherent. These results actually provided a sufficient condition for the following version of identifiability:

$$\mathcal{P}_\Omega(M') = \mathcal{P}_\Omega(M), M \text{ and } M' \text{ are incoherent and have rank} \leq r \Longrightarrow M' = M, \tag{1.4}$$

under a random model of Ω, e.g. each entry of $[n] \times [n]$ is independently included in Ω with probability K/n^2. A necessary condition for (1.4) has been provided in [3, Theorem 1.7] (for $n \times n$ matrix M), which shows that if $K < O(nr\mu \log n)$, then there are infinitely many μ-incoherent matrices M such that (1.2) fails with constant probability. In other words, to prove a high probability recovery result for any fixed μ-incoherent matrix, at least $O(nr\mu \log n)$ observations are needed.

The subtle difference between "a unique matrix" and "a unique incoherent matrix" brings up the following modelling consideration: if the groundtruth matrix M has a property other than low-rankness (including, but not limited to, incoherence), our goal should be to recover a low-rank matrix with this property. Another example of the additional property can be found in global positioning (or rigidity theory): the matrix should be a pair-wise distance matrix. Assuming an additional property of the true low-rank matrix, the identifiability question should be changed accordingly: is there a unique low-rank matrix M with a particular property that matches partial observations?

Under a very mild assumption that M is generic[1] , we obtain another weaker version of the identifiability question (1.3):

$$\text{What is the condition on } \Omega \text{ so that (1.2) holds for generic matrix } M? \tag{1.5}$$

It can be easily shown (see [36, Theorem 2.6]) that for generic M whether M can be recovered by $\mathcal{P}_\Omega(M)$ depends only on Ω, i.e. the positions of the observations. Since Ω corresponds to a bipartite graph with $m + n$ nodes, the identifiability of Ω is an intrinsic property of the corresponding bipartite graph. When $r = 1$, it can be easily shown that Ω leads to identifiability iff the corresponding graph is connected, but for $r > 1$ it not clear what graph property is sufficient and necessary. Kiraly and Tomika [36] provided a sufficient condition based on a very simple algorithm: find a submatrix of size $(r + 1) \times (r + 1)$ with exactly one missing position, then add this position to Ω; perform this step recursively until all positions are added to Ω (return success) or no such submatrix exists (return nil). If success, then a generic M can be uniquely determined by $\mathcal{P}_\Omega(M)$; however, no conclusion can be made if nil is returned.

[1] A property holds for generic matrix M means that there is a nonzero polynomial f on the entries of M such that the property holds in set $\{M | f(M) \neq 0\}$. It implies that the property holds with probability 1 when the entries of M are generated from continuous random distributions, which is stronger than a high probability result. For many engineering applications, proving that the results hold for generic parameters is enough (e.g. [35]).

Several simple necessary conditions, such as $|\Omega| \geq r(m+n-r)$ and "r-connectivity", are provided in [36, Proposition 2.13]. Singer and Cucuringu [37] studied (1.5) using tools from rigidity theory. They proposed a randomized algorithm [37, Algorithm 4] to test a sufficient condition for Ω, which makes no conclusion if it fails. Note that the sufficient conditions and necessary conditions provided in [36,37] are separated, and the gap between the sufficient conditions and the necessary conditions might be very large. Closing this gap, i.e. providing a sufficient *and* necessary condition for the identifiability seems a difficult task. We refer the interested readers to [38] for more discussions on the identifiability issue, including various open questions along this line.

One important theoretical question related to the identifiability is the sample complexity (or phase transition boundary) of matrix completion: how many samples are needed to recover a low-rank matrix, if *any* algorithm is allowed? A naive lower bound is $r(m+n-r) = O(nr)$ (when $m = n$), since this is the degrees of freedom of a rank-r $m \times n$ matrix (see, e.g. [38, Proposition 2.1.13]). For any μ-incoherent matrix, the current best upper bound is $O(nr\mu^2 \log^2 n)$[2] [4,5]. As mentioned above, [3, Theorem 1.7] proved that for a fixed μ-incoherent matrix, at least $O(nr\mu \log n)$ observations are needed to guarantee recovery with high probability. Empirically we find that $O(n \max\{r, \log n\})$ entries seem to be enough for exact recovery (assume μ is small). We conjecture that for a *generic* μ-incoherent matrix, $O(n\mu^2 \max\{r, \log n\})$ entries are enough for exact recovery, which is smaller than the required sample complexity for *any* μ-incoherent matrix.

In practice, inexact recovery is often satisfactory. Consider the following situation: there are only two rank-r matrices $M \neq M'$ that are consistent on a subset Ω, and $\|M - M'\|_F < \epsilon$ where ϵ is a very small positive number; now $\mathcal{P}_\Omega(M)$ does not uniquely determine M, but for practical purposes finding M' is also acceptable since it is very close to the groundtruth M. We can ask an inexact version of the identifiability question, though there is little research on this question. A related, but rather different, complexity theoretical question has been studied recently by Hardt et al. [39].

[2] To be precise, there are two different incoherence constants μ_0 and μ_1, and the bound is $O(nr \max\{\mu_0, \mu_1^2\} \log^2 n)$; see [4, Section I.B] and [5, Theorem 1.1].

1.4 Models of Matrix Completion

In this section, we discuss how to derive the mathematical (optimization) formulations of the matrix completion problem. There are two requirements: low-rankness and consistency with the partial observation. [3] The goal is to find a feasible point in the intersection of two sets $\mathcal{L} = \{Z \in \mathbb{R}^{m\times n} : \text{rank}(Z) \le r\}$ and $\mathcal{S} = \{Z \in \mathbb{R}^{m\times n} : Z_{ij} = M_{ij}, \forall (i,j) \in \Omega\}$. In other words, we try to solve a non-convex feasibility problem

$$\begin{aligned} \text{Find} \quad & Z \in \mathbb{R}^{m\times n}, \\ \text{s.t.} \quad & \text{rank}(Z) \le r, \\ & Z_{ij} = M_{ij}, \ \forall \ (i,j) \in \Omega. \end{aligned} \tag{1.6}$$

In general, constraints are difficult to handle, thus one may want to reformulate (1.6) as an optimization problem with fewer constraints.

There are two different approaches to reformulate (1.6). The first approach is to remove the constraint $\text{rank}(Z) \le r$ and add an objective function $\text{rank}(Z)$, obtaining a rank-minimization problem

$$\begin{aligned} \min_{Z\in\mathbb{R}^{m\times n}} \quad & \text{rank}(Z), \\ \text{s.t.} \quad & Z_{ij} = M_{ij}, \ \forall \ (i,j) \in \Omega. \end{aligned} \tag{1.7}$$

In the language of optimization, this approach is to penalize the violation of the first constraint $Z \in \mathcal{L}$ while keeping the second constraint $Z \in \mathcal{S}$. Problem (1.7) is equivalent to (1.6) in the sense that the optimal solution to (1.7) is either a feasible solution to (1.6) or an infeasibility certificate, and solving (1.6) for different r can lead to the optimal solution of (1.7). As the rank function is a discrete function and thus intractable, one may further relax this problem to a nuclear norm minimization problem

$$\min_{Z\in\mathbb{R}^{m\times n}} \quad \|Z\|_*, \quad \text{s.t.} \quad Z_{ij} = M_{ij}, \ \forall \ (i,j) \in \Omega, \tag{1.8}$$

where $\|Z\|_*$ is the nuclear norm of Z, i.e. the sum of all singular values of Z. See more discussions of this approach in Section 1.4.1.

[3] As argued before, this may not be a good starting point for modelling, since these two requirements may not be enough to uniquely determine the original matrix. We could, for example, start from three requirements: low-rankness, consistency with the partial observations, incoherence. Nevertheless, for simplicity we just start from these two basic requirements.

The second approach is to penalize the violation of the second constraint $Z \in \mathcal{S}$ while keeping the constraint $Z \in \mathcal{L}$, obtaining the following problem (using square loss function, but other loss functions can be used as well):

$$\begin{aligned} \min_{Z \in \mathbb{R}^{m \times n}} \quad & \|\mathcal{P}_\Omega(M - Z)\|_F^2, \\ \text{s.t.} \quad & \operatorname{rank}(Z) \leq r. \end{aligned} \tag{1.9}$$

Although the rank function is still intractable, a simple trick can make it tractable: any matrix Z with rank no more than r can be represented as $Z = XY^T$, where $X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}$. Then (1.9) can be rewritten as

$$P0: \quad \min_{X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}} \quad F(X, Y) = \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2. \tag{1.10}$$

MF based formulation has gained great popularity in the recommender systems field and served as the basic building block of many competing algorithms for the Netflix Prize [1, 40]. See more discussions on this formulation in Section 1.4.2

The major difference of these two approaches lies in how to deal with the rank function: the first approach uses the nuclear norm (a convex function) to approximate the rank, while the second approach encodes the low-rankness explicitly through the product of two small matrices. As a result, the difficulties of proving recovery for the two approaches are different: the first formulation is convex and can be solved to global optima, thus the difficulty is to show that the true low-rank matrix is indeed the unique global optimal solution of (1.8); in contrast, the true low-rank matrix must be a global optimal solution of the second formulation, thus the main difficulty is how to find the global optimal solution of this non-convex problem (assuming the identifiability holds). We will discuss the existing algorithms and theoretical analysis related to these two approaches, as well as their pros and cons in the following.

1.4.1 Nuclear Norm Formulation

The nuclear norm approach has a nice connection with compressive sensing [41–44] (or a related subject Lasso [45]). Specifically, the nuclear norm of a matrix can be viewed as the ℓ_1-norm of the vector consisting of all singular values of a matrix, while the rank of the matrix is the ℓ_0-norm (i.e. the number of non-zero entries) of that vector. The

ℓ_1-norm has been shown to be a good convex surrogate for the ℓ_0-norm in compressive sensing, thus it is natural to use the nuclear norm as a surrogate for the rank function.

In a series of remarkable papers [2–5], it has been shown that given a rank-r matrix $M \in \mathbb{R}^{n\times n}$ satisfying an incoherence condition (will define later), solving (1.8) will exactly reconstruct M with high probability provided that $\tilde{O}(rn\log^2 n)$ entries are uniformly randomly revealed, here $\tilde{O}$ notation hides the dependence on the incoherence constant of M. A key step in these papers is to construct a dual certificate satisfying certain conditions via an iterative procedure. To show this construction works, one needs to bound $\|\mathcal{P}_\Omega(W_k) - W_k\|$ where W_k is the iterate, but due to the possible dependency of the iterates W_k on the sample set Ω, the first two papers [2, 3] use rather intricate techniques, yet obtaining suboptimal sample complexity bounds (though only an extra polylogarithmic factor of n in [3]). Gross et al. [46] introduced an elegant idea called "golfing scheme" which uses independent sample sets $\Omega_k \subseteq \Omega$ at each iteration when constructing a dual certificate. This preprint only considered a matrix sensing type problem (Pauli measurements), not the case with missing entries. The golfing scheme was applied to to matrix completion in Gross [4] and Recht [5], which improved the sample complexity bound to $\tilde{O}(nr\log^2 n)$ with much simpler proofs than those of [2, 3].

Another proof for the recovery is given by Negahban and Wainwright [32]. An earlier paper Negahban et al. [47] showed that the proof of recovery can be boiled down to "restricted strong convexity", which is a variant of the RIP condition (1.1) with a major difference that the inequality does not hold for any low-rank matrix Z, but holds for Z in a certain restricted set (related to the descent cone of the optimization problem). The proof strategy relies on several probabilistic techniques besides the concentration inequalities, such as discretization and bounding the covering number, symmetrization by a Rademacher sequence, etc. This proof is not based on the original formulation (1.8), but the regularized version

$$\min_{Z\in\mathbb{R}^{m\times n}} \quad \|Z\|_* + \lambda \sum_{(i,j)\in\Omega} (Z_{ij} - M_{ij})^2, \tag{1.11}$$

with extra constraints on each entry of the matrix $|Z_{i,j}| \le c, \forall i, j$, where c is a certain parameter. The extra constraints on the infinity norm of the matrix (called "spikeness" in that paper) may require extra projection steps when solving the optimization problem.

For noisy matrix completion, Candes and Plan [3] considered a quadratically constrained minimization problem

$$\min_{Z\in\mathbb{R}^{m\times n}} \quad \|Z\|_*, \quad \text{s.t.} \sum_{(i,j)\in\Omega} (Z_{ij} - M_{ij})^2 \le \epsilon, \tag{1.12}$$

and proved that the reconstruction error is proportional to the initial observation error. The restricted strong convexity (RSC) of [32] also leads to a bound of the error between the solution to (1.11) and the true matrix. Possibly due to the lack of RIP-type condition (or RSC), the error bound in [23] can be $\sqrt{n}$ times worse than the bound of [32].

On the computational side, problems (1.8) and (1.12) can be reformulated as a semidefinite program (SDP) and solved to global optima by standard SDP solvers when the matrix dimension is smaller than, say, 500. To solve problems with larger size, researchers have developed first order algorithms, including the SVT (singular value thresholding) algorithm for the formulation (1.8) [48], and several variants of the proximal gradient method for the formulation (1.11) [49, 50]. The linear convergence of the proximal gradient method has been established for the formulation (1.11) under certain conditions [51, 52]. The linear convergence only implies that the number of iterations is not large, but the per-iteration cost of the proximal gradient type methods for the nuclear formulation is still rather high. In particular, each iteration requires computing SVD (Singular Value Decomposition), either exactly or inexactly, and the computation cost of SVD increases rapidly as the dimension of the problem increases, making these algorithms rather slow or even useless for problems of huge size. The other major drawback is the memory requirement of storing a large m by n matrix.

1.4.2 Matrix Factorization Formulation

In the matrix factorization approach, the unknown rank r matrix is expressed as the product of two much smaller matrices XY^T, where $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}$, so that the low-rank requirement is automatically fulfilled. In addition to the basic formulation (1.10), another popular formulation is to add regularizers to control the norms of X and Y:

$$\text{P0}' : \quad \min_{X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{n\times r}} \quad \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 + \lambda(\|X\|_F^2 + \|Y\|_F^2). \tag{1.13}$$

This formulation can be obtained by MAP (maximum a posteriori) estimation under a certain probabilistic model of the low-rank matrix [53]. There are a few variants of this formulation: the coefficient λ can be zero (reduced to (1.10)) [24, 54–56] or different for each row of X, Y [57]; each square loss term $[M_{ij} - (XY^T)_{ij}]^2$ can have different weights [1]; an additional matrix variable $Z \in \mathbb{R}^{n\times r}$ can be introduced [58].

Problem (1.13) is a non-convex fourth-order polynomial optimization problem. In general, fourth-order polynomial minimization is NP-hard, and without resorting to the structure of the problem one cannot hope to find a global optimal solution. The stationary points of (1.13) can be found by many standard nonlinear optimization algorithms such as the gradient methods, AltMin (alternating minimization) [1] and SGD (stochastic gradient descent) [1, 40, 59, 60]. The formulation (1.13) is suitable for AltMin since the objective function is convex with respect to either X or Y, even though not jointly convex over (X, Y), thus minimizing over one of the factors with another fixed is easy. It is also suitable for SGD since the loss function $\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 = \sum_{(i,j)\in\Omega}[M_{ij} - (XY^T)_{ij}]^2$ is decomposable across the samples, thus at each iteration one can use the error for the (i, j)-th entry $M_{ij} - (XY^T)_{ij}$ to update the i-th row of X and the j-th row of Y. AltMin and SGD (and their variants) are the two most popular methods to solve matrix factorization based formulations for the Netflix prize [1].

MF based formulation is very popular for large scale matrix completion and other related applications due to several reasons. First, the compact representation of the unknown matrix greatly reduces the per-iteration computation cost as well as the storage space (requiring essentially linear storage of $O((m + n)r)$ for small r). Second, the per-iteration computation cost is rather small and people have found in practice that huge size optimization problems based on the factorization model can be solved very fast. Third, as elaborated in [1], the factorization model can be easily modified to incorporate additional application-specific requirements.

Although (1.13) can be solved by standard AltMin or SGD, the original versions of the two algorithms may not be efficient for solving some large scale problems, and a great deal of recent effort has been devoted to an even faster algorithm. The obvious pros and cons of AltMin and SGD are the following: AltMin is easily parallelizable but has higher per-iteration computation cost than SGD; in contrast, SGD requires little computation per iteration, but its parallelization is a bit challenging. Recently several

parallelizable variants of the SGD [61–63] and AltMin [64,65] with very low per-iteration cost have been developed. Some of these algorithms have been tested in distributed computation platforms and can achieve good performance and high efficiency, solving very large problems with more than a million rows and columns in just a few minutes For instance, [62] reported to complete a matrix with $m = n = 10^6, r = 10$ in under 18 minutes; for the Netflix problem their algorithm took under 3 minutes on a 12-core workstation.

The key to parallelize SGD lies in how to pick the order of the component functions. In particular, [61–63] proposed to divide Ω into the union of "diagonal sets", where each diagonal set consists of entries that are not in the same row or column, and view the sum of the component functions $(M - XY^T)_{ij}^2$ corresponding to one diagonal set as a new component function. By this partition, each row of X and Y appears in one group at most once, thus all rows of X and Y based on the component functions in one group can be updated in parallel.

Alternating minimization in the context of matrix completion usually refers to the two-block coordinate descent (BCD) method that updates X and Y alternately. Since the objective function is decomposable across rows of X with fixed Y (or across rows of Y with fixed X), one can also view each row of X and Y as a block. Other choices of blocks lead to different and possibly faster algorithms. The CCD algorithm [64] updates each entry of X, Y cyclically according to a specific order. In the CCD++ algorithm [65], the blocks are chosen to be the columns of X, Y; this choice is interpreted as the "feature-wise" update order and is shown to be much faster than AltMin or CCD algorithm in some real data sets.

Empirically people have found that the algorithms for the MF model work very well. In particular, in the noiseless case one can always exactly recover the original matrix by algorithms for the MF model when there are enough samples, even though the MF formulation is non-convex. Then a question of significant interest is to provide a theoretical understanding of this phenomenon. In the language of optimization, the question is why a non-convex optimization problem can be solved to global optima. This is the main theoretical question we try to answer in the thesis.

1.5 Existing Results on MF Formulation

1.5.1 Brief Overview

The first recovery guarantee for the factorization based matrix completion is provided in [33], where Keshavan, Montanari and Oh considered a factorization model in Grassmannian manifold. In their formulation, the unknown rank r matrix is expressed as the product of three smaller matrices XSY^T with $X \in \mathbb{R}^{m\times r}, X^TX = mI$, $Y \in \mathbb{R}^{n\times r}, Y^TY = nI$, which is similar to SVD except that S may not be a diagonal matrix. The considered optimization problem is

$$\begin{aligned} &\min_{X\in\mathbb{R}^{m\times r}, X^TX=mI, Y\in\mathbb{R}^{n\times r}, Y^TY=nI} F(X,Y), \\ \text{where} \quad &F(X,Y) = \min_{S\in\mathbb{R}^{r\times r}} \sum_{(i,j)\in\Omega} \|M_{ij} - (XSY^T)_{ij}\|^2. \end{aligned} \tag{1.14}$$

The objective function $F(X,Y)$ depends on X, Y only through their column spaces, thus it can be viewed as a function on Grassmannian manifolds (a Grassmannian manifold consists of linear subspaces of a certain dimension). The resulting optimization problem $\min_{X,Y} F(X,Y)$ is in fact a two-level optimization problem. They showed that the unknown matrix can be recovered by a proper initialization and a gradient descent method on Grassmannian manifold. Besides being quite complicated, this model is not as flexible as the factorization model in Euclidean space (cannot add application-specific requirements on the factors), and it cannot be solved by many advanced large-scale optimization algorithms such as BCD-type and SGD-type methods.

The factorization model in Euclidean space was first analyzed in an unpublished work [24] of Keshavan[4] , as well as a later work of Jain et al. [54]. Both works considered AltMin with resampling, a special variant of the original AltMin. The sample complexity bounds were later improved by Hardt [55] and Hardt and Wooters [56], where in the latter work, notably, they devised an algorithm with a corresponding sample complexity bound independent of the condition number. However, these improvements are obtained for more sophisticated versions of resampling-based AltMin, not the original AltMin.

[4] Reference [24] is a PhD thesis that discusses various algorithms including the algorithm proposed in [33] and AltMin. In this thesis when we refer to [24], we are only referring to [24, Ch. 5] which presents resampling-based AltMin and the corresponding result.

There are a few very subtle issues with the resampling scheme, and we will discuss these issues in the following subsection.

1.5.2 Issues of Resampling

Resampling is quite subtle, as discussed in the recent work on phase retrieval by Candès, Li and Soltanolkotabi [66]; here we make some additional comments since resampling is used more widely in matrix completion.

Resampling scheme can be used at almost no cost for the nuclear norm approach [4,5, 67], but for AltMin it causes many issues. At first, it may seem that for both approaches resampling is a cheap way to get around a common difficulty: the dependency of the iterates on the sample set. However, there is a crucial difference: for the nuclear norm approach, resampling is just a proof technique used in a "conceptual" algorithm for constructing the dual certificate, while for AltMin, resampling is used in the actual algorithm. This difference causes some issues of resampling-based AltMin at conceptual, practical and theoretical levels.

1) Gap between theory and algorithm. Algorithmically, an easy resampling scheme is to randomly partition the given set Ω into non-overlapping subsets $\Omega_k, k = 1, \ldots, L$, as proposed in [24,54] [5] . However, the results in [24,54–56] actually require a generative model of independent Ω_k's, instead of sampling Ω_k's based on a given Ω. Therefore, the results in [24, 54–56] do not directly apply to the resampling scheme that is commonly used.

This issue has been discussed by Hardt and Wooters in [56, Section D], and they proposed a new resampling scheme [56, Algorithm 6] to which the results in [24, 54–56] can apply, provided that the generative model of Ω is exactly known. In practice, the underlying generative model of Ω is usually unknown, in which case the scheme [56, Algorithm 6] does not work. In contrast, the classical results in [2–5] and our result herein are robust to the generative model of Ω: these results actually prove that for, say, 99% of Ω (actually, $(1 - 1/n^c)$ portion) with a given size, one can recover M through

[5] The description in [54] has some ambiguity: in [54, Algorithm 2] the authors wrote "partition" Ω into a few subsets, but they also wrote "sampling with replacement". From our understanding, "partition" refers to Model 2 and "sampling with replacement" refers to Model 3 in Section 1.5.3; anyhow, under either model Ω_k's are dependent and the claimed result [54, Theorem 2.5] does not apply. See more discussions in Section 1.5.3.

a certain algorithm, thus for many reasonable probability distributions of Ω a high probability result holds. See more discussions in Section 1.5.3.

2) Impracticality. As argued previously, assuming a generative model of Ω_k's is not practical since Ω is usually given. For given Ω, the only known validated resampling scheme [56, Algorithm 6], besides not being robust to the underlying generative model of Ω, might be too complicated to use in practice. Even the simple resampling scheme of partitioning Ω (which has not been validated yet) is rather unrealistic. First, each sample is used only once during the algorithm, which is a waste of resources. Second, different accuracy requirements will lead to different pre-partition of the samples, and thus different forms of the algorithm. If the algorithm has produced an estimate of M and one asks for a more accurate estimate, then one has to re-partition Ω and re-run the algorithm from the beginning.

3) Inexact recovery. A theoretical consequence of the resampling scheme is that the required sample complexity $|\Omega|$ becomes dependent on the desired accuracy ϵ, and goes to infinity as ϵ goes to zero. This is different from the classical results (and ours) where exact reconstruction only requires finite samples. While it is common to see the dependency of *time complexity* on the accuracy ϵ, it is relatively uncommon to see the dependency of *sample complexity* on ϵ.

4) Randomized algorithm v.s. deterministic algorithm. Given Ω, resampling-based AltMin is a randomized algorithm that can only achieve the desired performance with a certain probability; in contrast, the alternating-type methods we consider are deterministic. Such a difference is hidden in the high probability statement: in our result,"with probability 99%" means that out of all possible sets Ω, 99% of them can lead to exact recovery by the algorithm *for sure*; for resampling-based AltMin, "with probability 99%" means that "for 99% of Ω, resampling-based AltMin can recover M *with probability* 99%" (the precise statement would be long, so we present a simpler one).

In a recent work [68] the authors have managed to remove the dependency of the required sample size on ϵ by using a singular value projection algorithm. However, [68] considers a matrix variable of the same size as the original matrix, which may not have the same advantage in memory and modeling flexibility as provided by the matrix factorization approach considered in this thesis. Moreover, the algorithm in [68] still uses independent samples in a number of iterations (though not all iterations), which

may suffer from the same issues we discussed above.

We remark that the gap between the theory and the actual algorithm also exists in a recent work [69] on SGD. In particular, the authors showed the global convergence of a special variant of SGD [69, Algorithm 1] for low-rank matrix problems, including matrix completion. However, [69, Algorithm 1] requires the samples used in each iteration to be independent, which is difficult (if not impossible) to implement for matrix completion: the independence of the iterates does not hold for a typical SGD that randomly picks a sample at a time (either sampling with or without replacement). Note that resampling-based AltMin can be viewed as a hybrid of mini-batch SGD and AltMin, where each Ω_k is a mini-batch; [69, Algorithm 1] is similar to the resampling-based AltMin in the sense that each Ω_k consists of just one sample. Thus it is not surprising that [69, Algorithm 1] suffers from the same issue as resampling-based AltMin. See Section 1.5.3 for more discussions.

1.5.3 More Discussions on Resampling and SGD

We will discuss various resampling schemes and SGD in detail, and we hope that these discussions could inspire future research.

More discussions on resampling

As discussed earlier, there is a gap between theory and algorithm for resampling-based AltMin. In short, the results in [24,54–56] do not apply to the simple resampling scheme that partitions Ω into subsets Ω_k's, since they require Ω_k's to be independent and each entry of the matrix is included in Ω_k with probability p_k. We elaborate the two random models below.

- Model 1 (generative model of independent Ω_k's): generate each Ω_k independently by a Bernoulli model where each entry is included in Ω_k with probability p_k, $k = 1, \ldots, L$. Note that one entry can appear in multiple Ω_k's. The results in [24, 54] apply to AltMin based on this random model. However, this random model is impractical: usually the observation set Ω is given and one cannot just "assume" Ω_k's follow a certain distribution.

- Model 2 (generate Ω first, then randomly partition Ω): For given Ω generated by a certain underlying random model, randomly partition Ω into L sets $\Omega = \Omega_1 \cup \cdots \cup \Omega_L$ (assign each entry a label randomly drawn from $\{1, \ldots, L\}$), where L is a certain constant. It might be tempting to think that Ω_k's are independent provided that each entry of M is included in Ω independently. However, these Ω_k's are not statistically independent: they are always non-overlapping, while for any two independent sample sets there is a positive possibility that they are overlapping. To put in another way, in this model one entry cannot appear in multiple Ω_k's while in Model 1 one entry can appear in multiple Ω_k's. Therefore, the results in [24, 54] do not apply to AltMin based on this resampling scheme.

Now we have discussed two random models: one leads to a "theoretical algorithm" (an algorithm that cannot be implemented), and the other leads to a practical algorithm with no theoretical guarantee. It is natural to ask whether one can close the gap between these two random models, which motivates the following question [56, Section D]:

Given a sample set Ω generated from a random model, can we generate subsets $\Omega_k \subseteq \Omega, k = 1, \ldots, L$ that follow Model 1 (L independent Bernolli distributions)? (1.15)

Although Model 2 does not achieve this goal, one may think that a simple variant will work. Now we discuss several possibilities.

- Model 3 (generate Ω, then sample Ω_k with replacement): for given Ω generated by a certain underlying random model, sample the entries of Ω with replacement to obtain $\Omega_k, k = 1, \ldots, T$. This sampling with replacement model is used in randomized BCD [70] and SGD; for the matrix completion problem, it is more practical than Model 1 since the corresponding algorithm can work on a given Ω. Different from Model 2, here one entry can appear in multiple Ω_k's; different from Model 1, here one entry can appear multiple times in one Ω_k. These Ω_k's are still dependent [6] : the probability of two sets Ω_1, Ω_2 having common entries

[6] Note that if we assume Ω is fixed and the "ambient space" is Ω, then Ω_k's are independent; in other words, Ω_k's are *conditionally independent* given Ω. However, in [24, 54–56] the ambient space is $[m] \times [n]$ and Ω itself is random, then Ω_k's are dependent.

is higher than a generative model of independent Ω_k's. Thus this model, again, is not covered by the results in [24, 54–56].

- Model 4 (for Ω from Bernolli model): this model is described in [56, Algorithm 6]. The construction there is delicate and very different from a simple resampling scheme one can think of, such as Model 2 or Model 3. For instance, when $p_1 = \cdots = p_L$, [56, Algorithm 6] requires computing $q_k = \frac{1}{1-(1-p_1)^L}\binom{mn}{k}p_1^k(1-p_1)^{mn-k}, k = 1, \ldots, L$. In addition, this scheme is not practical since it is not robust to the unknown generative model of Ω, as discussed in Section 1.5.2.

- Model 5 (sample Ω' with replacement, then randomly partition Ω'): assume Ω' is generated by sampling with replacement, and we record not only the samples but also the number of times a sample is drawn. One can think of Ω' as a set with possibly repeated elements. Partition Ω' uniformly randomly into L sets $\Omega_1, \ldots, \Omega_L$ with equal size. This random model is different from Model 1 as it allows repetitions in each set Ω_k, while Model 1 does not; nevertheless, when performing AltMin one can ignore the repetition in Ω_k. This random model is equivalent to generating Ω_k's independently from a sampling with replacement model, thus the results in [24,54] probably apply to this model (though not directly applicable as they are tailored for the Bernolli model of Ω_k). This model is still not practical for AltMin: given the set of non-repeated samples Ω, it is not clear how to generate Ω' so that Ω' follows the distribution of a sampling with replacement model. However, as we will discuss shortly, this model can be used for the nuclear norm approach [4, 5].

One might argue that Model 2 and Model 3 are just approximations of Model 1 in practice. However, it is necessary to validate this "approximation" rigorously[7] . We will explain below how the resampling scheme for the nuclear norm approach is validated and why the validation of the resampling scheme for AltMin is more difficult.

Take [5] as an example. [5, Proposition 3.1] first established the "equivalence" of the sampling with replacement model of generating Ω' and the uniform model of generating

[7] "Approximation" is too vague and does require justification. For example, it seems that Model 2 (sampling without replacement) and Model 3 (sampling with replacement) are quite similar; however, in a different setting (multi-block ADMM), it has been observed that Model 3 leads to *divergence* of the algorithm and Model 2 leads to *convergence* [71].

Ω from the collection of sets with a given size (not really equivalence, but just that if the desired result applies to the first random model then it applies to the second random model). Note that Ω' is allowed to contain duplicated entries. When constructing dual certificate, one can partition Ω' to obtain independent Ω_k, as in Model 5 described above. This partition is legitimate since it just happens in the theory, not in an actual algorithm. Another way to understand this process is the following: given Ω, there exists a way of labelling the entries of Ω by the number of repetitions so that one can obtain Ω' that follows the distribution of Model 5; one can then partition Ω' based on Ω and these labels. The key is that one does not need to know the true labelling, but just the existence of such a labelling.

The validation of resampling for the nuclear norm approach in [5] is relatively simple since [5] can work with an artificial generative model of Ω' (just a proof technique). For AltMin, one needs to consider a *mixture* of the generative model of Ω and a practical resampling scheme for Ω_k's, which seems quite difficult.

Issue of SGD: samples are dependent

Now we discuss the gap between theory and algorithm for SGD in [69]; these discussions will be based on the above discussions for resampling. In SGD, at each iteration we randomly pick one entry (i, j) from Ω and update the corresponding $X^{(i)}, Y^{(j)}$. There are two popular versions of SGD: RP-SGD (randomly permuted SGD; i.e. sampling (i, j) without replacement) and R-SGD (randomized SGD; i.e. sampling (i, j) with replacement).

The theory of [69] requires the samples to be selected "uniformly and independently at random" from a sampling distribution. We argue that in both versions of SGD, the samples are *not* independent uniform samples, thus the theory of [69] does not apply. We first look at RP-SGD that randomly permutes the available $|\Omega|$ samples and performs a gradient step for each sample once (i.e. one epoch; more epoch's will definitely lead to dependence of iterates). This one-pass algorithm is not practical: as shown in Figure 3.2 of Section 2.3, one epoc of SGD only leads to a test error of more than 1.2 [8] .

[8] The algorithm proposed in [69] is a variant of SGD and may perform better than the vanilla SGD; however, it is unlikely that one epoc of that algorithm can obtain, say, a test error < 0.1 in the setting of Figure 3.2.

Putting aside the impracticality, even within one epoc we cannot say the samples are independent: for a given Ω generated by a certain random model, the elements of Ω are not independent as discussed in Model 2 above.

One may still wonder: can we just assume a generative model of independent samples? In other words, given the observation set Ω consisting of K positions, can we just *assume* the K samples (positions) are independently generated from a certain distribution on $[m] \times [n]$? The answer is no, if K is reasonably large. To see why, suppose ξ follows a uniform distribution on $[n] \times [n]$ (assume $m = n$), and we independently draw K samples $\xi_1, \ldots, \xi_K$, then the probability of repetition is close to 1 in a reasonable setting $K \geq 4n$[9] . Unless we can "mimic" the repetition for given distinct positions, we cannot assume that the distinct observations are generated independently (see also the discussion in Model 5). This is different from traditional stochastic optimization where K samples drawn independently from a *continuous* distribution are distinct with probability 1.

We then discuss R-SGD, i.e. at each iteration randomly pick one sample s_i from Ω and perform a gradient step for this sample. As discussed before in Model 3, these samples $s_1, s_2, \ldots$ are conditionally independent given Ω, but are not necessarily independent when the ambient space is $[m] \times [n]$ and Ω itself is randomly generated. Therefore, again there is a gap between R-SGD and the theory of [69].

This gap may seem counter-intuitive for researchers who are familiar with SGD: it is a common assumption that the samples are independent in the analysis of SGD (see, e.g. [72]), why does the independence assumption fail here? A short answer is that the traditional analysis only requires the "conditional" independence of the samples. More specifically, consider a stochastic optimization problem

$$\min_x E_\xi f(x; \xi), \tag{1.16}$$

where ξ is a random variable. One way to solve this problem is by SAA (sample average

[9] The probability can be computed as follows: when $K = n$, the probability that there are at least two repeated samples is $1 - (1 - \frac{1}{n^2})(1 - \frac{2}{n^2}) \ldots (1 - \frac{n-1}{n^2}) \geq 1 - \sqrt{\frac{1}{(1+1/n)^{n-1}}} \approx 1 - \frac{1}{\sqrt{e}} \approx 0.4$; a similar computation shows that when $K = 4n$, the probability of repetition is approximately (at least) $1 - \frac{1}{\sqrt{e^{16}}} > 99.9\%$. It is reasonable to assume $K \geq 4n$: the minimal number of samples for recovering a rank-r matrix is $r(2n - r) \approx 2rn$, thus when $r \geq 2$ at least $4n$ observations are needed.

approximation), i.e. approximating (1.16) by a finite sum

$$\min_x \frac{1}{K}\sum_{k=1}^{K} f(x;\xi_k), \tag{1.17}$$

where $\xi_1,\dots,\xi_K$ are independent samples of ξ. The problem (1.17) can be solved by R-SGD that updates x based on samples $s_1, s_2, \dots$ drawn independently from a fixed distribution on $\{\xi_1,\dots,\xi_K\}$. Note that $s_1, s_2, \dots$ are conditionally independent given $\{\xi_1,\dots,\xi_K\}$, but not necessarily independent when the ambient space is the sample space of ξ (e.g. when ξ follows a Gaussian distribution in $\mathbb{R}^n$, then the sample space is $\mathbb{R}^n$). The lack of independence of $s_1, s_2, \dots$ for the original problem (1.16) is not an issue, since we only care about the convergence of R-SGD to the solution of the finite sum problem (1.17). It is a different task to prove that the solution to (1.17) is a good approximate solution to (1.16), and has nothing to do with R-SGD.

In the matrix completion problem, the counterpart of (1.16) is

$$\min_{X,Y} \|M - XY^T\|_F^2, \tag{1.18}$$

and the counterpart of (1.17) is

$$\min_{X,Y} \|\mathcal{P}_\Omega(M - XY^T)\|_F^2. \tag{1.19}$$

If we only care about solving (1.19) (i.e. reduce the training error), then the samples used in R-SGD are (conditionally) independent given Ω. However, the goal of [69] is to directly show that SGD approximately solves (1.18) (i.e. reduce the test error); for this purpose, conditional independence given Ω is not enough, and the theory of [69] requires independent samples in the ambient space $[m]\times[n]$. Since R-SGD does not use independent samples in $[m]\times[n]$, it is not covered by the theory of [69].

Our analysis of SGD in this thesis does not suffer from the issues discussed above; the reason is that we decouple optimization and approximation. Roughly speaking, we show that the solution of the finite sum problem (1.19) is an approximate solution of the "expectation" problem (1.18), thus any algorithm that finds the solution to (1.19) can solve (1.18). This framework is similar to the traditional framework in stochastic optimization or learning theory that decouples the "optimization error" and "approximation error". Therefore, for any version of SGD (no matter RP-SGD, R-SGD, or the

cyclic version of incremental gradient method), to show it solves (1.18) one only needs to prove a pure optimization result that it converges to the solution of (1.19) [10] .

1.6 Contributions

1.6.1 Summary of Contributions

Despite the great empirical success, the theoretical understanding of the algorithms for the MF formulation is fairly limited. More specifically, the fundamental question of whether these algorithms (including many recently proposed ones) can *exactly* recover the true low-rank matrix remains largely open. We summarize below our contributions in both algorithm design and in theory.

Novel numerical finding. We investigate the performance of various standard algorithms via numerical experiments. A novel finding is that for all algorithm we have studied, the maximum row-norm (or equivalently, the incoherence constants) of the iterates is the key indicator of whether it can recover the matrix or not, provided that the number of samples exceeds the fundamental limit. For example, we show that the regularizer $\lambda(\|X\|_F^2+\|Y\|_F^2)$ can improve the recovery performance even in the noiseless case, and we argue that this is *not* because it helps control the *norms* of the iterates, but because it helps control the *row-norms* of the iterates.

New practical formulation and algorithms. Inspired by the numerical findings, we propose a new formulation that adds constraints and/or regularizers to control the *row-norms* of the iterates, as well as additional constraints/regularizers on the *norms* of the iterates. Simulations show that this new formulation can exactly recover the true matrix even when the number of samples is close to the fundamental limit. Note that these algorithms only take little extra computation cost compared to existing algorithms in their "failed regimes", and usually do not require any extra step in their "successful regimes". The reason is that our regularizers are of penalty-function type and do not affect the algorithms when the iterates do not violate the desired constraints.

Recovery result. We provide a theoretical justification of the new nonconvex MF

[10] Note that the statements in this paragraph are not precise; for example, "solution of (1.19)" actually refers to " stationary point in a certain region", not "global optimal solution". Anyhow, the key point is the decoupling of optimization and approximation.

formulation. More specifically, we show that under similar conditions to those used in previous works, for a specific initialization many standard optimization algorithms for this new formulation indeed converge to the global optima and recover the true low-rank matrix (see Theorem 4.2.1). Our result applies to a large class of algorithms including the gradient methods, SGD and many block coordinate descent type methods such as two-block AltMin and block coordinate gradient descent. To the best of our knowledge, this is the first result that provides exact recovery guarantee for these standard algorithms. In addition, our result also provides the first recovery guarantee for AltMin without resampling (i.e. without using independent samples in different iterations). We elaborate these two contributions in light of the existing works below.

1) Our result provides a validation of the matrix factorization based formulation rather than a validation of a single algorithm. In other words, the success of many algorithms attributes mostly to the *geometry* of the problem, rather than the specific algorithms being used. As a result, it is easy to apply our result to many recently proposed algorithms (e.g. [65,73]), which are variants of classical optimization methods.

2) Our result applies to the standard forms of the algorithms (though our optimization formulation is a bit different), which do not require the additional resampling scheme used in other works [24, 54–56]. We obtain a sample complexity bound that is independent of the recovery error ϵ, while all previous sample complexity bounds for the matrix factorization based formulation (in Euclidean space) depend on ϵ. See more discussions on the resampling scheme in Section 1.5.

1.6.2 Proof Techniques

Below we briefly describe the main techniques and the proof framework.

Difference of two existing approaches. Let us briefly discuss the difference of the proof strategy of [33] for Grassmannian manifold and that of [24] [54] for resampling-based AltMin. Roughly speaking, both approaches need to bound $\mathcal{P}_\Omega(Z)$, where Z is a certain matrix related to the iterates (see, e.g. equation (16) in [54]). The first challenge is how to deal with the dependency of Z on Ω. One simple strategy is to use a resampling scheme to decouple Z and the observation set as in [24,54], and the subsequent analysis can be relatively easy. This strategy artificially avoids the difficulty, and causes a few issues discussed earlier in Section 1.5. Another strategy, as employed in [33], is to use a

random graph lemma proved by Feige and Ofek [74] that implies a bound on $\|\mathcal{P}_\Omega(Z)\|_F$ when Z is rank-1 and possibly dependent on Ω.

Coupled perturbation analysis. The dependency of iterates on Ω is just the first barrier, which we will overcome using the random graph lemma of [33, 74]. There are other difficulties besides the probability tools. The complications of the proof in [33] are mostly due to the Grassman manifold model. It includes heavy computation of various quantities in Grassman manifold; in addition, much effort is spent in estimating the terms related to the extra factor S which enables the decoupling of X and Y ([33] actually uses a three-factor decomposition XSY^T).

For our problem, the difficulty of dealing with the factorization model in Euclidean space is very different from that of Grassman manifold [33]. We avoid the computation in Grassman manifold as well as the estimation of various terms related to the extra factor S, but the price to pay is the coupling of X and Y. The main technical challenge is the "coupled perturbation analysis": given X, Y such that $\|XY^T - M\|_F$ is small, find a decomposition $M = UV^T$ such that U, V satisfy a few conditions including being close to X and Y respectively (Proposition 4.3.1 and Proposition 4.3.2). The difference from traditional perturbation analysis in [75] (i.e. if two matrices are close then their SVD factor spaces are close) is that in [75] the SVD factor spaces are fixed and have closed-form expression, while in our problem U, V are up to our choice. As a result, [75] only requires a "verification" proof that bounds a given error, while we need a "constructive" proof that designs a factorization $M = UV^T$ and shows it works. Naive factorizations of M such as SVD does not work; in fact, we need to factorize $M = UV^T$ according to the structure of X and Y. For Proposition 4.3.1, utilizing a coarse structure of X, Y is enough. For Proposition 4.3.2, it turns out that we need an iterative procedure to construct the factorization $M = UV^T$; moreover, the preliminary analysis in Appendix A.4.2 illustrates that a simple one-step construction probably does not work and a sophisticated iterative procedure is necessary.

Proof outline. The overall proof framework can be summarized as follows: we prove a certain type of local convexity (though quite different) of the objective function around the global optima, thus starting from a good initial point a locally convergent algorithm will converge to the global optima. The proof can be divided into two parts: the problem property (Lemma 4.2.1) and the algorithm property (Lemma 4.2.2). For

the problem property, Lemma 4.2.1 states that the objective function behaves like a convex function in a certain neighborhood of the global optima (more precisely, it is an "incoherent bounded neighborhood" of M, and we will call it "basin of attraction", or simply "basin"), thus there is no other stationary point in this basin. The main technical difficulty of proving Lemma 4.2.1 lies in Proposition 4.3.1 and 4.3.2, which can be viewed as coupled perturbation analysis. For the algorithm property, Lemma 4.2.2 states that starting from a certain initial point (easily computable), many standard algorithms generate a sequence that are inside the basin and these algorithms also convergence to stationary points.

Algorithm requirements. To guarantee that the iterates stay in the basin, it is not enough to have a descent algorithm since the decrease of $\|\mathcal{P}_\Omega(M - X_k Y_k^T)\|$ does not imply the same in $\|M - X_k Y_k^T\|_F$. We provide three conditions and show that if an algorithm satisfies either of them, then with specific initialization the iterates will stay in the desired basin (see Proposition 4.4.1). A special case of the third condition has been used in [33] for Grassman manifold optimization. Together, these three conditions cover a wide spectrum of algorithms including GD, SGD and alternating type methods.

Problem property and perturbation analysis. The problem property proved in Lemma 4.2.1 is that for any (X, Y) in a certain basin, we have

$$\langle \nabla_X f(X), X-U\rangle + \langle \nabla_Y f(Y), Y-V\rangle \geq c(\|X-U\|_F^2 + \|Y-V\|_F^2), \text{ for some } (U,V) \in \mathcal{X}^*, \tag{1.20}$$

where f is the objective function, and $\mathcal{X}^* = \{(U,V) \mid UV^T = M, U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}\}$ can be viewed as the set of global optimizers. To understand this relation, denoting $\boldsymbol{x} = (X, Y), \boldsymbol{x}^* = (U, V)$ and using $\nabla f(\boldsymbol{x}^*) = 0$, we obtain that for any $\boldsymbol{x}$ in a certain basin,

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle \geq c\|\boldsymbol{x} - \boldsymbol{x}^*\|^2, \text{ for some } \boldsymbol{x}^* \in \mathcal{X}^*. \tag{1.21}$$

It links the local optimality measure $\|\nabla f(\boldsymbol{x})\| = \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*)\|$ with the global optimality measure $\mathrm{dist}(\boldsymbol{x}, \mathcal{X}^*) = \min_{\boldsymbol{x}^*\in\mathcal{X}^*} \|\boldsymbol{x} - \boldsymbol{x}^*\|$, thus for any stationary point $\hat{x}$ in the basin we have $\mathrm{dist}(\hat{x}, \mathcal{X}^*) = 0$, i.e. $\hat{x}$ is a global optimum. The problem property (1.21) is related to "local strong convexity" but quite different. The relation is the following: f is strongly convex if (1.21) holds for arbitrary $\boldsymbol{x}, \boldsymbol{x}^*$; here we have restricted $\boldsymbol{x}$ to be close to $\boldsymbol{x}^*$, thus if there is only one global minimizer $\boldsymbol{x}^*$, we can

view (1.21) as the strong convexity of f relative to $\boldsymbol{x}^*$ in a local region. However, in our problem $\mathcal{X}^*$ contains infinite points and, in fact, it is a nonconvex set, thus it is not precise to say (1.21) reveals the local strong convexity of f.

The fact that $\mathcal{X}^*$ has infinite elements not only causes the conceptual difference from local convexity, but also results in the main difficulty of "coupled perturbation analysis" mentioned earlier. In fact, to prove Lemma 4.2.1 we need to find one point $(U, V) \in \mathcal{X}^*$, i.e. constructing a factorization $M = UV^T$, such that (1.20) holds. Using several probability tools including the random graph lemma, we can transform (1.20) to some simple conditions on U, V, then the task becomes to construct a factorization $M = UV^T$ so that U, V satisfy a few conditions including being close to X, Y respectively. Such a step is what we call the "coupled perturbation analysis".

We show that the success of many algorithms is due to the geometry of the problem. This viewpoint is different from other works [24,54–56] that interpreted alternating minimization as a perturbed version of the power method (though rely on resampling and is actually different from power method). Power method is known to converge to the global optimal solution in settings that are more general than eigenvector computation; see, e.g. [76–78]. Our geometric point of view covers a much broader range of algorithms; in fact, many algorithms, such as gradient descent and multi-block alternating minimization, perhaps cannot be viewed as perturbed versions of the power method.

Optimization Interpretation. Another interpretation of (1.21) is that it links the local optimality measure $\|\nabla f(x)\|$ with the global optimality measure $\mathrm{dist}(\boldsymbol{x}, \mathcal{X}^*) = \min_{\boldsymbol{x}^* \in \mathcal{X}^*} \|\boldsymbol{x} - \boldsymbol{x}^*\|$. From an optimization point of view, (1.21) is a quite natural step to connect stationary points and global optima. More specifically, standard first-order methods are actually solving the equation $\nabla f(\boldsymbol{x}) = 0$, instead of directly reducing the optimality gap $\mathrm{dist}(\boldsymbol{x}, \mathcal{X}^*)$, thus to show the global convergence we need to show $\nabla f(\boldsymbol{x}) = 0 \Longrightarrow \mathrm{dist}(\boldsymbol{x}, \mathcal{X}^*) = 0$. To achieve this, it suffices to show $\|\nabla f(\boldsymbol{x})\|^2 \geq c \cdot \mathrm{dist}(\boldsymbol{x}, \mathcal{X}^*)^2$ (or replace the exponent 2 by any positive number).

The relation (1.21) is closely related to the so-called "cost-to-go estimate" $\|\nabla f(\boldsymbol{x}_k)\|^2 \geq c_2[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$; see Lemma 4.2.3. Thus (1.21) can be used to prove the linear convergence of various methods, as long as "sufficient decrease" $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq c_1\|\nabla f(\boldsymbol{x}_k)\|^2$ can also be established; see Section 4.2.2.

The property (1.21) is related to, and quite different from, the error bound (a very

useful property that can leads to the linear convergence for *non-strongly* convex functions) [79] and Lojasiewicz inequality (a well-known inequality in real algebraic geometry, which has been widely used in optimization; see, e.g. [80]). In the simplest setting of unconstrained smooth optimization, the error bound and Lojasiewicz inequality both lower bound $\|\nabla f(\boldsymbol{x})\|$ by $\text{dist}(\boldsymbol{x}, \hat{\mathcal{X}})^\alpha$ or $(f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}))^\alpha$, where α is a positive constant, and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}$ denotes the set of *stationary points*. In contrast, we lower bound $\|\nabla f(\boldsymbol{x})\|$ by $\text{dist}(\boldsymbol{x}, \mathcal{X}^*)^\alpha$ or $(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)^\alpha$, where $\boldsymbol{x}^* \in \mathcal{X}^*$ denotes the set of *global optimizers*. Therefore, for non-convex problems the error bound and Lojasiewicz inequality can only be used to prove convergence to the set of stationary points, while our result (1.21) can be used to prove convergence to the set of global optimizers.

1.7 Organization and Notations

Organization. In Chapter 2, we will simulate and analyze several standard optimization algorithms for the traditional MF formulations from a new perspective. In Section 2.1, we observe that while the vanilla AltMin works well when the number of samples is much larger than the fundamental limit, it can fail when there are not that many samples. We show that the reason it fails is because it cannot control the row-norms (or incoherence constants) of the iterates. We show that AltMin for the regularized version can help control the row-norms (but not the norms) of the iterates. In Section 2.2, we analyzed the gradient methods (with constant stepsize or BB stepsize), and show that the LBB method (one type of the BB method) can converge to a quite accurate solution very fast. We found that the incoherence constants again play a key role in indicating whether the gradient methods succeed or fail. In Section 2.3, we investigate SGD and show that it is faster than GD and AltMin since it allows a very large stepsize compared to GD; however, a byproduct of the large stepsize is the divergence with a certain probability.

In Chapter 3, we present a new formulation that adds either penalty-function-type regularizers or constraints to control the incoherence constants. Preliminary simulation results show that the algorithms for the new formulation can significantly outperform the existing algorithms for the unregularized formulation or the formulation with regularizers $\lambda(\|X\|_F^2 + \|Y\|_F^2)$. We present several standard algorithms for solving this

formulation.

In Chapter 4, we establish the theory for the new formulation. We show that under similar conditions to those used in previous works, many standard optimization algorithms for this new formulation indeed converge to the global optima and recover the true low-rank matrix (see Theorem 4.2.1). Linear convergence of some algorithms can be proved with some extra effort (see Theorem 4.2.2). In Section 4.1, we formally define the problem formulation and four typical algorithms. In Section 4.2, we present the main results and the main lemmas used in the proofs for the main results. The proof of the two lemmas used in proving Theorem 4.2.1 are given in Section 4.3 and Section 4.4 respectively. The proof of the first lemma depends on two "coupled perturbation analysis" results Proposition 4.3.1 and Proposition 4.3.2, the proofs of which are given in Appendix A.3 and Appendix A.4 respectively. The proof of a lemma used in proving Theorem 4.2.2 is given in Appendix A.6.

Notations. Throughout the paper, $M \in \mathbb{R}^{m\times n}$ denotes the unknown data matrix we want to recover, and $r \ll \min\{m,n\}$ is the rank of M. The SVD of M is $M = \hat{U}\Sigma\hat{V}^T$, where $\hat{U} \in \mathbb{R}^{m\times r}, \hat{V} \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with diagonal entries $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_r$. We denote the maximum and minimum singular value as $\Sigma_{\max}$ and $\Sigma_{\min}$, respectively, and denote $\kappa \triangleq \Sigma_{\max}/\Sigma_{\min}$ as the condition number of M. Define $\alpha = m/n$, which is assumed to be bounded away from 0 and ∞ as $n \longrightarrow \infty$. Without loss of generality, assume $m \geq n$, then $\alpha \geq 1$.

Define the short notations $[m] \triangleq \{1,2,\ldots,m\}, [n] \triangleq \{1,2,\ldots,n\}$. Let $\Omega \subseteq [m]\times[n]$ be the set of observed positions, i.e. $\{M_{ij} \mid (i,j) \in \Omega\}$ is the set of all observed entries of M, and define $p \triangleq \frac{|\Omega|}{mn}$ which can be viewed as the probability that each entry is observed. For a linear subspace $\mathcal{S}$, denote $\mathcal{P}_{\mathcal{S}}$ as the projection onto $\mathcal{S}$. By a slight abuse of notation, we denote $\mathcal{P}_{\Omega}$ as the projection onto the subspace $\{W \in \mathbb{R}^{m\times n} : W_{i,j} = 0, \forall (i,j) \notin \Omega\}$. In other words, $\mathcal{P}_{\Omega}(A)$ is a matrix where the entries in Ω are the same as A while the entries outside of Ω are zero.

For a vector $x \in \mathbb{R}^n$, denote $\|x\|$ as its Euclidean norm. For a matrix X, denote $\|X\|_F$ as its Frobenius norm, and $\|X\|_2$ as its spectral norm (i.e. the largest singular value). Denote $\sigma_{\max}(X), \sigma_{\min}(X)$ as the largest and smallest singular values of X, respectively. Let $X^\dagger$ denote the pseudo inverse of a matrix X. The standard inner product between vectors or matrices are written as $\langle x,y\rangle$ or $\langle X,Y\rangle$, respectively. Denote $A^{(i)}$ as the

ith row of a matrix A. We will use C, C_1, C_T, C_d, etc. to denote universal numerical constants.

Chapter 2

A New Perspective on Standard Optimization Algorithms

In this chapter, we will discuss various standard optimization algorithms tailored for solving the matrix factorization formulations (1.10) and (1.13). We will argue that a crucial indicator of whether an algorithm succeeds in recovering the original matrix is the incoherence constants (or the row norms) of the iterates.

2.1 Alternating Minimization

Alternating minimization belongs to the class of block coordinate descent (BCD) type methods. While the original BCD algorithm cyclically updates each block of variables by solving the subproblem exactly, one can update the blocks in different orders (e.g. essentially cyclic [81], randomized [70] or parallel) and solve the subproblem inexactly. Commonly used inexact BCD type algorithms include BCGD (block coordinate gradient descent, which updates each variable by a single gradient step [70]) and BSUM (block successive upper bound minimization, which updates each variable by minimizing an upper bound of the objective function [82]). BCD-type methods have been widely used in engineering (e.g. [83–85]). In the context of matrix completion, reference [73] proposed an algorithm that could be viewed as a BSUM algorithm. Just considering different choices of the blocks will lead to different algorithms for the matrix completion problem [65]. Different versions of BCD algorithms may have quite different computational time

and recovery ability; however, for simplicity, we only consider the two-block alternating minimization method (i.e. AltMin) in this chapter.

2.1.1 AltMin for Unregularized Formualtion

AltMin, in the context of matrix completion, usually refers to the algorithm that alternates between X and Y by updating one factor at a time with the other factor fixed. Although the overall objective function is non-convex, each subproblem of X or Y is convex and thus can be solved efficiently. The details are given in Table 4.3.

Table 2.1: AltMin (Two-block Alternating Minimization)

Initialization: randomly generate (X_0, Y_0).
The k-th iteration:
$X_k \longleftarrow \arg\min_X F(X, Y_{k-1})$,
$Y_k \longleftarrow \arg\min_Y F(X_{k-1}, Y)$.

The objective function $F(X, Y)$ is quadratic with respect to X or Y and updating X_k, Y_k can be done in closed form. In fact, $\arg\min_X F(X, Y)$ is the solution to the equation $\nabla_X F(X, Y) = 0$, i.e. $\mathcal{P}_\Omega(XY^T - M)^T X = 0$, and $\arg\min_Y F(X, Y)$ is the solution to $\mathcal{P}_\Omega(XY^T - M)Y = 0$. By some algebraic computation, one can show that $\arg\min_X F(X, Y)$ and $\arg\min_Y F(X, Y)$ can be computed in closed form. Specifically, suppose $X^T = (x_1, \ldots, x_m)$ and $Y^T = (y_1, \ldots, y_n)$, where $x_i, y_j \in \mathbb{R}^{r\times 1}$. Then $(x_1^*, \ldots, x_m^*) \triangleq (\arg\min_X F(X, Y))^T$ and $(y_1^*, \ldots, y_n^*) \triangleq (\arg\min_Y F(X, Y))^T$ are given by

$$\begin{aligned} x_i^* &= (\sum_{j\in\Omega_i^x} y_j y_j^T)^\dagger (\sum_{j\in\Omega_i^x} M_{ij} y_j),\ i = 1, \ldots, m, \\ y_j^* &= (\sum_{i\in\Omega_j^y} x_i x_i^T)^\dagger (\sum_{i\in\Omega_j^y} M_{ij} x_i),\ j = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where $\Omega_i^x = \{j \mid (i, j) \in \Omega\}, \Omega_j^y = \{i \mid (i, j) \in \Omega\}$, and $A^\dagger$ denotes the pseudo inverse of a matrix A.

To compute one x_i^* according to (4.15), the memory requirement is $O(|\Omega_i^x| + r^2)$, and the computation time can be calculated as follows. The time for computing $\sum_{j\in\Omega_i^x} y_j y_j^T$ is $r^2|\Omega_i^x|$ (it takes r^2 time to form each matrix $y_j y_j^T$, and $r^2(|\Omega_i^x| - 1)$ time to sum these matrices up); the time for computing $\sum_{j\in\Omega_i^x} M_{ij} y_j$ is $2r|\Omega_i^x|$; computing $A^{-1}v$ for

$A \in \mathbb{R}^{r\times r}$ and $v \in \mathbb{R}^{r\times 1}$ roughly takes time $O(r^3)$; adding them together, the time cost for computing one x_i^* is $O(r^3 + r^2|\Omega_i^x|)$. To compute all x_i^*, the memory requirement is $O(|\Omega| + mr^2)$ and the time cost is $O(mr^3 + r^2|\Omega|)$. Similarly, to compute all y_j^*, the memory requirement is $O(|\Omega| + nr^2)$ and the time cost is $O(nr^3 + r^2|\Omega|)$. Therefore, for each iteration of AltMin, the memory requirement is $O(|\Omega|+(m+n)r^2)$ and the time cost is $O((m+n)r^3 + 2r^2|\Omega|)$. For a reasonable recovery, the number of observations should be at least the degrees of freedom of a rank-r matrix, which is about $O(mr)$ (assume $m \geq n$), then $2r^2|\Omega| \geq 2mr^3 \geq (m+n)r^3$. Thus the time cost for each iteration of AltMin is roughly $O(|\Omega|r^2) \geq O(nr^3)$.

If we view r as $O(1)$ and $|\Omega|$ as $O(n)$, then each iteration of AltMin requires memory space of size $O(n)$ and computation time $O(n)$, which are much better than the memory requirement of $O(n^2)$ and computation time of $O(n^3)$ for each iteration of first order methods for the nuclear norm formulation. In practice, the size of r matters, and the $O(r^2|\Omega|)$ time for solving linear systems of equations might still be a bit expensive in certain scenarios. Thus one may consider even cheaper iterations such as the gradient step; more discussions are provided later.

For numerical experiments, we consider a randomly generated rank-10 1000×1000 matrix M, and try to recover it from p-portion of entries, where $p = 0.1$ and $p = 0.05$ respectively. The original matrix M is generated by $M = UV^T$ where $U, V \in \mathbb{R}^{1000\times10}$ has independent Gaussian entries with zero mean and variance $1/1000$, and Ω is generated by a Bernolli model where each entry of M is included in Ω with probability p. We consider random initial point $X_0, Y_0 \in \mathbb{R}^{1000\times10}$ with independent Gaussian entries (also zero mean and variance $1/1000$). The figures are obtained by averaging 100 Monte Carlo runs.

Note that Keshavan [24] has extensively studied many algorithms for the setting $m = n = 1000, r = 10$, and many figures in this thesis can be compared with the simulation results of [24] (such as Figure 6.1 there). However, to the best of our knowledge, all previous works, including [24], do not investigate why AltMin (and other algorithms) succeed or fail, which is the focus of our investigation in this chapter.

We use RMSE (Root Mean Square Error) as a performance measure. The training error for an estimated matrix $\hat{M}$ indicates RMSE for the training samples (i.e. the

observed entries), and is defined as [1]

$$\text{RMSE}_{\text{train}} = \|\mathcal{P}_\Omega(M - \hat{M})\|_F / \|\mathcal{P}_\Omega(M)\|_F. \tag{2.2}$$

The test error indicates RMSE for the whole matrix, and is defined as

$$\text{RMSE}_{\text{test}} = \|M - \hat{M}\|_F / \|M\|_F. \tag{2.3}$$

In practical scenarios that the full matrix is unknown, one may use another set of samples as the test set. In our definition, the training error is scaled by $1/\|\mathcal{P}_\Omega(M)\|_F$ and test error is scaled by $1/\|M\|_F$, therefore for random M and random $\hat{M}$ the two RMSE measures $\text{RMSE}_{\text{train}}$ and $\text{RMSE}_{\text{test}}$ are roughly the same (not a rigourous statement, but just based on intuition and verified by experiments).

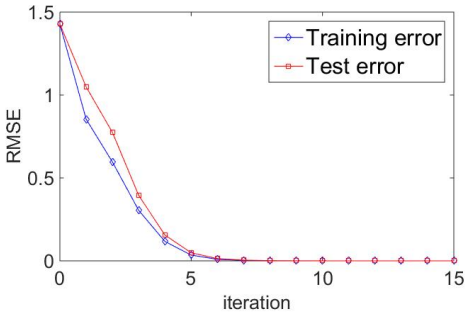


Figure 2.1: $m = n = 1000, r = 10, p = 0.1$

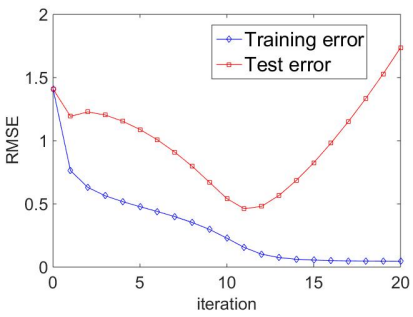


Figure 2.2: $m = n = 1000, r = 10, p = 0.05$

AltMin performs quite well when there are enough samples, but fails when there are not too many samples. In particular, Figure 2.1 shows that when there are enough samples ($p = 0.1$), AltMin converges in fewer than 10 iterations, and the training error and the test error both converge to 0. Note that zero test error means exact recovery of the original matrix, thus AltMin can recover the original matrix in fewer than 10 iterations. When there are fewer observations ($p = 0.05$), Figure 2.2 shows that AltMin still reduces the training error to a small amount, but the test error can diverge. Figure 2.1 and Figure (2.2) can be viewed as the case of "many samples" and the case of "not enough samples" respectively, and we can summarize the performance of AltMin as follows.

[1] Note that the traditional definition of RMSE is slightly different; for example, $\text{RMSE}_{\text{train}}$ is defined as $\|\mathcal{P}_\Omega(M - \hat{M})\|_F / \sqrt{mn}$, which differs from our definition by a constant ratio. However, if one multiplies all entries of M by 100, such defined RMSE also increases by 100 times. In contrast, our definition of RMSE can be viewed as the relative error and is scaling-invariant.

Observation 2.1.1 *For p large enough, AltMin converges to the global optimizer of* (1.10) *and recovers the underlying matrix M with probability close to* 1. *For p that is moderately large (say, between $4\max\{r, \log n\}/n$ and $8\max\{r, \log n\}/n$), AltMin may generate a sequence such that the training error converges to a neighborhood of* 0 *and the test error is larger than* 1.

One may explain the large gap between the training error and test error in Fig. 2.2 by overfitting: when there are not enough data, the iterates X_k, Y_k tend to fit the observation set too well, resulting in a larger test error. However, we will argue that the gap in 2.2 is *not* due to overfitting, but a more subtle issue as discussed below. Averaging might hide important information, so we take a closer look at each implementation of 10 experiments: it turns out that in 3 realizations the training error is below 10^{-4} while the test error is also below 10^{-4}; in other 7 realizations the training error is between 0.05 and 0.13 while the (final) test error becomes larger than 2; see the first two rows of the table in Figure 2.3. There seems to be a phase transition region of training error v.s. test error:

Observation 2.1.2 *(For p not too small) When applying AltMin, if the training error is very small ($< 1/n$), then the test error is also very small ($< 1/n$); if the training error is greater than c/n (say, $10/n$), then the test error can be larger than* 1.

The above observation implies that the gap between the training error and the test error in Fig. 2.2 is not because of overfitting, but rather because of "under-fitting"; if the training error is small enough, then the test error will be small enough as well. This observation is also related to the identifiability issue: it implies that the test error will go to zero as the training error goes to 0, which mean that M can be uniquely determined by $\mathcal{P}_\Omega$. Therefore, this observation can only be true in the "identifiable" region. We guess that this observation is true for any p in the "identifiable region" that $p > 2.5\max\{r, \log n\}/n$ (the coefficient 2.5 may not be accurate).

Observation (2.1.2) implies that solving (1.10) very accurately can recover the original matrix. From an algorithmic point of view, this implication seems to be useless: how can we control the accuracy of the solution obtained by solving a non-convex problem? An algorithm such as AltMin may be stuck at a training error of 0.02, how can we push the training error further down? To answer these questions, we should further explore

the reason why AltMin fails. A common guess is that AltMin has no control over the norms of X and Y, thus $\|X\|_F$ or $\|Y\|_F$ may diverge and AltMin may not perform well. We do find that $\|X\|_F\|Y\|_F$ in the successful instances are at least twice smaller than $\|X\|_F\|Y\|_F$ in the failed instances (see Figure 2.3). However, it is not entirely clear why the norm of X and Y should matter: what we care about is XY^T, not the individual factors X, Y.

Inspired by the incoherence condition [2], we guess that the true reason of the failure of AltMin is that the estimated matrix $\hat{M} = XY^T$ is not incoherent. Recall that the incoherence condition in [2] contains two parts: the row-norms of U, V are upper bounded where U, V are SVD factors of M, and each entry of M is upper bounded (also called spikeness in [32]). Here we do not consider the SVD factors of XY^T, but X and Y themselves. The maximum row-norm of a matrix $X \in \mathbb{R}^{K\times r}$ (we will call it "max-norm" for short in some places) is defined as

$$\|X\|_{2,\infty} = \max_{1\le i\le K} \|X^{(i)}\|. \tag{2.4}$$

We define the "spikeness" of a matrix $Z \in \mathbb{R}^{m\times n}$ as the maximum magnitude of its entries (similar to [32] and related to the second incoherence condition of [2]), i.e.

$$\|Z\|_\infty = \max_{i,j} |Z_{ij}|. \tag{2.5}$$

A small spikeness indicates that the entries of Z are evenly spread out. The max-norm of X, Y and the spikeness of XY^T are related to the incoherence constants, and we will discuss the incoherence constants later.

In Figure 2.3, we compare the test error, the training error, $\|X\|_F$, $\|Y\|_F$, spikeness of XY^T, max-norm of X, max-norm of Y. Column 1-3 indicate successful instances (small test error; can recover M), and column 8-10 indicate failed instances (large test error; cannot recover M). In all failed instances, the training error is at least 0.04, but in all successful instances the training error is smaller than 6×10^{-5}. This table shows that all the successful instances have smaller $\|X\|_F\|Y\|_F$ than the failed instances, and have much smaller training error, spikeness $\|XY^T\|_\infty$ and $\|X\|_{2,\infty}\|Y\|_{2,\infty}$ than failed instances. To see the difference more clearly, we average the quantities for the successful and failed instances and summarize the results in Table 2.1.1.

To obtain a quantitative comparison of successful and failed instances, we scale all quantities (except the training and test errors) by the corresponding quantities in

	1	2	3	4	5	6	7	8	9	10
Test error	1.2944e-04	9.9424e-05	1.0676e-04	2.1798	4.3319	5.1742	4.0432	2.0162	5.7910	3.1831
Training error	5.1353e-05	3.9386e-05	4.2138e-05	0.0531	0.0550	0.0794	0.0532	0.0440	0.0701	0.0735
Norm of X	2.0025	2.0150	2.0366	5.0610	4.6260	5.7422	8.1383	2.0594	4.1578	6.4534
Norm of Y	5.0591	5.0379	4.9793	4.9153	11.6634	13.1261	5.0005	11.5621	15.2944	5.2547
Spikeness	0.0224	0.0224	0.0224	0.7081	10.1310	12.4060	1.0852	0.6862	13.5670	0.8320
Max-Norm of X	0.1155	0.1147	0.1098	4.6159	4.1631	4.4401	5.9658	0.1099	3.6348	5.4330
Max-Norm of Y	0.3270	0.3131	0.2996	0.3225	10.5098	12.0917	0.3396	10.4355	10.3589	0.3315

Figure 2.3: Comparison of 10 experiments of AltMin for $m = n = 1000$, $r = 10$, $p = 0.05$. Column 1-3 indicate successful instances (small test error; can recover M), and column 8-10 indicate failed instances (large test error; cannot recover M). This table shows that all the successful instances have much smaller training error, spikeness $\|XY^T\|_\infty$ and the product of max-norms $\|X\|_{2,\infty}\|Y\|_{2,\infty}$ than failed instances. See an "averaged version" of this table in Table 2.1.1.

	Succeed	Fail
Average test error	1.1e-4	3.8
Average training error	4.4e-5	0.06
Average $\max\{\|X\|_F, \|Y\|_F\}$	5.0	9.6
Average $\max_i\{\|X^{(i)}\|, \|Y^{(i)}\|\}$	0.3	6.3

Table 2.2: Comparison of successful and failed instances for AltMin. This table is an "averaged version" of the table in Figure 2.3.

Column 1 (the successful instance), to obtain a table in Figure 2.4. In all failed instances, either the max-norm of X or the max-norm of Y or both are 30 times larger than the corresponding term of the successful instances; as a consequence, $\|X\|_{2,\infty}\|Y\|_{2,\infty}$ is at least $30 \approx O(\sqrt{n})$ times larger than that of the successful instances. The fact that the ratio is around $\sqrt{n}$ is reasonable: if X is highly unbalanced in the sense that its energy is concentrated in one row (or very few rows), then $\|X\|_{2,\infty} \approx \|X\|_F$; if X is balanced that each row has almost the same row norm, then $\|X\|_{2,\infty} \approx \|X\|_F/\sqrt{n}$, which is $\sqrt{n}$ times smaller than the unbalanced case. A graphical comparison of X with small max-norm and large max-norm is given in Figure 2.5. We further check the norms of all rows of X with a large max-norm, and find that only one of the row norms is larger than 1 (which determines the max-norm) and all other row norms are approximately $\|X\|_{2,\infty}/30 \approx \|X\|_{2,\infty}/\sqrt{n}$; such a phenomenon also happens for Y.

We summarize the above findings in the following:

Observation 2.1.3 *In all failed instances, at least one of X and Y has a max-norm*

	1	2	3	4	5	6	7	8	9	10
Test error	1.2944e-04	9.9424e-05	1.0676e-04	2.1798	4.3319	5.1742	4.0432	2.0162	5.7910	3.1831
Training error	5.1353e-05	3.9386e-05	4.2138e-05	0.0531	0.0550	0.0794	0.0532	0.0440	0.0701	0.0735
Norm of X, scaled	1	1.0063	1.0170	2.5274	2.3102	2.8676	4.0642	1.0284	2.0764	3.2227
Norm of Y, scaled	1	0.9958	0.9842	0.9716	2.3055	2.5946	0.9884	2.2854	3.0232	1.0387
Spikeness, scaled	1	1.0001	1.0001	31.6771	453.2070	554.9726	48.5455	30.6957	606.9364	37.2183
Max-Norm of X, scaled	1	0.9931	0.9511	39.9691	36.0482	38.4467	51.6575	0.9519	31.4739	47.0442
Max-Norm of Y, scaled	1	0.9574	0.9163	0.9863	32.1400	36.9777	1.0387	31.9129	31.6788	1.0136

Figure 2.4: A scaled version of the table in Figure 2.3, where all quantities in Column 2-10 (except the training and test errors) are scaled by the corresponding quantities in Column 1. Column 1-3 indicate successful instances, and column 8-10 indicate failed instances. This table shows that the ratio between a quantity of the successful instance and the corresponding quantity is at least $30 \approx \sqrt{n}$.

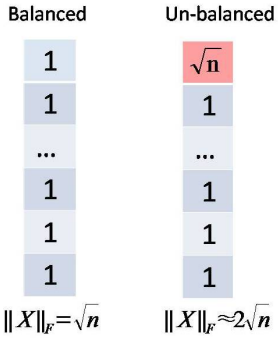


Figure 2.5: Illustration of balanced and unbalanced tall matrices. In balanced matrices, each row-norm is approximately 1; in unblanced matrices, there is one row-norm that is much larger than all other row-norms.

(i.e. maximum row-norm) of order $O(\sqrt{n})$*; in all successful instances, both* X *and* Y *have a max-norm of order* $O(1)$*. In addition, when* X *(or* Y*) has a large row-norm, only one (or very few) row-norm is of order* $O(\sqrt{n})$ *and all other row-norms are of order* $O(1)$*.*

Note that it is not precise to say X has a max-norm of order $O(1)$ when success; the right order should be $O(\frac{\sqrt{r}}{\sqrt{n}}\sqrt{\log n})$ which is the max-norm of the generative factors U, V. For our setting $O(\frac{\sqrt{r}}{\sqrt{n}}\sqrt{\log n})$ is approximately 0.1, which can be verified by Column 1-3 of Figure (2.4). Therefore, $O(\sqrt{n})$ and $O(1)$ in Observation (2.1.4) should be understood as the size scaled by the max-norm of the generative factors U, V.

Observation (2.1.4) helps explain quantitatively the difference of success/failure for other quantities in Figure 2.4. The spikeness of XY^T for all failed instances is at least $30 \approx \sqrt{n}$ times larger than the spikeness for successful instances; moreover, if

both $\|X\|_{2,\infty}$ and $\|Y\|_{2,\infty}$ are of order $O(\sqrt{n})$, then the spikeness of XY^T is of order $O(n) \approx O(1000)$. To explain this phenomenon, let $x_i = X^{(i)}, y_j = Y^{(j)}$ denote the rows of X and Y, $x_{i,\max} \in \mathbb{R}^{1\times r}$ and $y_{j,\max} \in \mathbb{R}^{1\times r}$ denote the row of X and the row of Y with the maximum row norm respectively, and θ denote the angle between $x_{i,\max}$ and $y_{j,\max}$. Since $\|XY^T\|_\infty = \max_{i,j} |x_i^T y_j| \approx x_{i,\max}^T y_{j,\max} = \|x_{i,\max}\|\|y_{j,\max}\| \cos(\theta)$ [2] , and $\|X\|_{2,\infty}\|Y\|_{2,\infty} = \|x_{i,\max}\|\|y_{j,\max}\|$, we have

$$\frac{\|XY^T\|_\infty}{\|X\|_{2,\infty}\|Y\|_{2,\infty}} \approx \cos(\theta) \leq 1.$$

This explains why in Figure 2.4 the scaled spikeness is smaller than the product of the scaled max-norms of X and Y. We also observe that the ratio $\frac{\|XY^T\|_\infty}{\|X\|_{2,\infty}\|Y\|_{2,\infty}}$ in Figure 2.4 (divide the number in the third last row by the product of the numbers in the last two rows) is always above 1/2. This is reasonable since if θ is a generic number in $[0, \pi/2]$, then one would expect $\cos(\theta)$ is larger than, say, $\cos(\pi/3) = 1/2$, which implies $\frac{\|XY^T\|_\infty}{\|X\|_{2,\infty}\|Y\|_{2,\infty}} \geq \frac{1}{2}$ [3]

Having argued that the large spikeness is due to the large max-norms of X and Y, we argue that the large norms of X or Y in failed instances are also due to the large max-norms of X or Y. In fact, it can be seen from Figure 2.3 that when the norm of X is large, we have $\|X\|_{2,\infty} \leq \|X\|_F \leq 1.5\|X\|_{2,\infty}$ (the same for Y). If the large norm of X is because of many row-norms being larger than usual, then we would not observe the phenomenon that one row norm is close to the norm of X.

In the end of this subsection, we show another version of the table in Figure 2.3, replacing the last three rows by the incoherence constants. In the original table, the meaning of the quantities in the last three rows (spikeness, maximum row-norm) is not clear without comparison; the goal of the new table is to replace them by relative quantities. Following the notion of incoherence in [2], we define the incoherence of a

[2] We have checked our numerical results and find that the relation $\max_{i,j} |x_i^T y_j| = |x_{i,\max}^T y_{j,\max}|$ is precise for all failed instances we encountered. Theoretically speaking, the approximation $\max_{i,j} |x_i^T y_j| \approx |x_{i,\max}^T y_{j,\max}|$ is not always true, since there may exist another pair (x_i, y_j) such that $\|x_i\| < \|x_{i,\max}\|$ or $\|y_j\| < \|y_{j,\max}\|$ but $x_i^T y_j > x_{i,\max}^T y_{j,\max}$. However, this does not happen in our experiments since, as indicated in Observation 2.1.4, other $\|x_i\|, \|y_j\|$ are much smaller than $\|x_{i,\max}\|$ and $\|y_{j,\max}\|$.

[3] This explanation is not rigorous: of course θ is not a random angle, and we cannot prove $\theta > \pi/3$ at this point.

matrix $Z \in \mathbb{R}^{m\times n}$ as the ratio of the "spikeness" and the average magnitude, i.e.

$$\mu(Z) \triangleq \sqrt{mn}\frac{\max_{i,j}|Z_{ij}|}{\|Z\|_F} = mn\frac{\|Z\|_\infty}{\|Z\|_F}, \tag{2.6}$$

and define the incoherence of $X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}$ as the ratio of the square of the maximum row-norm and the square of the norm, i.e.

$$\begin{aligned}\mu(X) &\triangleq m\frac{\max_{1\le i\le K}\|X^{(i)}\|^2}{\|X\|_F^2} = m\frac{\|X\|_{2,\infty}^2}{\|X\|_F^2},\\ \mu(Y) &\triangleq m\frac{\max_{1\le i\le K}\|X^{(i)}\|^2}{\|X\|_F^2} = n\frac{\|Y\|_{2,\infty}^2}{\|Y\|_F^2}.\end{aligned} \tag{2.7}$$

Note that the definition of $\mu(X)$ and $\mu(Y)$ in (2.7) involve square of the norms, while the definition of $\mu(Z)$ in (2.6) does not; this is consistent with the definition of incoherence constants in [2].

We have the following observation based on Fig. 2.4, which describes quantitatively the incoherence constants in successful and failed instances.

Observation 2.1.4 *In all successful instances, the incoherence constants of X and Y are below 4.5, and the incoherence constant of XY^T is below 7.5. In all failed instances, at least one of the incoherence constants of X and Y is above 500, and the incoherence constant of XY^T is at least 200.*

	1	2	3	4	5	6	7	8	9	10
Test error	1.2944e-04	9.9424e-05	1.0676e-04	2.1798	4.3319	5.1742	4.0432	2.0162	5.7910	3.1831
Training error	5.1353e-05	3.9386e-05	4.2138e-05	0.0531	0.0550	0.0794	0.0532	0.0440	0.0701	0.0735
Norm of X	2.0025	2.0150	2.0366	5.0610	4.6260	5.7422	8.1383	2.0594	4.1578	6.4534
Norm of Y	5.0591	5.0379	4.9793	4.9153	11.6634	13.1261	5.0005	11.5621	15.2944	5.2547
Incoherence of XY'	7.1671	7.1677	7.1680	227.0346	3.2483e+03	3.9777e+03	347.9451	220.0013	4.3500e+03	266.7497
Incoherence of X	3.3267	3.2402	2.9067	831.8406	809.8833	597.9004	537.3658	2.8478	764.2472	708.7650
Incoherence of Y	4.1778	3.8625	3.6203	4.3049	811.9673	848.6006	4.6122	814.6163	458.7355	3.9799

Figure 2.6: Another version of the table in Figure 2.3, where the last three rows are replaced by the corresponding incoherence constants.

2.1.2 AltMin for Regularized Formualtion

It is reasonable to add a regularizer $\|X\|_F^2 + \|Y\|_F^2$ to control the norms of X and Y; however, as we will see later, the reason why it helps is not because it controls the norms

of the iterates (in fact, it does not!), but it can control the row-norms of the iterates, though via an unknown mechanism.

Using AltMin to solve the regularized formulation (1.13) is quite similar to the original AltMin, and we denote this algorithm as AltMinReg; see details in Table 2.3.

Table 2.3: AltMinReg

Initialization: randomly generate (X_0, Y_0) .
The k-th iteration:
$X_k \longleftarrow \arg\min_X \ F(X, Y_{k-1}) + \lambda\|X\|_F^2,$
$Y_k \longleftarrow \arg\min_Y \ F(X_{k-1}, Y) + \lambda\|Y\|_F^2.$

Similar to AltMin, for AltMinReg each subproblem can still be solved in closed form. In fact, $(x_1^*, \ldots, x_m^*) \triangleq (\arg\min_X F(X,Y))^T$ and $(y_1^*, \ldots, y_n^*) \triangleq (\arg\min_Y F(X,Y))^T$ are given by

$$\begin{aligned} x_i^* &= (\sum_{j\in\Omega_i^x} y_j y_j^T + 2\lambda I)^\dagger (\sum_{j\in\Omega_i^x} M_{ij} y_j),\ i = 1, \ldots, m, \\ y_j^* &= (\sum_{i\in\Omega_j^y} x_i x_i^T + 2\lambda I)^\dagger (\sum_{i\in\Omega_j^y} M_{ij} x_i),\ j = 1, \ldots, n, \end{aligned} \tag{2.8}$$

For a given λ, the computation cost and the memory space for AltMinReg are almost the same as those for original AltMin. Nevertheless, it may take extra time to pick a good parameter λ (usually by cross validation).

While the SVD factors of the true matrix M form one global optimal solution of (1.10), they may not be the global optimal solution of (1.13). The global optimal value of (1.13) can differ from the global optimal value of (1.10), which is 0, by an amount proportional to λ. More precisely, suppose (U, V) is a global optimal solution of $F(X,Y) = \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2$, i.e. $UV^T = M$, and $(\hat{X}, \hat{Y})$ is a global optimal solution of $F_1(X,Y) = F(X,Y) + \lambda(\|X\|_F^2 + \|Y\|_F^2)$. Due to the optimality of $(\hat{X}, \hat{Y})$, we have

$$F_1(\hat{X}, \hat{Y}) \le F_1(U, V) = \lambda(\|U\|^2 + \|V\|^2),$$

which implies

$$F(\hat{X}, \hat{Y}) \le \lambda(\|U\|^2 + \|V\|^2) - \lambda(\|\hat{X}\|_F^2 + \|\hat{Y}\|_F^2).$$

This is true for all U, V such that $UV^T = M$, then we can pick $U = \hat{U}\Sigma^{1/2}, V = \hat{V}\Sigma^{1/2}$

(assuming $M = \hat{U}\Sigma\hat{V}^T$ is the SVD), and the above relation becomes

$$F(\hat{X}, \hat{Y}) \leq \lambda(2\|M\|_* - \|\hat{X}\|_F^2 - \|\hat{Y}\|_F^2).$$

The training error corresponding to $(\hat{X}, \hat{Y})$ is then bounded as

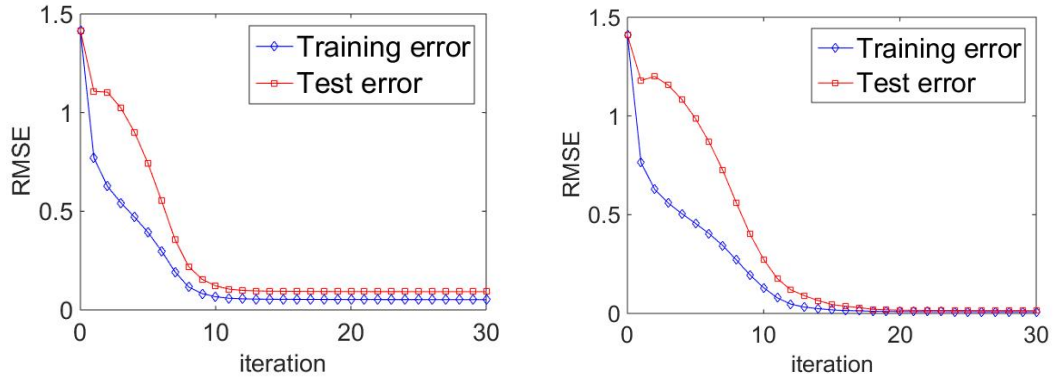
$$\text{RMSE}_{\text{train}}^2 = \frac{F(\hat{X}, \hat{Y})}{\|\mathcal{P}_\Omega M\|_F^2/2} = 2\lambda \frac{1}{\|\mathcal{P}_\Omega M\|_F^2}(2\|M\|_* - \|\hat{X}\|_F^2 - \|\hat{Y}\|_F^2). \tag{2.9}$$

The above relation is a necessary condition for $(\hat{X}, \hat{Y})$ to be a global optimizer of the regularized function, and can be used to test the global optimality of $(\hat{X}, \hat{Y})$. This criterion can help understand how well the regularized function has been solved.

We consider the same setting as before, i.e. the original matrix M is generated by $M = UV^T$ where $U, V \in \mathbb{R}^{1000\times 10}$ has independent Gaussian entries with zero mean and variance $1/1000$, and Ω is generated by a Bernolli model with parameter p. We consider random initial point $X_0, Y_0 \in \mathbb{R}^{1000\times 10}$ with independent Gaussian entries (also zero mean and variance $1/1000$). The figures are obtained by averaging 100 Monte Carlo runs.

Fig. 2.7 shows the performance of AltMinReg for $p = 0.05$, under two choices of λ: $\lambda = 10^{-3}$ and $\lambda = 10^{-4}$. It can be seen from the figure that for both choices of λ the training error and the test error both converge to a small number. This greatly improves the success probability compared to the unregularized AltMin (the success probability of AltMin for $p = 0.05$ is between 20% and 30%). We summarize the finding below:

Observation 2.1.5 *Regularizers $\lambda(\|X\|_F^2 + \|Y\|_F^2)$ can be very helpful for AltMin even in the noiseless case. In particular, in a certain setting the probability of successful recovery can be increased from less than 30% to more than 99% by using AltMinReg.*



(a) $\lambda = 10^{-3}$ (b) $\lambda = 10^{-4}$

Figure 2.7: Performance of AltMinReg, i.e. AltMin for the MF formulation with the regularizer $\lambda(\|X\|_F^2 + \|Y\|_F^2)$. $m = n = 1000, r = 10, p = 0.05$.

We present various quantities of the convergent solutions for 10 experiments in Fig. 2.8 and Fig. 2.9. As shown in the two tables, the incoherence constants of X are all between 2.8 and 3.2, and the incoherence constants of Y are all between 3.4 and 3.8, no matter what value λ is. The incoherence constant of XY^T is around 7.5 for $\lambda = 10^{-3}$ and around 6.5 for $\lambda = 10^{-4}$, both of which are rather small. The choice of $\lambda = 10^{-3}$ leads to a much larger test error than $\lambda = 10^{-4}$, because a large λ leads to a large distortion of the optimal solution.

	1	2	3	4	5	6	7	8	9	10
Test error	0.0915	0.0930	0.0914	0.0922	0.0920	0.0931	0.0928	0.0936	0.0912	0.0948
Training error	0.0512	0.0520	0.0515	0.0516	0.0514	0.0521	0.0518	0.0517	0.0516	0.0526
Norm of X	3.0278	3.0266	3.0275	3.0265	3.0273	3.0260	3.0270	3.0270	3.0278	3.0249
Norm of Y	3.0466	3.0441	3.0458	3.0454	3.0454	3.0442	3.0444	3.0438	3.0458	3.0419
Incoherence of XY'	7.7553	7.5939	7.5456	7.7787	7.5689	7.9267	7.5767	7.1996	8.0774	7.8672
Incoherence of X	2.9455	2.8680	2.9692	2.8677	2.8352	2.9873	2.8778	2.9454	2.8444	2.9212
Incoherence of Y	3.5767	3.5393	3.3341	3.5501	3.3557	3.4887	3.6498	3.2891	3.6372	3.3721

Figure 2.8: Various quantities related to the convergent solution of AltMinReg with $\lambda = 10^{-3}$ in 10 experiments. $m = n = 1000, r = 10, p = 0.05$. The incoherence constants of X and Y are all below 4, and the incoherence constant of XY^T is around 7.5.

We claim that AltMinReg improves upon AltMin because AltMinReg controls the incoherence constants, not because it controls the norms of X and Y. To validate this claim, we give an example that the norms of X and Y are well controlled but still the

	1	2	3	4	5	6	7	8	9	10
Test error	0.0148	0.0142	0.0145	0.0143	0.0146	0.0135	0.0137	0.0143	0.0144	0.0147
Training error	0.0085	0.0081	0.0083	0.0082	0.0083	0.0080	0.0080	0.0081	0.0084	0.0083
Norm of X	2.4263	2.4535	2.4258	2.4474	2.4451	2.4346	2.4505	2.4669	2.4296	2.4284
Norm of Y	4.1335	4.0786	4.1266	4.0909	4.0965	4.1090	4.0869	4.0594	4.1223	4.1217
Incoherence of XY'	6.4606	6.3559	6.4080	6.4563	6.3542	6.4766	6.4801	6.4303	6.4218	6.4518
Incoherence of X	2.8082	2.9703	2.9355	2.7108	2.8412	2.8920	2.7072	3.1340	3.0760	2.8152
Incoherence of Y	3.4147	3.6948	3.7559	3.5977	3.5766	3.3326	3.6020	3.6931	3.4536	3.5964

Figure 2.9: Various quantities related to the convergent solution of AltMinReg with $\lambda = 10^{-4}$ in 10 experiments. $m = n = 1000, r = 10, p = 0.05$. The incoherence constants of X and Y are all below 4, and the incoherence constant of XY^T is around 6.5.

test error is high. We run 10 experiments for $\lambda = 3 \times 10^{-5}$, and record the results for each experiment in the table of Fig. 2.10 and the average results for success/failure in Table 2.1.2. In Figure 2.10, Column 1-4 indicate failed instances (test error > 0.3),[4] and Column 5-10 indicate successful instances (test error < 0.006). The norms of X, Y for the failed instances are well controlled and almost indistinguishable from the norms of X, Y for the successful instances, thus the norms of X, Y are not indicators of success. In contrast, the incoherence constants for failed and successful instances are hugely different. More specifically, either $\mu(XY^T)$ or $\max\{\mu(X), \mu(Y)\}$ is much larger in failed instances than the corresponding quantity in successful instances.

	1	2	3	4	5	6	7	8	9	10
Test error	0.8328	0.6002	0.5043	0.4408	0.3353	0.0055	0.0052	0.0055	0.0053	0.0060
Training error	0.0905	0.0645	0.0662	0.0603	0.0343	0.0032	0.0030	0.0032	0.0031	0.0034
Norm of X	2.6698	2.4648	2.3075	2.2502	2.2255	2.1505	2.2318	2.1640	2.1856	2.1665
Norm of Y	4.7215	4.7212	4.8040	4.8159	4.7384	4.7235	4.5554	4.6680	4.6435	4.6999
Incoherence of XY'	117.9507	39.4705	37.8289	126.0118	29.9190	7.1809	7.1942	7.2044	7.1634	7.1984
Incoherence of X	183.7334	138.6052	117.8811	84.2748	72.2548	3.0860	2.8824	2.8230	3.2197	2.7691
Incoherence of Y	29.6985	3.0718	2.8786	41.6899	2.8603	3.1091	2.7992	2.8436	3.0263	2.8096

Figure 2.10: Results for AltMinReg with $\lambda = 3 \times 10^{-5}$ in 10 experiments. $m = n = 1000, r = 10, p = 0.05$. Column 1-5 indicate failed instances, and Column 6-10 indicate successful instances. The norms of X, Y for the failed and successful instances are rather close, but the incoherence constants for failed and successful instances are hugely different. See Table 2.1.2 for an averaged version of this table.

We summarize the observations obtained from Figure 2.10 and Table 2.1.2 below.

Observation 2.1.6 *When applying AltMinReg, the norms of X, Y for the failed instances are almost indistinguishable from the norms of X, Y for the successful instances. In contrast, either $\mu(XY^T)$ or $\max\{\mu(X), \mu(Y)\}$ is much larger in failed instances than the corresponding quantities in successful instances.*

[4] Note that in some applications the test error 0.3 might be acceptable, but since a much smaller test error < 0.01 can be achieved, we still view a test error of 0.3 as a failure.

	Succeed	Fail
Average test error	5.5e-3	0.54
Average $\max\{\|X\|_F, \|Y\|_F\}$	4.65	4.76
Average $\mu(XY^T)$	7.2	70.2
Average $\mu(X)$	2.96	120.3
Average $\mu(Y)$	2.92	16

Table 2.4: Comparison of successful and failed instances for AltMinReg, $\lambda = 3e-5$, $m = n = 1000$, $r = 10$, $p = 0.05$. This table is an averaged version of the table in Figure 2.10.

This observation clearly validates our previous claim, which is restated below.

Claim 2.1.1 *(Empirical claim) The reason why AltMinReg improves upon AltMin for some p is not that it can control the norms of X, Y, but that it can control the incoherence constants of X, Y, even though the underlying mechanism is unknown.*

In the above experiments where $p = 5r/n = 0.05$, we can obtain a test error $< 10^{-3}$ for properly chosen λ. However, as p becomes smaller and smaller, the regularizer coefficient λ should be larger and larger to control the incoherence, resulting in larger and larger training error and test error. Simulation results show that when $p = 4r/n = 0.04$, AltMinReg can achieve a training error below 0.05 and a test error below 0.1; when $p = 3r/n = 0.03$, AltMinReg can only achieve a test error of around 0.2.

2.2 Gradient Methods

Gradient descent (GD) refers to a class of iterative optimization algorithms [86] that uses the gradient information to update the iterates. It can be used to solve both convex and non-convex problems and the convergence to stationary points can be guaranteed under suitable stepsize rules. It has gained much interest in big data optimization due to its cheap iteration, which only involves a gradient evaluation.

For the MF formulation (1.10), the gradient $\nabla F = (\nabla_X F, \nabla_Y F)$ can be easily computed as follows:

$$\begin{aligned} \nabla_X F(X,Y) &= \mathcal{P}_\Omega(XY^T - M)Y, \\ \nabla_Y F(X,Y) &= \mathcal{P}_\Omega(XY^T - M)^T X. \end{aligned} \tag{2.10}$$

Note that one does not need to compute XY^T which requires storing a $m \times n$ matrix. Denote $x_i^T = X^{(i)} \in \mathbb{R}^{1\times r}, y_j^T = Y^{(j)} \in \mathbb{R}^{1\times r}$ as the rows of X and Y, and $\partial_i^x = \{j \mid (i,j) \in \Omega\}$ as the set of "column neighbors" of row i. Then the i-th row of $\nabla_X F$ can be computed by

$$(\nabla_X F)^{(i)} = \sum_{j\in\partial_i^x} (x_i^T y_j - M_{ij}) y_j,$$

which only requires memory size $O(|\partial_i^x|)$ and computation time about $4r|\partial_i^x|$. For computing $\nabla_X F(X,Y)$, the total memory size is $O(|\Omega|)$ and the total computation time is about $4r|\Omega|$. Similarly, for computing $\nabla_Y F(X,Y)$ the total memory size is $O(|\Omega|)$ and the total computation time is about $4r|\Omega|$. The total computation time of each gradient step is about $8r|\Omega|$. Compared to the $O((m+n)r^3 + 2r^2|\Omega|) \geq 2r^2|\Omega|$ time for each iteration of AltMin, the gradient evaluation takes less than $\frac{4}{r}$-fraction of time. For a large-scale matrix completion problem, the value of r can be as large as, say, 50, in which case GD takes no more than $4/r \approx 1/10$ fraction of time in each iteration compared to AltMin.

There are many choices of stepsizes for gradient descent method such as constant stepsize, line search (exact or limited), diminishing stepsize, Armijo rule [86], BB (Barzilai-Borwein) stepsize [87] and Nesterov's accelerated stepsize. Some of these stepsize rules (liner search and Armijo rule) require an inner loop, which may increase the per-iteration cost. Here we explore three step-size rules: constant stepsize, LBB (long BB) stepsize and SBB (short BB) stepsize, and we will give a short introduction below.

For an unconstrained smooth optimization problem $\min_x f(x)$, where $x \in \mathbb{R}^n$, the gradient descent method is given by

$$x_{k+1} = x_k - \eta_k \nabla f(x_k),$$

where η_k is the stepsize at the k-th iteration. For constant stepsize rule, η_k is chosen to be a fixed constant η. Suppose ∇f is Lipschitz continuous with constant L (f is not necessarily convex), i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y, \tag{2.11}$$

and the stepsize $\eta < 2/L$, then GD converges to stationary points (more precisely, each limit point of the sequence of iterates is a stationary point) [86, Proposition 1.2.3].

However, for some problems (including the matrix completion problem) it is not easy to obtain an accurate estimate of the Lipschitz constant L; in addition, the universal upper bound L might be too pessimistic for some iterations and resulting in a slower convergence. In practice, one can test various values of η and pick the best one.

The BB stepsize rule is a popular tuning-free step-size rule that is known to perform very well in practice (see, e.g., [87]). The motivation of the BB stepsize is to mimic the Newton method by approximating the Hessian matrix $\nabla^2 f(x_k)$ by a simpler matrix $\eta_k I$, where I denotes an identity matrix. There are two versions of BB stepsize rule: LBB (long BB) and SBB (short BB), which are defined as

$$\eta_k^{\mathrm{LBB}} = \frac{s_k^T s_k}{s_k^T d_k}, \quad \eta_k^{\mathrm{SBB}} = \frac{s_k^T d_k}{d_k^T d_k}, \\ \text{where } s_k = x_k - x_{k-1},\ d_k = \nabla f(x_k) - \nabla f(x_{k-1}). \tag{2.12}$$

The linear convergence of the BB method (i.e. the gradient method with BB stepsize) for strongly convex quadratic functions has been established [88], but for non-quadratic strongly convex functions it can be divergent [87]. Many variants have been proposed to improve the BB stepsize; for example, one can use an extra safeguard step:

$$\eta_k^{\mathrm{BB-SF}} \triangleq \min(\max(\eta_k^{\mathrm{BB}}, \eta_{\min}), \eta_{\max}),$$

where $0 < \eta_{\min} < \eta_{\max}$ are fixed constants. Note that the BB method is not a monotone method (i.e. the function values of the iterates do not necessarily decrease), thus rigorously speaking it is not a "gradient descent method", but just a "gradient method"; nevertheless, sometimes we will still call it GD with BB stepsize.

Now we present GD with different stepsize rules below; in the table $\boldsymbol{x}_k = (X_k, Y_k)$ denotes the k-th iterate.

In the numerical experiments, we consider the same setting as before, i.e. M is generated by $M = UV^T$ where $U, V \in \mathbb{R}^{1000\times 10}$ has independent Gaussian entries with zero mean and variance $1/1000$, and Ω is generated by a Bernolli model with parameter p. We use random initial point $X_0, Y_0 \in \mathbb{R}^{1000\times 10}$ with independent Gaussian entries (also zero mean and variance $1/1000$). The figures are obtained by averaging 100 Monte Carlo runs.

For GD with constant stepsize, we find that $\eta = 10$ or 11 is the best choice for this setting; meanwhile, $\eta = 13$ can lead to divergence, and $\eta < 8$ can lead to very slow

Table 2.5: GD (Gradient descent)

Initialization: randomly generate (X_0, Y_0).

The k-th iteration:

$$X_k \longleftarrow X_{k-1} - \eta_k \nabla_X F(X_{k-1}, Y_{k-1}),$$
$$Y_k \longleftarrow Y_{k-1} - \eta_k \nabla_Y F(X_{k-1}, Y_{k-1}),$$

where the stepsize η_k is chosen according to one of the following rules:

a) Constant stepsize: $\eta_k = \eta,\ \forall\ k$, where η is a constant.

b) BB stepsize with safeguard:

$$\eta_k = \min(\max(\eta^{\mathrm{BB}}, \eta_{\min}), \eta_{\max}),$$

where η^{BB} is computed by one of the following two rules

b1) LBB stepsize: $\eta_k^{\mathrm{BB}} = \frac{\boldsymbol{s}_k^T \boldsymbol{s}_k}{\boldsymbol{s}_k^T \boldsymbol{d}_k}$,

b2) SBB stepsize: $\eta_k^{\mathrm{BB}} = \frac{\boldsymbol{s}_k^T \boldsymbol{d}_k}{\boldsymbol{d}_k^T \boldsymbol{d}_k}$.

Here $\boldsymbol{s}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1} = (X_k - X_{k-1}, Y_k - Y_{k-1})$,

$\boldsymbol{d}_k = \nabla F(\boldsymbol{x}_k) - \nabla F(\boldsymbol{x}_{k-1}) = (\nabla_X F(\boldsymbol{x}_k) - \nabla_X F(\boldsymbol{x}_{k-1}), \nabla_Y F(\boldsymbol{x}_k) - \nabla_Y F(\boldsymbol{x}_{k-1}))$.

convergence. In other words, the performance of GD with constant stepsize is rather sensitive to the choice of the stepsize. There is no universal constant that works well for all settings; in fact, in the setting $m = n = 100, r = 5, p = 0.25$, the best constant stepsize is $\eta \approx 2$, and $\eta > 3$ will lead to divergence. To save the tuning time, one can pick two random points $\boldsymbol{x} = (X, Y)$ and $\hat{\boldsymbol{x}} = (\hat{X}, \hat{Y})$ and use $\frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_F}{\|\nabla F(\boldsymbol{x}) - \nabla F(\hat{\boldsymbol{x}})\|_F}$ as a reference value for the stepsize. The BB method with safeguard is almost tuning free, and the only parameters are the safeguard boundaries $\eta_{\min}$ and $\eta_{\max}$. The performance of the BB method is not very sensitive to the safeguard boundaries, and we set them to 0.2 and 25 in the experiments. We notice that the pure BB method without safeguard can fail in very rare cases (less than 5% of the time), and the reason of failure is that the BB stepsize can be very large. Thus the safeguard step (especially the upper bound of the stepsize) is needed for the convergence of the BB method in our problem.

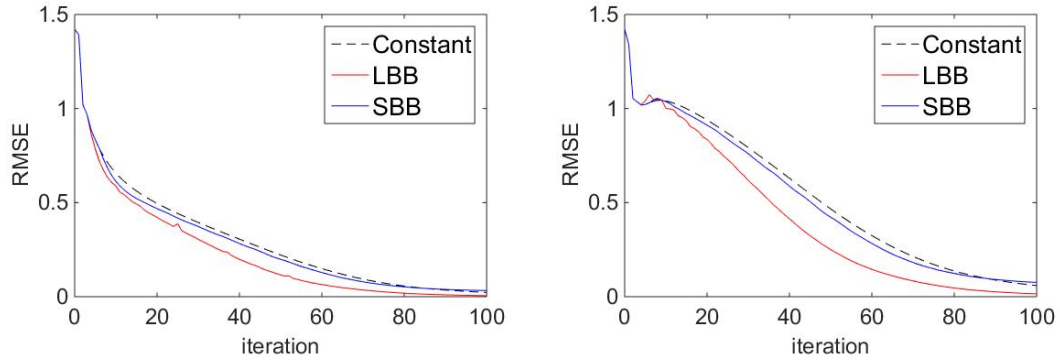
Fig. 2.11(a) and Fig. 2.11(b) show the training error and the test error of three gradient methods respectively. It can be seen that the LBB stepsize is the best among the three. The recovery rate (only require the final test error < 0.1) for the three methods is $99\%, 85\%, 83\%$. These recovery rates correspond to just choosing one random initial point for each realization of Ω, and they can be further improved by starting from

multiple random initial points and picking the best. Recall that for AltMin, we only obtain a recovery rate below 30% for $p = 0.05$, thus the gradient methods, even the version with the simplest constant stepsize rule, can greatly improve the recovery rate. One drawback of the gradient method with constant stepsize is that it converges slowly when close to optimum; in fact, this is a well known drawback of the first order methods. Nevertheless, the LBB gradient method converges fast even when close to optimum: in 100 iterations it can obtain an average test error of 0.015, and in 200 iterations it can obtain an average test error below 0.001. This phenomenon is also expected since the BB method is actually a quasi-Newton method which approximates the Hessian by a diagonal matrix with equal diagonal entries. AltMin converges faster to a test error of 10^{-4} if it succeeds; nevertheless, a recovery accuracy of 0.01 is acceptable for many practical scenarios with noise.

It is a bit tricky to compare the computation time of GD (including constant size and BB stepsize) and alternating minimization (including AltMin and AltMinReg). First, the success rates (assuming one random initial point) of these two methods are different, and a more fair comparison might require restarting many times until achieving success. Second, different choices of the target recovery accuracy ϵ lead to rather different conclusions: for $\epsilon = 10^{-6}$, AltMin (even considering restarting) probably takes less time than the gradient methods, but for $\epsilon = 0.2$ the gradient methods will probably win. Third, it is unfair to just compare the number of iterations since GD and alternating minimization have different per-iteration costs; roughly speaking one should multiply the number of iterations for AltMinReg by at least $r/4$ to compare with the gradient methods. With these considerations in mind, a coarse conclusion drawn from our simulation results is that alternating minimization and GD are comparable in terms of computation time.

Another important issue involved in the comparison of the time cost is the distributed/parallel implementation. It is widely known that AltMin-type methods for matrix completion are easily parallelizable since each row of X and Y can be updated independently. The gradient method with constant stepsize enjoys the same property since the gradient is decomposable across the rows of X and Y (this is exactly the reason why AltMin can be parallelizable). The BB method (i.e. the gradient method with BB stepszie) requires an additional global step of computing the stepsize η_k based on

$\boldsymbol{s}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ and $\boldsymbol{d}_k = \nabla_X F(\boldsymbol{x}_k) - \nabla_Y F(\boldsymbol{x}_{k-1})$. This compuation takes $O(|\Omega|r + r^2)$ time, but it can be performed in a distributed fashion. To illustrate this, note that $\|X_k\|_F^2$ can be computed in a parallel fashion: in a single core it takes time $2mr$; if there are 10 cores, we can divide $\{1, 2, \ldots, m\}$ into 10 disjoint sets $J_1, \ldots, J_{10}$, and each core can compute $\sum_{i \in J_t} \|X_k^{(i)}\|^2, t = 1, \ldots, 10$ in time $2mr/10$ and then sum them up. Similarly the inner products $\boldsymbol{s}_k^T \boldsymbol{d}_k, \boldsymbol{d}_k^T \boldsymbol{d}_k, \boldsymbol{s}_k^T \boldsymbol{s}_k$ can be computed in a parallel fashion since they are decomposable across rows, hence the BB stepsizes can be computed in a parallel fashion.

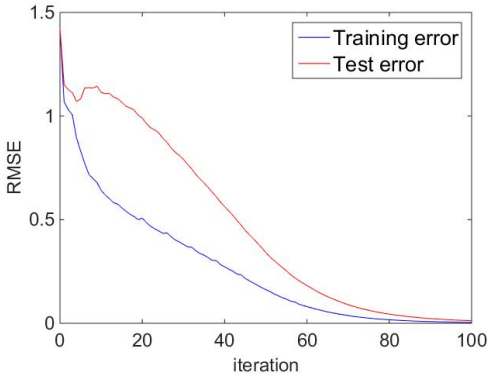


(a) Training error

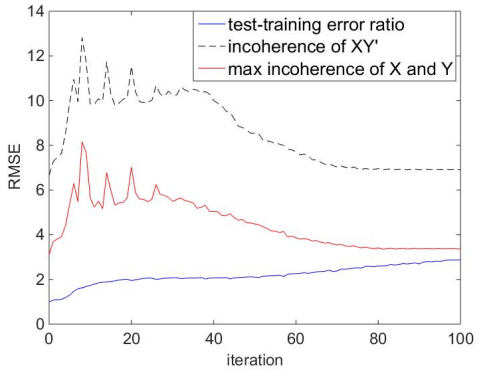
(b) Test error

Figure 2.11: Training and test errors v.s. iteration; for several gradient methods. $m = n = 1000, r = 10, p = 0.05$.

Next, we will show that the incoherence constants are indicators of the success or failure of the gradient methods. We consider a successful instance ($p = 0.05$) and a failed instance ($p = 0.03$) under the previous setting, and compare the RIP constant (the ratio of the test error over the training error) with the incoherence constants. Fig. 2.12(b) shows that for the successful instance ($p = 0.05$), $\max(\mu(X), \mu(Y))$ is kept below 4 and decreases to around 3; Fig. 2.13(b) shows that for the failed instance ($p = 0.03$), $\max(\mu(X), \mu(Y))$ grows slowly to about 8 in 500 iterations. Another interesting finding is that $\max(\mu(X), \mu(Y))$ matches very well with the RIP constant in Fig. 2.13(b). For the gradient methods, $\max(\mu(X), \mu(Y))$ may grow to $> 40 \approx O(\sqrt{n})$ after, say, 10^5 iterations according to the trend in Fig. 2.13(b). This is different from AltMin for which $\max(\mu(X), \mu(Y))$ can grow to $O(\sqrt{n})$ in just a few iterations.

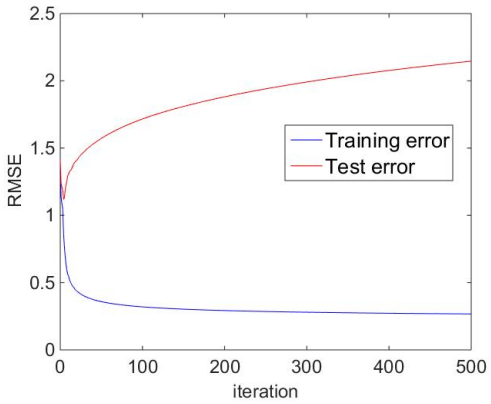


(a) Training and test error

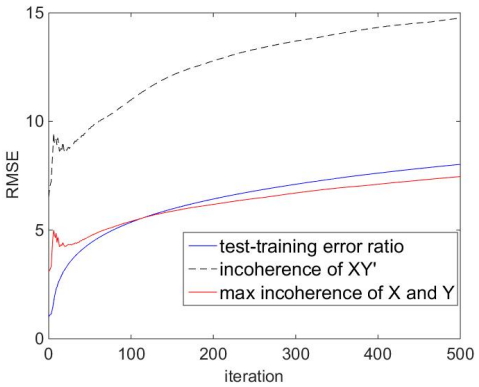


(b) Test-training error ratio and incoherence constants

Figure 2.12: Successful instance for the gradient method with LBB stepsize. $p = 0.05$ and $m = n = 1000, r = 10$. The figure shows that the incoherence constants do not grow.



(a) Training error



(b) Test error

Figure 2.13: Failed instance for the gradient method with LBB stepsize. $p = 0.03$ and $m = n = 1000, r = 10$. The figure shows that the incoherence constants grow slowly along with the test-training error ratio.

We summarize the findings of this subsection as follows.

Observation 2.2.1 *(Performance of GD) GD has better recovery ability than AltMin,*

since the row-norms of the iterates generated by GD change slowly compared to AltMin. When p is not very small, in failed instances of GD the incoherence constants of the iterates are larger than usual and monotonically increasing; in successful instances of GD the incoherence constants of the iterates are bounded from above.

We remark that when p is very small, e.g. $p = 0.025$, GD fails and both the norms and maximum row-norms of the iterates become very huge in a few iterations.

2.3 SGD (Stochastic Gradient Descent)

Stochastic gradient descent (SGD) is a popular method for minimizing the expected value of a function or the sum of finitely many component functions. In the optimization field, the algorithm for minimizing a finite sum is more commonly referred to as "incremental gradient method", while SGD represents the algorithm for minimizing the expectation of a function; nevertheless, in this thesis we will use "SGD" to denote the algorithm for minimizing a finite sum.

SGD has been very popular for solving the MF based formulation (1.10) and (1.13) [1, 40, 59–63]. The objective function $F(X, Y)$ can be decomposed as

$$F(X, Y) = \sum_{(i,j)\in\Omega} F_{ij}(X, Y),$$

where the component functions

$$F_{ij}(X, Y) = \frac{1}{2}[(XY^T - M)_{ij}]^2 = \frac{1}{2}[(X^{(i)})^T Y^{(j)} - M_{ij}]^2, \ (i, j) \in \Omega. \tag{2.13}$$

At each iteration one picks a component function and performs a gradient update. Similar to the BCD type methods where the blocks can be chosen in different orders, in SGD one can pick the component functions in a cyclic order, in an essentially cyclic order, or in a random order (can be either sampling with replacement or sampling without replacement). We only consider the vanilla SGD with constant stepsize, where the component functions are sampled either with or without replacement.

Theoretically speaking, SGD with constant stepsize may not converge and the error can be proportional to the stepsize (see, e.g., [89, Sec. 2]). Using a diminishing stepsize rule can guarantee convergence of SGD (in the sense that each limit point is a stationary

Table 2.6: SGD

Initialization: randomly generate (X_0, Y_0).
The $(k+1)$-th loop:
$X \longleftarrow X_k, \quad Y \longleftarrow Y_k$.
For $t = 1, \ldots, |\Omega|$:
Randomly pick (i, j) from Ω (either sampling with or without replacement);
$\epsilon_{ij} \longleftarrow (X^{(i)})^T Y^{(j)} - M_{ij}$,
$X^{(i)} \longleftarrow X^{(i)} - \eta \epsilon_{ij} Y^{(j)}$,
$Y^{(j)} \longleftarrow Y^{(j)} - \eta \epsilon_{ij} X^{(i)}$.
End For
$X_{k+1} \longleftarrow X, \quad Y_{k+1} \longleftarrow Y$.

point) [90]. For our problem, SGD with diminishing stepsize converges much slower than GD, thus we only consider SGD with constant stepsize.

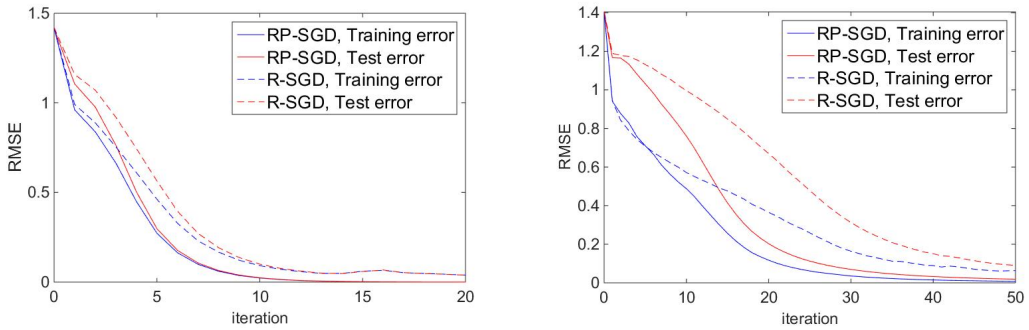
We consider the same setting as before, i.e. M is generated by $M = UV^T$ where $U, V \in \mathbb{R}^{1000\times 10}$ have independent Gaussian entries with zero mean and variance $1/1000$, and Ω is generated by a Bernolli model with parameter p. We use random initial point $X_0, Y_0 \in \mathbb{R}^{1000\times 10}$ with independent Gaussian entries (also zero mean and variance $1/1000$). The figures are obtained by averaging 100 Monte Carlo runs.

For the matrix completion problem SGD allows a much bigger stepsize than GD in many cases, making SGD much faster than GD. In fact, SGD with stepsize 70 can converge, while GD diverges for stepsize > 13. This explains why SGD in Figure 2.14(a) takes fewer than 10 iterations to converge, while in Figure 2.11 GD takes about 70 iterations to converge (both for the case $p = 0.05$). Thus for $p = 0.03$, SGD is much better than GD and AltMin since GD and AltMin almost always fail.

Note, however, that SGD with a large stepsize does not always converge; in fact, SGD with stepsize 70 only converges in 70%-80% of the experiments, and SGD diverges in 2 or 3 iterations (norms of X, Y suddenly jump to $> 10^{200}$) in other 20%-30% experiments. For smaller p (e.g. $p = 0.025$), the failure probability is very high (larger than, say, 70%). One way to save the failed instances is to re-run the algorithm from different random initial points, but this will cost extra time and may not work when the failure probability is high ($p = 0.025$). The success probability of SGD for $p = 0.03$ can be increased to almost 100% if we use a much smaller stepsize, but the byproduct is the slower convergence; for $p = 0.025$ we have not found a stepsize for SGD that enables

convergence. We will propose a simple correction scheme that improves the success probability without sacrificing the convergence speed in Chapter 3.

Another observation is that RP-SGD (randomly permuted SGD, i.e. sampling without replacement) converges much faster than R-SGD (randomized SGD, i.e. sampling with replacement); in addition, RP-SGD converges to very accurate solutions (error $< 1e{-}3$) while R-SGD converges only to an approximate solution (error > 0.05). The superiority of random permutation (sampling without replacement) over sampling without replacement has been reported experimentally for many algorithms such as SGD [62] and ADMM [71], but the theoretical analysis of this phenomenon seems to be rather difficult (see [71, 91] for some analysis).



(a) $p = 0.05$ (b) $p = 0.03$

Figure 2.14: Successful instances of SGD. Consider two variants: in RP-SGD (randomly permuted SGD) we use sampling with replacement; in R-SGD (randomized SGD) we use sampling without replacement. Note that RP-SGD and R-SGD diverge in about 30% and 22% of all experiments respectively, which are not shown in the above figures.

We summarize the findings of this subsection as follows.

Observation 2.3.1 *(Performance of SGD) SGD with a large stepsize performs much better than AltMin and GD: it takes much less time and can recover the matrix M with fewer samples. However, a byproduct of the large stepsize is that SGD diverges with certain probability when p is small.*

Chapter 3

Incoherence Control: New Formulation and Algorithms

The main lesson we learned from Chapter 2 is: bounded row-norms (or incoherence constants) is the key indicator of successful recovery for various algorithms. When there are abundant samples, the incoherence constants will be bounded (according to simulation) for either AltMin or the gradient methods, though the mechanism is unknown yet. However, when the number of samples is limited, the incoherence constants may go unbounded. Adding a regularizer $\lambda(\|X\|_F^2 + \|Y\|_F^2)$ is a way to indirectly control the incoherence constants, and it does help; however, it may distort the optimal solution.

In this chapter, we propose new formulations and methods that directly control the row-norms. Simulation results show that the new methods can recover the original matrix even when the number of observations is close to the fundamental limit.

3.1 New Formulation

As mentioned before, the traditional regularizer $\lambda(\|X\|_F^2 + \|Y\|_F^2)$ indirectly control the incoherence constants. It is natural to consider a more direct way to control the incoherence constants, while not causing distortion to the optimal solution. One simple

method is to add constraints

$$\begin{aligned}(X,Y) \in K_2 \triangleq &\{\|X^{(i)}\| \le \beta_1,\ \forall\ i;\ \|Y^{(j)}\| \le \beta_2,\ \forall j\},\\ &\text{where } \beta_1 = \beta_T \frac{\sqrt{\mu}}{\sqrt{m}},\ \beta_2 = \beta_T \frac{\sqrt{\mu}}{\sqrt{n}},\end{aligned} \tag{3.1}$$

in which β_T is roughly the norm of the groundtruth factors X or Y, and μ is the desired incoherence constant. Then the problem formulation becomes

$$\begin{aligned}\min_{X,Y} \quad & \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2,\\ \text{s.t.} \quad & \|X^{(i)}\| \le \beta_1,\ \forall\ i; \quad \|Y^{(j)}\| \le \beta_2,\ \forall\ j.\end{aligned} \tag{3.2}$$

This formulation has appeared in [92] (see equation (6) there), and the constraint is called a "max-norm" constraint in that paper. Note that the motivation of [92] is not to control the incoherence constants. Problem (4.9) can be solved by a simple gradient projection algorithm, as suggested by [92]: after performing the gradient step, one can simply add an additional projection step if the row norms of the iterates are larger than the thresholds (β_1 for X, and β_2 for Y). Unfortunately, this simple gradient projection method does not work for $p = 0.03, m = n = 1000, r = 10$. According to our theoretical analysis in Chapter 4, this failure is possibly because the direction to go in the gradient projection method (the difference between concecutive itereates) is not positively related to the "global descent direction" $\boldsymbol{x}^* - \boldsymbol{x}_k$ (i.e. the angle is larger than $\pi/2$), where $\boldsymbol{x}_k$ is the k-th iterate and $\boldsymbol{x}^*$ is one global optimum; see more details in the proof of (4.52). Note that the direction to go in the gradient projection method is positively related to the opposite direction of gradient direction, which is a locally descent direction; however, due to the non-convex nature of the problem, the locally descent direction can be very different from the global descent direction $\boldsymbol{x}^* - \boldsymbol{x}_k$.

3.1.1 New Formulation with Both Row-norm and Norm Constraints

A seemingly non-intuitive way to improve the performance is to add an additional constraint

$$(X,Y) \in K_1 \triangleq \{\|X\|_F \le \beta_T,\ \|Y\|_F \le \beta_T\}. \tag{3.3}$$

In other words, we try to solve the following constrained problem:

$$\begin{aligned} \min_{X,Y} \quad & \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2, \\ \text{s.t.} \quad & \|X\|_F \le \beta_T, \quad \|Y\|_F \le \beta_T; \\ & \|X^{(i)}\| \le \beta_1, \ \forall\ i; \quad \|Y^{(j)}\| \le \beta_2, \ \forall\ j. \end{aligned} \tag{3.4}$$

Note that if X and Y satisfy (3.1), then $\|X\|_F, \|Y\|_F$ satisfy a weaker upper bound $\|X\|_F \le \mu\beta_T, \|Y\|_F \le \mu\beta_T$ than the bound given in (3.3). Why does a stronger bound (3.3) help? Intuitively, with the new constraint (3.1), the direction to go in a gradient projection method will become positively related to the global descent direction $\boldsymbol{x}^* - \boldsymbol{x}_k$. Technically, the extra constraint (3.3) plays a crucial role in the proof of (4.52). Here is an interesting phenomenon: this extra constraint does not only help the theoretical analysis, but also greatly improves the emprical performance. Therefore, our work provides a good theoretical guidance for practical algorithm design.

When performing the algorithms, one can first project the iterates to K_1 (i.e. the ball specified in (3.3)), and then project them to K_2 (the ball specified in (3.1)). This is not a precise gradient projection algorithm which requires projection to $K_1 \cap K_2$; here, we just perform two consecutive projections to K_1 and K_2. We call it an AGP (approximate gradient projection) algorithm.

Table 3.1: AGP (Approximate Gradient Projection) for solving (3.4)

Initialization: randomly generate (X_0, Y_0).
The k-th iteration:
Gradient step: $X_k \longleftarrow X_{k-1} - \eta_k \nabla_X F(X_{k-1}, Y_{k-1})$,
$Y_k \longleftarrow Y_{k-1} - \eta_k \nabla_Y F(X_{k-1}, Y_{k-1})$.
Full-scaling step: If $\|X_k\|_F > \beta_T$, define $X_k \longleftarrow X_k \frac{\beta_T}{\|X_k\|_F}$;
if $\|Y_k\|_F > \beta_T$, define $Y_k \longleftarrow Y_k \frac{\beta_T}{\|Y_k\|_F}$.
Row-scaling step: If $\|X_k^{(i)}\| > \beta_1$, define $X_k^{(i)} \longleftarrow X_k^{(i)} \frac{\beta_1}{\|X_k^{(i)}\|}$, $i = 1, \ldots, m$;
if $\|Y_k^{(j)}\| > \beta_2$, define $Y_k^{(j)} \longleftarrow Y_k^{(j)} \frac{\beta_2}{\|Y_k^{(j)}\|}$, $j = 1, \ldots, n$.

It is not clear whether AGP converges to KKT points of (3.4). We then consider a

partially regularized version of (3.4)

$$\begin{aligned} \min_{X,Y} \quad & \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 + \rho\max(\|X\|_F^2, \beta_T^2) + \rho\max(\|Y\|_F^2, \beta_T^2), \\ \text{s.t.} \quad & \|X^{(i)}\| \le \beta_1, \ \forall\ i; \quad \|Y^{(j)}\| \le \beta_2, \ \forall\ j. \end{aligned} \tag{3.5}$$

Compared to the traditional regularizer $\lambda(\|X\|_F^2 + \|Y\|_F^2)$ that is effective for all X, Y, the new regularizer $\rho\max(\|X\|_F^2, \beta_T^2) + \rho\max(\|Y\|_F^2, \beta_T^2)$ only penalizes X, Y that are outside of the desired feasible region K_1. From an optimization point of view, these two regularizers belong to the Lagrangian multiplier method and the penalty method respectively. We can also view the regularizer $\lambda(\|X\|_F^2 + \|Y\|_F^2)$ as a "soft regularizer", and our new regularizer as a "hard regularizer". The advantage of the hard regularizer is that it does not distort the optimal solution.

Table 3.2: GP (Gradient Projection) for solving (3.5)

Initialization: randomly generate (X_0, Y_0).
The k-th iteration:
Gradient step: $X_k \longleftarrow X_{k-1} - \eta_k\left(\nabla_X F(X_{k-1}, Y_{k-1}) + \lambda X_{k-1} I_{\beta_T}(X_{k-1})\right),$
$Y_k \longleftarrow Y_{k-1} - \eta_k(\nabla_Y F(X_{k-1}, Y_{k-1} + +\lambda Y_{k-1} I_{\beta_T}(Y_{k-1})).$
Here, $I_{\beta_T}(X) = 0$, if $\|X\|_F \le \beta_T; I_{\beta_T}(X) = 1$, otherwise.
Row-scaling step: If $\|X_k^{(i)}\| > \beta_1$, define $X_k^{(i)} \longleftarrow X_k^{(i)} \frac{\beta_1}{\|X_k^{(i)}\|}$, $i = 1, \ldots, m$;
if $\|Y_k^{(j)}\| > \beta_2$, define $Y_k^{(j)} \longleftarrow Y_k^{(j)} \frac{\beta_2}{\|Y_k^{(j)}\|}$, $j = 1, \ldots, n$.

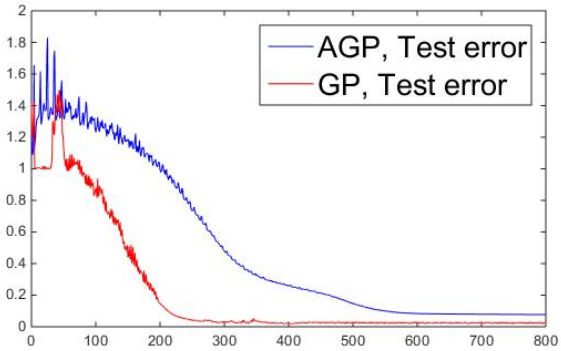


Figure 3.1: Performance of AGP and GP for $m = n = 1000$, $r = 10$, $p = 0.03$. The fluctuation is because we use the BB stepsize.

In the numerical experiments, we consider $m = n = 1000, r = 10, p = 0.3$ and the same random models of M, Ω and initial points as before. We choose $\mu = 4 = 2^2$, which

is close to the incoherence constant of groudtruth factors X or Y (about 1.9^2). For the synthesis data set, we can choose β_T to be close to $\sqrt{\|M\|_*}$ (or $c\sqrt{\|M\|_*}$ where $c > 1$), since $\sqrt{\|M\|_*}$ is the norm of the SVD factors $\hat{U}\Sigma^{1/2}$ and $\hat{V}\Sigma^{1/2}$ when the SVD of M is $M = \hat{U}\Sigma\hat{V}^T$. In our setting, $\sqrt{\|M\|_*} \approx 3.13$, thus we choose $\beta_T = 3.2$. In practical scenarios where the groundtruth M is unknown, we guess that β_T^2 should be chosen in the region $[\frac{\|\mathcal{P}_\Omega\|_F}{\sqrt{p}}, 2\frac{\|\mathcal{P}_\Omega\|_F}{\sqrt{p}}\sqrt{r}]$, since ideally $\frac{\|\mathcal{P}_\Omega\|_F}{\sqrt{p}} \approx \|M\|_F \in [\frac{\|M\|_*}{\sqrt{r}}, \|M\|_*]$.

Figure 3.6 shows the simulation results for AGP and GP (using LBB stepsize). AGP for the formulation (3.2) converges to a test error that is below < 0.1 in fewer than 500 iterations; GP for the formulation (3.5) converges to an average test error < 0.01 in 200 iterations. Both methods work quite well, while GP for the partially regularized formulation converges faster than AGP for the constrained version. We guess this is because AGP does not perform exact projection to the constrained set at each iteration, resulting in a slower convergence. We remark that while the presented simulation results are for the LBB stepsize rule, AGP and GP with constant stepsize (say, stepsize 12) are slower than the versions with LBB stepsize and converge in 500-1500 iterations. Figure 3.6 only shows the average performance. A closer look reveals that the probability of success (generating a test error < 0.01) for AGP and GP are both above 90% (but below 95%), and we expect that several random starts can improve the success probability to more than 99%.

A naive fundamental limit is $pmn \approx |\Omega| > r(m + n - r)$ (the number of degrees of freedom for a rank-r matrix), which means $p \geq 2r/n = 0.02$ (more precisely, $p \geq 0.0199$) in the setting $m = n = 1000, r = 10$ [1] . Thus, $p = 0.03$ is already close to the fundamental limit. For such a small fraction of observations, all traditional methods, including AltMinReg and SGD, cannot converge to a reasonable solution with high probability. In constrast, the proposed AGP and GP can recover the original matrix for $p = 0.03$ with high probability.

[1] Using the algorithm in [37], Keshavan plots the fundamental limit in [24, Figure 6.1] for the setting $m = n = 1000, r = 10$ and the same random models of M, Ω as ours; roughly speaking, the fundamental limit is between 0.02 and 0.03.

3.1.2 SGD with Row-norm Projection

In the previous subsection, we mentioned that GD for the formulation (3.2) does not work when $p = 0.03$, and an extra norm-scaling step is needed. Interestingly, SGD for the formulation (3.2) works very well for $p = 0.03$ in our experiments. To be precise, the algorithm we consider is a variant of SGD in which we add an additional projection step (row-scaling) after the gradient step (without norm-scaling step); we call this new algorithm SGP (Stochastic Gradient Projection). The details of SGP is given in Table 3.3.

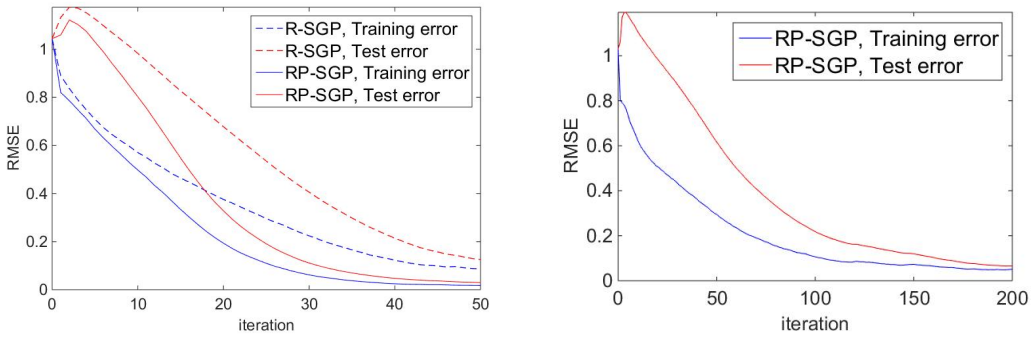
Table 3.3: SGP (Stochastic Gradient Projection)

Initialization: randomly generate (X_0, Y_0).
The $(k+1)$-th loop:
$X \longleftarrow X_k, \quad Y \longleftarrow Y_k$.
For $t = 1, \ldots, |\Omega|$:
Randomly pick (i, j) from Ω (either sampling with or without replacement);
$\epsilon_{ij} \longleftarrow (X^{(i)})^T Y^{(j)} - M_{ij}$,
Gradient step: $X^{(i)} \longleftarrow X^{(i)} - \eta \epsilon_{ij} Y^{(j)}, \quad Y^{(j)} \longleftarrow Y^{(j)} - \eta \epsilon_{ij} X^{(i)}$,
Projection step: if $\|X^{(i)}\| > \beta_1$, define $X^{(i)} \longleftarrow X^{(i)} \frac{\beta_1}{\|X^{(i)}\|}$,
if $\|Y^{(j)}\| > \beta_2$, define $Y^{(j)} \longleftarrow Y^{(j)} \frac{\beta_2}{\|Y^{(j)}\|}$,
where β_1, β_2 are defined in (3.1).
End For
$X_{k+1} \longleftarrow X, \quad Y_{k+1} \longleftarrow Y$.

In the numerical experiments, we consider $m = n = 1000, r = 10$ and the same random models of M, Ω and initial points as before; we choose $\beta_T = 3.2$ and $\mu = 3.6$. Figure 3.3 shows that SGP (either with or without resampling) converges to the original matrix in fewer than 50 iterations for $p = 0.03$. Although this figure looks similar to Figure 3.2(a) which shows the performance for the vanilla SGD, we emphasize that the vainlla SGD fails with probablity 20%-30% that is not shown in Figure 3.2(a), while SGP suceeds with probablity close to 1 (all 100 instances we have tested are successful). Therefore, a simple row-scaling step significantly improves the recovery ability of SGD.

The superiority of SGP over the vanilla SGD is more noticable when there are fewer observations. When $p = 0.025$, the vanilla SGD almost always diverges; in constrast, SGP with stepsize 75 converges in no more than 200 iterations to a solution with test

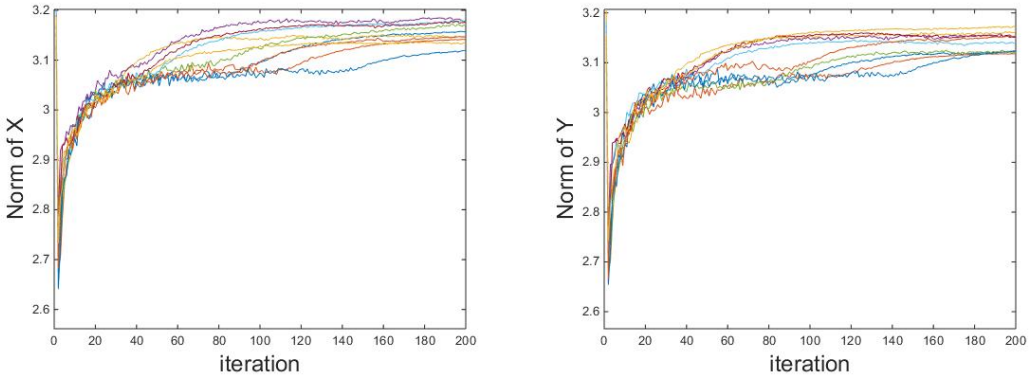
error < 0.1 (more iterations can further reduce the test error). See the simulation results in Figure 3.2(b).



(a) $p = 0.03$ (b) $p = 0.025$

Figure 3.2: Performance of SGP (Stochastic Gradient Projection) for $m = n = 1000$, $r = 10$. Two versions of SGP: in RP-SGP (randomly permuted SGP) we use sampling with replacement; in R-SGP (randomized SGP) we use sampling without replacement.

One natural question is: for the formulation (3.2), why does SGP succeed, while GP fails? In other words, why do we *not* need to control the norms of the iterates for SGP, while for GP we have to control the norms as well as the row-norms? It turns out that SGP controls the norms of the iterates automatically: the norms are always below $3.2 = \beta_T$ in our numerical experiments; see Figure 3.3. This phenomenon seems mysterious since in SGP we do not impose constraints $\|X\|_F \leq \beta_T, \|Y\|_F \leq \beta_T$. In the worst case, all the row-norms $\|X^{(i)}\|, \|Y^{(j)}\|$ are equal to the upper bounds β_1 and β_2 respectively, resulting in $\|X\|_F = \|Y\|_F = \sqrt{\mu}\beta_T > \beta_T$. We guess that the nature of SGD-type methods prevents this worst-case situation: in each iteration (here one "iteration" refers to one pass of all component functions) $X^{(i)}$ and $Y^{(j)}$ are updated multiple times, and it seems unplausible that the new $\|X^{(i)}\|, \|Y^{(j)}\|$ become large each time. In other words, SGD-type methods average out the bad cases (i.e. the row-norms become large), thus after each iteration (i.e. each pass) the norms of X and Y stay bounded above. In contrast, GD-type methods update all rows at the same time, and it is possible that all rows become large at the same time, thus only row-scaling does not guarantee $\|X\|_F \leq \beta_T, \|Y\|_F \leq \beta_T$.



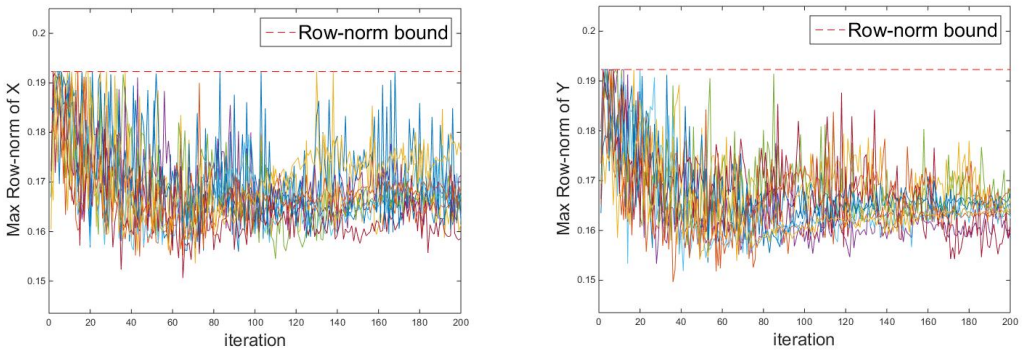
(a) Norm of X_k v.s. iteration k

(b) Norm of Y_k v.s. iteration k

Figure 3.3: Norms of the itereates generated by SGP (Stochastic Gradient Projection) for $p = 0.025$, $m = n = 1000$, $r = 10$, $\beta_T = 3.2$, $\mu = 1.9^2$. The 10 lines in 10 different colors represent 10 different experiments.

At this point we do not know whether this phenomenon is universal: if M and Ω are generated from different models, SGP might fail to control the norms of X, Y and extra control on the norms of the itereates is needed. We can add an additional norm-scaling step after each pass of SGP, i.e. project X_k, Y_k to the ball $K_1 = \{(X, Y) \mid \|X\|_F \leq \beta_T, \|Y\|_F \leq \beta_T\}$. Another possiblity is to consider the partially regularized formulation (3.5) and perform SGP for that formulation. In Chapter 4, we will provide a theoretical gurantee for the version of SGD with control on both the row-norms and the norms of the itereates, instead of the simpler version we consider here.

Note that our observation is that SGD-type methods can control the norms of X, Y provided that the row-norms are controlled. However, the vanilla SGD cannot control the row-norms $\|X^{(i)}\|, \|Y^{(j)}\|$ (with aggresive stepsize), thus the extra row-scaling step is helpful. Figure 3.4 shows how the maximum row-norms of X_k, Y_k evolve as the algorithm SGP proceeds for 10 instances. In the first 50 iterations, the maximum row-norms are often equal to the row-norm bound, meaning that the row-scaling step is effective in the early stage of the algorithm. In the later iterations (iteration 50-200), the maximum row-norms are strictly below the row-norm bound (in very few cases equal to the bound), which implies that the row-scaling step is ineffective and SGP actually becomes SGD in most instances.



(a) Max row-norm of X_k v.s. iteration k (b) Max row-norm of Y_k v.s. iteration k

Figure 3.4: Max row-norms of the itereates generated by SGP (Stochastic Gradient Projection) for $p = 0.025$, $m = n = 1000$, $r = 10$, $\beta_T = 3.2$, $\mu = 1.9^2$. The 10 lines in 10 different colors represent 10 different experiments.

As for the parameter μ which controls the row-norms of the iterates, we find that $1.8^2 = 3.24 \leq \mu \leq 4.84 = 2.2^2$ works, and $\mu \geq 2.3^2$ leads to divergence. This choice of μ is close to the groundtruth: for a random matrix $X \in \mathbb{R}^{n\times r}$ with Gaussian entries where $n = 1000, r = 10$, the incoherence constant $\mu \approx 1.9^2$. While Figure 3.4 is generated by picking $\mu = 3.61 = 1.9^2$, we find that picking a larger μ leads to faster convergence. In particular, Figure 3.5 shows that by picking $\mu = 2.1^2$ SGP converges to a test error of < 0.1 in fewer than 100 iterations; in contrast, in Figure 3.4 where $\mu = 1.9^2$ SGP converges to a test error of < 0.1 in 150-200 iterations.

3.2 Special Initialization

For non-convex problems, good initialization may improve the performance of an algorithm in two ways: increase the convergence speed, and/or improve the quality of the convergent solution. In this section, we will present a special intialization that can greatly improve the quality of the convergent solutions of GD and SGD, for the sythesis data sets.

A popular initilization procedure is the so-called spectral method [24, 33, 93], which in our context refers to the following method: use the top r singular vectors of the

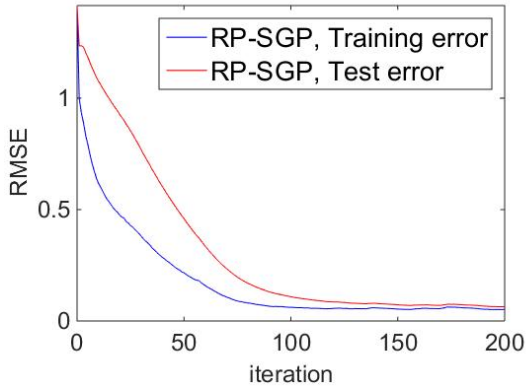


Figure 3.5: Performance of SGP for $\mu = 4.41 = 2.1^2$, $p = 0.025$, $m = n = 1000$, $r = 10$.

partial observation matrix (missing entries imputed by the mean of the matrix) as the initial point. We have shown that AltMin and GD diverge when there are not enough entries (say, $p = 0.03$ for $m = n = 1000, r = 10$). Does the spectral method save AltMin and GD in the same setting? Unfortunately, the answer is no: AltMin and GD still diverge. This is not surprising since the with so few observations, the partial observation matrix is almost full rank and the top r singular vectors seem to carry little information about the original matrix M. In fact, the distance between $\mathrm{P}_r(\mathcal{P}_\Omega(M))$ (the best rank-r approximation of $\mathcal{P}_\Omega(M)$) and M is not too much smaller than the distance between a random matrix and M. Keshavan et al. [33] suggested a trimming step that set to zero all rows and columns with too many observations (more than twice the average). Theoretically speaking, in the regime $p = O(1/n)$ and $p < O(\log n/n)$, the trimming step improves the solution produced by the spetral method [33]. However, in the setting $m = n = 1000$ and $p = 0.03 = 3r/n > 3\log n/n$, trimming does not make any difference since none of the rows and columns have too many observations.

Inspired by the analysis of the importance of incoherence in Chapter 2 and the proposed formulation (3.2), we propose a new initialization procedure that adds an additional row-scaling step after the spectral method. Denote the best rank-r approximation of a matrix A as $\mathrm{P}_r(A)$. Define an operation SVD_r that maps a matrix A to the SVD components (X, D, Y) of its best rank-r approximation $\mathrm{P}_r(A)$, i.e.

$$\mathrm{SVD}_r(A) \triangleq (X, D, Y), \text{ where } XDY^T \text{ is the compact SVD of } \mathrm{P}_r(A). \tag{3.6}$$

The details of the new initialization method is given in Table 3.4. Rather surprisingly, this method saves AltMin, GD (makes them converge to original matrix) and improves SGD (naive SGD can fail with probablity more than 20% for $p = 0.03$, now it always succeeds). See the simulation results in Figure 3.6.

Table 3.4: Initialization procedure (INITIALIZE)

Input: $\mathcal{P}_\Omega(M)$, target rank r, target row norm bounds β_1, β_2.

Algorithm INITIALIZE$(\mathcal{P}_\Omega(M), p, r)$.

1. Compute $(\bar{X}_0, D_0, \bar{Y}_0) = \text{SVD}_r\left(\frac{1}{p}\mathcal{P}_\Omega(M)\right)$, as defined in (3.6).
   Compute $\hat{X}_0 = \bar{X}_0 D_0^{1/2}, \hat{Y}_0 = \bar{Y}_0 D_0^{1/2}$.
2. For each row of $\hat{X}_0$ (resp. $\hat{Y}_0$) with norm larger than β_1 (resp. β_2), scale it to make the norm of this row equal β_1 (resp. β_2) to obtain X_0, Y_0, i.e.

$$\begin{aligned} X_0^{(i)} &= \frac{\hat{X}_0^{(i)}}{\|\hat{X}_0^{(i)}\|} \min\left\{\|\hat{X}_0^{(i)}\|, \beta_1\right\}, i = 1, \ldots, m. \\ Y_0^{(j)} &= \frac{\hat{Y}_0^{(j)}}{\|\hat{Y}_0^{(j)}\|} \min\left\{\|\hat{Y}_0^{(j)}\|, \beta_2\right\}, j = 1, \ldots, n. \end{aligned} \tag{3.7}$$

Output $X_0 \in \mathbb{R}^{m\times r}, Y_0 \in \mathbb{R}^{n\times r}$.

We have shown that empirically either the specfic initialization or directly controlling the row-norms can lead to accurate recovery when p is very small (close to the fundamental limit). Moreover, with the specfic initialization GD and SGD converge faster than directly controlling the row-norms. However, we suspect that directly controlling the row-norms is better than the specfic initialization for other generative models of M and Ω or the real data sets. This is because directly controlling the row-norms seems to be more robust and generative-model-indepent, while the success of the prposed initialization for the synthesis data sets may rely on the specific generative model. Nevertheless, experiments on more data sets are needed to support this claim.

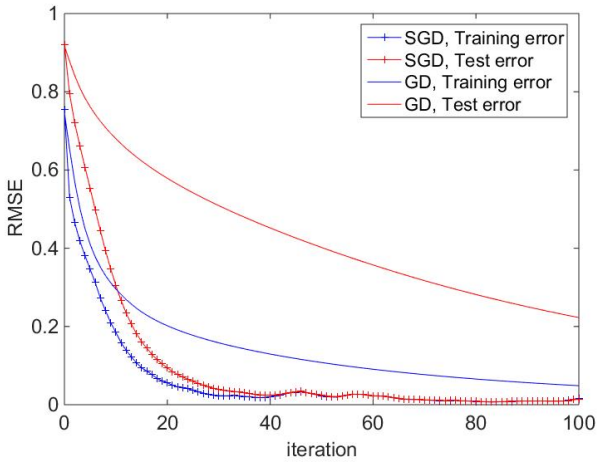


Figure 3.6: Performance of SGD and GD with the proposed initialization in Table 3.4, for $m = n = 1000$, $r = 10$, $p = 0.03$. Here SGD refers to RP-SGD, i.e. using sampling without replacement when selecting the component functions.

Chapter 4

Recovery Result

In this chapter, we provide a theoretical justification for the new MF (matrix factorization) formulation and the corresponding algorithms proposed in Chapter 3. For simplicity of analysis, we consider the regularized version of (4.9) by penalizing the constraints (3.3) and (3.1); nevertheless, our theoretical results can be easily modified to cover the constrained version (4.9) or the partially regularized version (3.5).

The algorithms for our formulation have almost the same computation cost as the algorithms for the traditional formulation with a regularizer $\lambda(\|X\|_F^2+\|Y\|_F^2)$ or without regularizer. In fact, our regularizers (or constraints) serve as a "safeguard": when p is large enough that traditional algorithms (e.g. AltMin, SGD) successfully recover M, our regularizers (or constraints) are inactive and our algorithms are the same as the traditional algorithms; when the traditional algorithms fail, our regularizers become active and save the algorithms. In some sense, our algorithms are "better" versions of the traditional naive algorithms, and our theoretical results can also be viewed as a validation of the traditional algorithms in the "large-p regime".

4.1 Formulation and Algorithms for Theoretical Analysis

To enable a rigourous theoretical analysis, we will specificially define all the parameters involved in the formulation, which requires some assumptions on the problem.

4.1.1 Assumptions

Incoherence condition. The incoherence condition for the matrix completion problem is first introduced by Candès and Recht in [2] and has become a standard assumption for low-rank matrix recovery problems (except a few recent works such as [94, 95]). We will define an incoherence condition for a $m \times n$ matrix M which is the same as that in [33].

Definition 4.1.1 *We say a matrix $M = \hat{U}\Sigma\hat{V}^T$ (compact SVD of M) is μ-incoherent if:*

$$\sum_{k=1}^{r} \hat{U}_{ik}^2 \leq \frac{\mu r}{m}, \quad \sum_{k=1}^{r} \hat{V}_{jk}^2 \leq \frac{\mu r}{n}, \quad 1 \leq i \leq m, 1 \leq j \leq n. \tag{4.1}$$

It can be shown that $\mu \in [1, \frac{\max\{m,n\}}{r}]$. For some popular random models for generating M, the incoherence condition holds with a parameter scaling as $\sqrt{r \log n}$ (see [33]). In this thesis, we just assume that M is μ-incoherent.

Note that the above incoherence condition is defined for an $m \times n$ matrix M, and $\hat{U}, \hat{V}$ are orthogonal matrices where each column has unit norm. Now we define an incoherence condition for an arbitrary $K \times r$ matrix (not necessarily orthogonal or with unit column norm).

Definition 4.1.2 *Suppose $K \in \{m, n\}$. We say $X \in R^{K \times r}$ is c-incoherent if:*

$$\|X^{(i)}\|^2 \leq c\frac{r}{K}, \; i = 1, 2, \ldots, K. \tag{4.2}$$

Here, $X^{(i)}$ denotes the i-th row of X.

One interpretation of the incoherence condition (4.1) is provided in [2]: a small μ_0 implies the incoherence between $\hat{U}$ (resp. $\hat{V}$) and the standard basis of $\mathbb{R}^m$ (resp. $\mathbb{R}^n$). In fact, (4.1) can be expressed as

$$\max_{1 \leq i \leq m} \|\mathcal{P}_{\hat{U}}(e_i)\|^2 \leq \mu_0, \max_{1 \leq j \leq n} \|\mathcal{P}_{\hat{V}}(e_j)\|^2 \leq \mu_0,$$

where $\mathcal{P}_{\hat{U}}, \mathcal{P}_{\hat{V}}$ denote the projection onto the column spaces of $\hat{U}$ and $\hat{V}$, respectively, and e_i, e_j are the standard basis vectors of $\mathbb{R}^m, \mathbb{R}^n$, respectively. This expression shows that μ_0 measures the "coherence" (in fact, the angle) between each standard basis vector

and $\mathrm{col}(\hat{\mathrm{U}}), \mathrm{col}(\hat{\mathrm{V}})$ (the column spaces of $\hat{U}, \hat{V}$). Therefore, a small μ_0 implies the angle between each standard basis vector and $\mathrm{col}(\hat{\mathrm{U}}), \mathrm{col}(\hat{\mathrm{V}})$ is large, i.e. $\mathrm{col}(\hat{\mathrm{U}}), \mathrm{col}(\hat{\mathrm{V}})$ are "incoherent" with the standard basis.

Another interpretation of the incoherence condition (4.1) is the following. Note that the sum of all row-norm-squares of $\hat{U}$ is $\|\hat{U}\|_F^2 = \sum_{i=1}^{m} \sum_{k=1}^{r} \hat{U}_{i,k}^2 = r$, thus the average of row-norm-squares of $\hat{U}$ is $\frac{r}{m}$. The first inequality of (4.1) means that each row-norm-square of $\hat{U}$ does not deviate from the average row-norm-square too much; in fact, the deviation is bounded by a factor of μ_0. Thus μ_0 is small implies that the row-norms of $\hat{U}(\hat{V})$ are evenly spread. Note that this interpretation of (4.1) does not utilize the fact that $\hat{U}$ and $\hat{V}$ are orthogonal matrices, i.e. their columns are orthogonal.

Based on this interpretation, a direct generalization of (4.1) to any $K \times r$ matrix should be to require each row-norm to be bounded by a constant times the average row-norm. However, the incoherence condition (4.2) is defined in a slightly different way: each row-norm needs to be bounded by a constant that does not depend on the average row-norm. As a result, it is possible that one row of X has norm $\frac{cr}{K}$ while all other rows are zero, still satisfying (4.2).

Random sampling model. Throughout this thesis, the probability is taken with respect to the uniform random model of $\Omega \subseteq [m] \times [n]$ with fixed size $|\Omega| = S$, i.e. Ω is generated uniformly at random from set $\{\Omega' \subseteq [m] \times [n] : \text{ the size of } \Omega' \text{ is } S\}$. Our results also hold for a Bernolli model that each entry of M is independently included in Ω with probability p. Under this model, the size of Ω is close to pmn; in fact, $pmn - C'\sqrt{n \log n} \leq |\Omega| \leq pmn + C'\sqrt{n \log n}$ for some numerical constant C', with high probability [2]. These two random models have been shown to be equivalent (i.e. any result for one random model holds for the other) except for different numerical constants [2].

4.1.2 A New Problem Formulation

We consider a regularized version of (4.9) by penalizing the constraints (3.3) and (3.1). In other words, we introduce two types of regularization terms besides the square loss function: the first type is designed to force X_k, Y_k (the produced solution in the k-th iteration) to be incoherent (i.e. with bounded row norm), and the second type is designed to upper bound the norm of X_k and Y_k. Reference [33] also used the first type

of regularizers, but not the second type; this difference with our work is mainly because their formulation forced X, Y to have a fixed norm.

The regularization function G is defined as follows:

$$G(X,Y) \triangleq \rho \sum_{i=1}^{m} G_0\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right) + \rho \sum_{j=1}^{n} G_0\left(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2}\right) + \rho G_0\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right) + \rho G_0\left(\frac{3\|Y\|_F^2}{2\beta_T^2}\right), \tag{4.3}$$

where $A^{(i)}$ denotes the ith row of a matrix A,

$$G_0(z) \triangleq I_{[1,\infty]}(z)(z-1)^2 = \max\{0, z-1\}^2, \tag{4.4}$$

$$\beta_T \triangleq \sqrt{C_T r \Sigma_{\max}},\ \beta_1 \triangleq \beta_T\sqrt{\frac{3\mu r}{m}} = \sqrt{C_T r \Sigma_{\max}}\sqrt{\frac{3\mu r}{m}},\ \beta_2 \triangleq \beta_T\sqrt{\frac{3\mu r}{n}} = \sqrt{C_T r \Sigma_{\max}}\sqrt{\frac{3\mu r}{n}}. \tag{4.5}$$

Here, $I_{\mathcal{C}}$ is the indicator function of a set $\mathcal{C}$, i.e. $I_{\mathcal{C}}(z)$ equals 1 when $z \in \mathcal{C}$ and 0 otherwise. ρ is a constant specified as follows. Throughout the thesis, δ and δ_0 are defined as

$$\delta \triangleq \frac{\Sigma_{\min}}{C_d r^{1.5}\kappa}, \quad \delta_0 \triangleq \frac{\delta}{6}, \tag{4.6}$$

where C_d is some numerical constant. The coefficient ρ is defined as (a larger ρ also works)

$$\rho \triangleq \frac{2p\delta_0^2}{G_0(3/2)} = 8p\delta_0^2. \tag{4.7}$$

The numerical constant $C_T > 5$ will be specified in the proof of our main result. The parameter β_T is chosen to be of the same order as $\|\hat{U}\Sigma^{1/2}\|_F$ and $\|\hat{V}\Sigma^{1/2}\|_F$, and β_1, β_2 are chosen to be of the same order as $\sqrt{r}\|(\hat{U}\Sigma^{1/2})^{(i)}\|, \sqrt{r}\|(\hat{V}\Sigma^{1/2})^{(j)}\|$. The additional factor $\sqrt{3r}$ is due to technical consideration (to prove (A.189)).

It is easy to verify that G_0 is continuously differentiable. The choice of function G_0 is not unique; in fact, we can choose any G_0 that satisfies the following requirements: a) G_0 is convex and continuously differentiable; b) $G_0(z) = 0, z \in [0,1]$. In [33], G_0 is chosen as $G_0(z) = I_{[1,\infty]}(z)(e^{(z-1)^2} - 1)$, which also satisfies these two requirements. Choosing different G_0 does not affect the proof except the change of numerical constants (which depend on $G_0(3/2), G_0'(3/2), G_0''(3/2)$). In the partially regularized formulation (3.5) proposed in Chapter 3, we use a different penalty function $\max\{0, \cdot\}$, which is

not differentiable at only one point 0; although our theoretical result cannot be directly applied, we believe a simple modification of our proof can cover this penalty function.

Denote the square loss term in (P0) as $F(X,Y) \triangleq \sum_{(i,j)\in\Omega}[M_{ij} - (XY^T)_{ij}]^2 = \|\mathcal{P}_\Omega(M - XY^T)\|_F^2$. Replacing the objective function of (P0) by $\tilde{F}(X,Y) \triangleq F(X,Y) + G(X,Y)$, we obtain the following problem:

$$\text{P1}: \quad \min_{X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{n\times r}} \quad \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 + G(X,Y). \tag{4.8}$$

For G_0 given in (4.4), (P1) can be interpreted as the penalized version of the constrained problem (3.4), which we restate below (with slightly different thresholds):

$$\begin{aligned} \min_{X,Y} \quad & \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2, \\ \text{s.t.} \quad & \|X\|_F^2 \le \frac{2}{3}\beta_T^2, \quad \|Y\|_F^2 \le \frac{2}{3}\beta_T^2; \\ & \|X^{(i)}\|^2 \le \frac{2}{3}\beta_1^2, \ \forall\ i, \quad \|Y^{(j)}\|^2 \le \frac{2}{3}\beta_2^2, \ \forall\ j. \end{aligned} \tag{4.9}$$

Let us illustrate the connection of (4.9) and (P1). The constraint $f_1(X) \triangleq \frac{3\|X\|_F^2}{2\beta_T^2} - 1 \le 0$ corresponds to the penalty term $\rho G_0(f_1(X)+1) = \rho\max\{0, f_1(X)\}^2$ which appears as the third term in $G(X,Y)$; similarly, other constraints in (4.9) correspond to other terms in $G(X,Y)$. In other words, the regularization function $G(X,Y)$ is just a penalty function for the constraints of the problem (3.4). The function $\max\{0,\cdot\}^2$ is a popular choice for the penalty function in optimization (see, e.g. [96]), which motivates our choice of G_0 in (4.4).

It is easy to check that the optimal value of (P1) is zero and $(X,Y) = (\hat{U}\Sigma^{1/2}, \hat{V}\Sigma^{1/2})$ is an optimal solution to (P1), provided that M is μ-incoherent. In fact, since $\tilde{F}$ is a nonnegative function, we only need to show $\tilde{F}(X,Y) = 0$ for this choice of (X,Y). As $XY^T = M$ implies $\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 = 0$, we only need to show $G(X,Y) = G(\hat{U}\Sigma^{1/2}, \hat{V}\Sigma^{1/2})$ equals zero. In the expression of $G(X,Y)$, the third and fourth terms $G_0(\frac{3\|X\|_F^2}{2\beta_T^2})$ and $G_0(\frac{3\|Y\|_F^2}{2\beta_T^2})$ equal zero because $\|X\|_F^2 = \|Y\|_F^2 \le r\Sigma_{\max} < \frac{2}{3}\beta_T^2$. The first and second terms $\sum_i G_0(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})$ and $\sum_j G_0(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})$ equal zero because $\|X^{(i)}\|^2 \le \Sigma_{\max}\|\hat{U}^{(i)}\|^2 \le \Sigma_{\max}\frac{\mu r}{m} \le \frac{2}{3}\beta_1^2$, for all i and, similarly, $\|Y^{(j)}\|^2 \le \frac{2}{3}\beta_2^2$, for all j, where we have used the incoherence condition (4.1).

One commonly used assumption in the optimization literature is that the gradient of the objective function is Lipschitz continuous. For any positive number β, define a

bounded set

$$\Gamma(\beta) \triangleq \{(X,Y) | X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \|X\|_F \leq \beta, \|Y\|_F \leq \beta\}. \tag{4.10}$$

The following result shows that this assumption (Lipschitz continuous gradients) holds for our objective function within a bounded set.

Claim 4.1.1 *Suppose* $\beta_0 \geq \beta_T$ *and*

$$L(\beta_0) \triangleq 4\beta_0 + 54\rho\frac{\beta_0^2}{\beta_T^4}. \tag{4.11}$$

Then $\nabla\tilde{F}(X,Y)$ *is Lipschitz continuous over the set* $\Gamma(\beta_0)$ *with Lipschitz constant* $L(\beta_0)$*, i.e.*

$$\|\nabla\tilde{F}(X,Y) - \nabla\tilde{F}(U,V)\|_F \leq L(\beta_0)\|(X,Y)-(U,V)\|_F, \ \ \forall (X,Y),(U,V) \in \Gamma(\beta_0),$$

where $\|(X,Y)-(U,V)\|_F = \sqrt{\|X-U\|_F^2 + \|Y-V\|_F^2}$.

The proof of Claim 4.1.1 is given in Appendix A.1.

4.1.3 Initialization

For technical reasons, our results require the initial point to be close enough to the global optima. To be more precise, we want the initial point to be in an incoherent neighborhood of the original matrix M (this neighborhood will be specified later). Special initialization is also required in other works on non-convex formulations [24, 33, 54–56, 66, 97]. As discussed in Chapter 3, in our simulations a special initialization is not necessary since the proposed algorithms with random initial points do converge; nevertheless, a special initialization does help improve the convergence speed.

We will show that such an initial point can be found through a simple procedure. This procedure consists of two steps: first, using the spectral method (see, e.g. [33]), we obtain $M_0 = \hat{X}_0\hat{Y}_0^T$ which is close to M; second, we modify $(\hat{X}_0, \hat{Y}_0)$ to make it incoherent (i.e. with bounded row norm). Denote the best rank-r approximation of a matrix A as $\mathrm{P}_r(A)$. Define an operation SVD_r that maps a matrix A to the SVD components (X, D, Y) of its best rank-r approximation $\mathrm{P}_r(A)$, i.e.

$$\mathrm{SVD}_r(A) \triangleq (X,D,Y), \text{ where } XDY^T \text{ is the compact SVD of } \mathrm{P}_r(A). \tag{4.12}$$

The initialization procedure is given in Table 4.1; note that this is almost the same as Table 3.4, except that the scaling threshold is slightly different. As mentioned before, the row-scaling step is crucial in our experiments since simply initializing via the spectral method does not improve the recovery performance of the algorithms. The property of the initial point generated by this procedure will be presented in Claim 4.4.2.

Table 4.1: Initialization procedure (INITIALIZE)

Input: $\mathcal{P}_\Omega(M)$, target rank r, target row norm bounds β_1, β_2.

Algorithm INITIALIZE($\mathcal{P}_\Omega(M), p, r$).

1. Compute $(\bar{X}_0, D_0, \bar{Y}_0) = \mathrm{SVD}_r\left(\frac{1}{p}\mathcal{P}_\Omega(M)\right)$, as defined in (4.12).
   Compute $\hat{X}_0 = \bar{X}_0 D_0^{1/2}, \hat{Y}_0 = \bar{Y}_0 D_0^{1/2}$.
2. For each row of $\hat{X}_0$ (resp. $\hat{Y}_0$) with norm larger than $\sqrt{\frac{2}{3}}\beta_1$ (resp. $\sqrt{\frac{2}{3}}\beta_2$), scale it to make the norm of this row equal $\sqrt{\frac{2}{3}}\beta_1$ (resp. $\sqrt{\frac{2}{3}}\beta_2$) to obtain X_0, Y_0, i.e.

$$\begin{aligned} X_0^{(i)} &= \frac{\hat{X}_0^{(i)}}{\|\hat{X}_0^{(i)}\|} \min\left\{\|\hat{X}_0^{(i)}\|, \sqrt{\frac{2}{3}}\beta_1\right\}, i = 1, \ldots, m. \\ Y_0^{(j)} &= \frac{\hat{Y}_0^{(j)}}{\|\hat{Y}_0^{(j)}\|} \min\left\{\|\hat{Y}_0^{(j)}\|, \sqrt{\frac{2}{3}}\beta_2\right\}, j = 1, \ldots, n. \end{aligned} \tag{4.13}$$

Output $X_0 \in \mathbb{R}^{m\times r}, Y_0 \in \mathbb{R}^{n\times r}$.

4.1.4 Standard Algorithms for the New Formulation

Our result applies to many standard algorithms such as gradient descent, SGD and block coordinate descent type methods (including alternating minimization, block coordinate gradient descent, block successive upper bound minimization, etc.). We will describe several typical algorithms in this subsection.

The gradient $\nabla\tilde{F} = \nabla F + \nabla G = (\nabla_X F + \nabla_X G, \nabla_Y F + \nabla_Y G)$ can be easily computed

as follows:

$$\begin{aligned}
\nabla_X F(X,Y) &= \mathcal{P}_\Omega(XY^T - M)Y, \\
\nabla_Y F(X,Y) &= \mathcal{P}_\Omega(XY^T - M)^T X, \\
\nabla_X G(X,Y) &= \rho \sum_{i=1}^{m} G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})\frac{3\bar{X}^{(i)}}{\beta_1^2} + \rho G_0'(\frac{3\|X\|_F^2}{2\beta_T^2})\frac{3X}{\beta_T^2}, \\
\nabla_Y G(X,Y) &= \rho \sum_{j=1}^{n} G_0'(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})\frac{3\bar{Y}^{(j)}}{\beta_2^2} + \rho G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3Y}{\beta_T^2},
\end{aligned} \tag{4.14}$$

where $G_0'(z) = I_{[1,\infty]}(z)2(z-1)$, and $\bar{X}^{(i)}$ (resp. $\bar{Y}^{(j)}$) denotes a matrix with the i-th (resp. j-th) row being $X^{(i)}$ (resp. $Y^{(j)}$) and the other rows being zero.

We first present a gradient descent algorithm in Table 4.2. There are many choices of stepsizes such as constant stepsize, exact line search, limited line search, diminishing stepsize and Armijo rule [86]. We present three stepsize rules here: constant stepsize, restricted Armijo rule and restricted line search (the latter two are the variants of Armijo rule and exact line search). Note that the restricted line search rule is similar to that used in [33] for the gradient descent method over Grassmannian manifolds. In Chapter 2 we use BB stepsize for GD which perform very well in practice (see Table 2.5); here we do not present the BB stepsize since the convergence analysis of the BB method for non-convex problems is not that straightforward and left as future work. To simplify the notations, we denote $\boldsymbol{x}_k(\eta) \triangleq (X_k(\eta), Y_k(\eta))$ and $d(\boldsymbol{x}_k(\eta), \boldsymbol{x}_0) \triangleq \sqrt{\|X_k(\eta) - X_0\|_F^2 + \|Y_k(\eta) - Y_0\|_F^2}$.

Alternating minimization belongs to the class of block coordinate descent (BCD) type methods. As mentioned in Section 2.1, there are many BCD-type methods, and some of them have been applied to matrix completion. Our result applies to many BCD-type methods, including the two-block alternating minimization, BCGD and BSUM. While it is not very interesting to list all possible algorithms to which our results are applicable, we just present two specific algorithms for illustration.

The first BCD type algorithm we present is (two-block) alternating minimization, which, in the context of matrix completion, usually refers to the algorithm that alternates between X and Y by updating one factor at a time with the other factor fixed. Although the overall objective function is non-convex, each subproblem of X or Y is convex and thus can be solved efficiently. The details are given in Table 4.3.

Table 4.2: Algorithm 1 (Gradient descent)

Initialization: $(X_0, Y_0) \longleftarrow \textsc{Initialize}(\mathcal{P}_\Omega(M), p, r)$.
The k-th iteration:

$$X_k \longleftarrow X_k(\eta_k) \triangleq X_{k-1} - \eta_k \nabla_X \tilde{F}(X_{k-1}, Y_{k-1}),$$
$$Y_k \longleftarrow Y_k(\eta_k) \triangleq Y_{k-1} - \eta_k \nabla_Y \tilde{F}(X_{k-1}, Y_{k-1}),$$

where the stepsize η_k is chosen according to one of the following rules:

a) Constant stepsize: $\eta_k = \eta < \bar{\eta}_1,\ \forall\ k$ ($\bar{\eta}_1$ is a constant specified in Appendix A.5.4).

b) Restricted Armijo rule: Let $\sigma \in (0,1), \xi \in (0,1), s_0$ be fixed scalars.

b1) Find the smallest nonnegative integer i such that $d(\boldsymbol{x}_k(\xi^i s_0), \boldsymbol{x}_0) \leq 5\delta/6$ and $\tilde{F}(\boldsymbol{x}_k(\xi^i s_0)) \leq \tilde{F}(\boldsymbol{x}_{k-1}) - \sigma \xi^i s_0 \|\nabla \tilde{F}(\boldsymbol{x}_{k-1})\|_F^2$.

b2) Let $\eta_k = \xi^i s_0$.

c) Restricted line search: $\eta_k = \arg\min_{\eta \in \mathbb{R}, d(\boldsymbol{x}_k(\eta), \boldsymbol{x}_0) \leq 5\delta/6} \tilde{F}(\boldsymbol{x}_k(\eta))$.

Table 4.3: Algorithm 2 (Two-block Alternating Minimization)

Initialization: $(X_0, Y_0) \longleftarrow \textsc{Initialize}(\mathcal{P}_\Omega(M), p, r)$.
The k-th iteration:

$$X_k \longleftarrow \arg\min_X \tilde{F}(X, Y_{k-1}),$$
$$Y_k \longleftarrow \arg\min_Y \tilde{F}(X_{k-1}, Y).$$

For the case without the regularization term $G(X,Y)$, the objective function becomes $F(X,Y)$ and is quadratic with respect to X or Y. Thus X_k, Y_k have closed form update. Suppose $X^T = (x_1, \ldots, x_m)$ and $Y^T = (y_1, \ldots, y_n)$, where $x_i, y_j \in \mathbb{R}^{r \times 1}$. Then $(x_1^*, \ldots, x_m^*) \triangleq (\arg\min_X F(X,Y))^T$ and $(y_1^*, \ldots, y_n^*) \triangleq (\arg\min_Y F(X,Y))^T$ are given by

$$\begin{aligned} x_i^* &= (\sum_{j \in \Omega_i^x} y_j y_j^T)^\dagger (\sum_{j \in \Omega_i^x} M_{ij} y_j),\ i = 1, \ldots, m, \\ y_j^* &= (\sum_{i \in \Omega_j^y} x_i x_i^T)^\dagger (\sum_{i \in \Omega_j^y} M_{ij} x_i),\ j = 1, \ldots, n, \end{aligned} \tag{4.15}$$

where $\Omega_i^x = \{j \mid (i,j) \in \Omega\}, \Omega_j^y = \{i \mid (i,j) \in \Omega\}$, and $A^\dagger$ denotes the pseudo inverse of a matrix A. For our problem with the regularization term $G(X,Y)$, we no longer have closed form update of X_k, Y_k. One way to solve the convex subproblems is to start from the solution given in (4.15) and then perform gradient update until convergence. The details for solving $\min_X \tilde{F}(X,Y)$ is given in Table 4.4 (the stepsize can be chosen by one of the standard rules of the gradient descent method), and the other subproblem

$\min_Y \tilde{F}(X,Y)$ can be solved in a similar fashion.

Theoretically speaking, AltMin for our formulation (P1) is not as efficient as the vanilla AtlMin for (P0) since an extra inner loop is needed to solve the subproblem. However, we remark that in the regimes of p that the vanilla AltMin works, the least square solution X (resp. Y) is always bounded and incoherent (empirical observation), in which case the regularizer G is inactive; therefore, the gradient updates in Table 4.4 do not happen. In the regimes of p that the vanilla AltMin fails, G is active and the gradient updates do happen; however, instead of solving the subproblem exactly, one could perform one gradient step and the algorithm becomes the popular variant BCGD [70]. Our main result of exact recovery still holds for BCGD (the proof for Algorithm 3 in Claim 4.4.3 can be applied to BCGD since BCGD is a special case of BSUM).

Table 4.4: Solving subproblem of Algorithm 2

Solving subproblem of Algorithm 2: $\min_X \tilde{F}(X,Y)$.
Input: $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{n\times r}$.
Initialization: $X = (x_1, \ldots, x_m)$, where $x_i = (\sum_{j\in\Omega_i^x} y_j y_j^T)^\dagger (\sum_{j\in\Omega_i^x} M_{ij} y_j)$, $i = 1, \ldots, m$,
Repeat:
$X \longleftarrow X - \eta \nabla_X \tilde{F}(X,Y)$,
Until Stopping criterion is met.

In the second BCD type algorithm called row BSUM, we update the rows of X and Y cyclically by minimizing an upper bound of the objective function; see Table 4.5. The extra terms $\frac{\lambda_0}{2}\|X^{(i)} - X_{k-1}^{(i)}\|^2$ or $\frac{\lambda_0}{2}\|Y^{(j)} - Y_{k-1}^{(j)}\|^2$ are added to make the subproblems strongly convex, which help prove convergence to stationary points. Such a technique has also been used in the alternating least square algorithm for tensor decomposition [82]. Note that for the two-block BCD algorithm, convergence to stationary points can be guaranteed even when the subproblems are not strongly convex [98], thus in Algorithm 2 we do not add the extra terms. The benefit of cyclically updating the rows is that each subproblem can be solved efficiently using a simple binary search; see Appendix A.2 for the details. We remark again that instead of solving the subproblem exactly, one could just perform one gradient step to update each row of X and Y (with $\lambda = 0$) and our result still holds.

Table 4.5: Algorithm 3 (Row BSUM)

Initialization: $(X_0, Y_0) \longleftarrow$ INITIALIZE$(\mathcal{P}_\Omega(M), p, r)$.
Parameter: $\lambda_0 > 0$.
The k-th loop:
For $i = 1$ to m:
$X_k^{(i)} \longleftarrow \arg\min_{X^{(i)}} \tilde{F}(X_k^{(1)}, \ldots, X_k^{(i-1)}, X^{(i)}, X_{k-1}^{(i+1)} \ldots, X_{k-1}^{(m)}, Y_{k-1}) + \frac{\lambda_0}{2}\|X^{(i)} - X_{k-1}^{(i)}\|^2$,
For $j = 1$ to n:
$Y_k^{(j)} \longleftarrow \arg\min_{Y^{(j)}} \tilde{F}(X_k, Y_k^{(1)}, \ldots, Y_k^{(j-1)}, Y^{(j)}, Y_{k-1}^{(j+1)}, \ldots, Y_{k-1}^{(m)}) + \frac{\lambda_0}{2}\|Y^{(j)} - Y_{k-1}^{(j)}\|^2$.

The fourth algorithm we present is SGD (stochastic gradient descent) [1,59] tailored for our problem (P1). In SGD, at each time we pick a component function and perform a gradient update. Similar to the BCD type methods where the blocks can be chosen in different orders, one can pick the component functions in a cyclic order, in an essentially cyclic order, or in a random order (either sampling with or without replacement). In this thesis we only consider the cyclic order. The objective function $\tilde{F}(X, Y)$ can be decomposed as follows:

$$\tilde{F}(X,Y) = \sum_{(i,j)\in\Omega} F_{ij}(X,Y) + \sum_{i=1}^{m} G_{1i}(X) + \sum_{j=1}^{n} G_{2j}(Y) + G_3(X) + G_4(Y)$$
$$= \sum_{k=1}^{|\Omega|+m+n+2} f_k(X,Y),$$

where the component functions

$$F_{ij}(X,Y) = [(XY^T - M)_{ij}]^2 = [(X^{(i)})^T Y^{(j)} - M_{ij}]^2,\ (i,j) \in \Omega,$$
$$G_{1i}(X) = \rho G_0(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}), 1 \le i \le m,\ G_{2j}(Y) = \rho G_0(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2}), 1 \le j \le n, \quad (4.16)$$
$$G_3(X) = \rho G_0(\frac{3\|X\|_F^2}{2\beta_T^2}),\ G_4(Y) = \rho G_0(\frac{3\|Y\|_F^2}{2\beta_T^2})$$

and $\{f_k(X,Y)\}_{k=1}^{|\Omega|+m+n+2}$ denotes the collection of all component functions. With these definitions, the SGD algorithm is given in Table (4.6). Note that we use a standard stepsize rule for SGD [90,99] which requires the stepsizes $\{\eta_k\}$ go to zero as $k \to \infty$, but neither too fast nor too slow (this choice guarantees convergence to stationary points

even for nonconvex problems). One such choice of stepsizes is $\eta_k = O(1/k)$. In our simulations, choosing a constant stepsize seems to work quite well; the analysis of the constant stepsize SGD is left as future work.

Table 4.6: Algorithm 4 (SGD)

Initialization: $(X_0, Y_0) \longleftarrow \text{INITIALIZE}(\mathcal{P}_\Omega(M), p, r)$.
Parameters: $\eta_k, k = 0, 1, \ldots$ satisfying $\sum_k \eta_k = \infty, \sum_k \eta_k^2 < \eta_{\text{sum}}$ and $0 < \eta_k \leq \bar{\eta}$, where η_{sum} and $\bar{\eta}$ are constants specified in Appendix A.5.4.
The $(k+1)$-th loop:
$X_{k,0} \longleftarrow X_k, \quad Y_{k,0} \longleftarrow Y_k$.
For $i = 1$ to $|\Omega| + m + n + 2$:
$X_{k,i} \longleftarrow X_{k,i-1} - \eta_k \nabla_X f_i(X_{k,i-1}, Y_{k,i-1})$,
$Y_{k,i} \longleftarrow Y_{k,i-1} - \eta_k \nabla_Y f_i(X_{k,i-1}, Y_{k,i-1})$.
End
$X_{k+1} \longleftarrow X_{k,|\Omega|+m+n+2}, \quad Y_{k+1} \longleftarrow Y_{k,|\Omega|+m+n+2}$.

4.2 Main Results

Our main result is that Algorithms 1-4 (standard optimization algorithms) will converge to the global optima of problem (P1) given in (4.8) and reconstruct M exactly with high probability, provided that the number of revealed entries is large enough. Similar to the results for nuclear norm minimization [2–5], the probability is taken with respect to the random choice of Ω, and "with probability 99%" means that out of all possible sets Ω with a given size, 99% of them can lead to exact reconstruction by Algorithm 1-4 *for sure*. We will prove this theorem in Section 4.2.1.

Theorem 4.2.1 *(Exact Recovery) Assume a rank-r matrix $M \in \mathbb{R}^{m \times n}$ is μ-incoherent. Suppose the condition number of M is κ and $\alpha = m/n \geq 1$. Then there exists a numerical constant C_0 such that: if Ω is uniformly generated at random with size*

$$|\Omega| \geq C_0 \alpha n r \kappa^2 \max\{\mu \log n, \sqrt{\alpha}\mu^2 r^6 \kappa^4\}, \tag{4.17}$$

then with probability at least $1 - 2/n^4$, each of Algorithms 1-4 reconstructs M exactly. Here, we say an algorithm reconstructs M if each limit point (X^*, Y^*) of the sequence $\{X_k, Y_k\}$ generated by this algorithm satisfies $X^*(Y^*)^T = M$.

This result is rather surprising since it applies to the non-convex formulation (4.8), as opposed to the convex formulation considered in [2]. Different from all previous works on alternating minimization for matrix completion, our result does not require the algorithm to use independent samples in different iterations. To the best of our knowledge, our result is the first one that provides theoretical guarantee for alternating minimization without resampling. In addition, this result also provides the first exact recovery guarantee for many algorithms such as gradient descent, SGD and BSUM.

As demonstrated in [2] (and proved in [3, Theorem 1.7]), $O(nr \log n)$ entries are the minimum requirement to recover the original matrix: $O(nr)$ is the number of degrees of freedom of a rank r matrix M, and the additional $\log n$ factor is due to the coupon collector effect [2]. For $r = O(1)$ and κ bounded, Theorem 4.2.1 is order optimal in terms of the sample complexity since only $O(n \log n)$ entries are needed to exactly recover M. For $r = \mathcal{O}(\log n)$, however, our result is suboptimal by a polylogarithmic factor. The initialization has contributed $r^4\kappa^4$ to the sample complexity bound, and we expect that using other initialization procedures (e.g. the one proposed in [55]) can reduce the exponents of r and κ.

Theorem 4.2.1 only establishes the convergence, but not the convergence speed. With some extra effort, we can prove the linear convergence of the gradient descent method (see Theorem 4.2.2 below). Again, this result can be extended beyond the gradient descent method. In fact, by a standard optimization argument, we can prove the linear convergence of any algorithm that satisfies "sufficient decrease" (i.e. $\tilde{F}(\boldsymbol{x}^k) - \tilde{F}(\boldsymbol{x}^{k+1}) \geq O(\|\nabla \tilde{F}(\boldsymbol{x}^k)\|_F^2)$) and the requirements in Lemma 4.2.2; see Corollary 4.2.2. Many first order methods, including alternating type methods (e.g. BCGD, two-block BCD), can be shown to have the sufficient decrease property under mild conditions. For space reason, we do not verify all these methods in this thesis, but only present the linear convergence result for the gradient descent method. The proof of Theorem 4.2.2 is given in Section 4.2.2.

Theorem 4.2.2 *(Linear convergence) Under the same condition of Theorem 4.2.1, with probability at least* $1 - 2/n^4$*, Algorithm 1a (gradient descent with constant stepsize) converges linearly; more precisely, the sequence* $\{X_k, Y_k\}$ *generated by Algorithm 1a*

satisfies

$$\tilde{F}(X_k, Y_k) \leq (1 - \frac{1}{2}\eta_1 \xi)^k, \tag{4.18}$$

where $\xi = \frac{1}{C_g r^5 \kappa^3} p\Sigma_{\min}$ *(here* C_g *is a numerical constant),* η_1 *is the stepsize and* $\eta_1 \xi < 1$.

The linear convergence will immediately lead to a time complexity of $\tilde{O}(\text{poly}(n) \log \frac{1}{\epsilon})$ for achieving any ϵ-optimal solution, where the $\tilde{O}$ notation hides factors polynomial in r, κ, α. We believe that the time complexity bound can be improved to $\tilde{O}(|\Omega| \log(1/\epsilon))$ as observed in practice. However, finding the optimal time complexity bound is not the focus of this thesis, and is left as future work.

The above result shows that $\tilde{F}(X_k, Y_k)$ converges to zero at a linear speed. Note that $\tilde{F}(X, Y) = 0$ (global convergence) only implies $\mathcal{P}_\Omega(M - XY^T) = 0$, not necessarily $M = XY^T$ (exact reconvery). The following lemma implies that with high probability (for random Ω) the global convergence implies the exact recovery, or equivalently, the training error equals zero implies that the test error equals zero. In fact, it shows that the observed loss $\|\mathcal{P}_\Omega(M - XY^T)\|_F^2$ is on the order of the recovery error $p\|M - XY^T\|_F^2$ if (X, Y) lies in an incoherent neighborhood of M.

Claim 4.2.1 *Under the same condition of Theorem 4.2.1, with probability at least* $1 - 1/(2n^4)$*, we have*

$$\frac{1}{3} p\|M - XY^T\|_F^2 \leq \|\mathcal{P}_\Omega(M - XY^T)\|_F^2 \leq 2p\|M - XY^T\|_F^2, \quad \forall (X, Y) \in K_1 \cap K_2 \cap K(\delta). \tag{4.19}$$

The proof of this claim is given in Appendix A.5.2. This result is a simple corollary of several intermediate bounds established in the proof of Lemma 4.2.1.

4.2.1 Proof of Theorem 4.2.1 and main lemmas

To prove Theorem 4.2.1, we only need to prove two lemmas which describe the properties of the problem formulation (P1) and the properties of the algorithms respectively. Roughly speaking, the first lemma shows that any stationary point of (P1) in a certain region is globally optimal, and the second lemma shows that each of Algorithms 1-4 converges to stationary points in that region. This region can be viewed as an "incoherent neighborhood" of M, and can be formally defined as $K_1 \cap K_2 \cap K(\delta)$, where K_1, K_2

are defined as

$$K_1 \triangleq \{(X,Y)|X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \|X^{(i)}\| \le \beta_1, \|Y^{(j)}\| \le \beta_2, \forall i \in [m], j \in [n]\},$$
$$K_2 \triangleq \{(X,Y)|X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \|X\|_F \le \beta_T, \|Y\|_F \le \beta_T\}. \tag{4.20}$$

and $K(\delta)$ is defined as

$$K(\delta) \triangleq \{(X,Y)|X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}, \|M - XY^T\|_F \le \delta\}. \tag{4.21}$$

Note that $K_2 = \Gamma(\beta_T)$ by our definition of Γ in (4.10)

The first lemma describes the property of the problem formulation (P1) and is stated below. An immediate corollary of this result is that any stationary point (X,Y) in $K_1 \cap K_2 \cap K(\delta)$ satisfies $XY^T = M$. The proof of Lemma 4.2.1 will be given in Section 4.3.

Lemma 4.2.1 *There exist numerical constants C_0, C_d such that the following holds. Assume δ is defined by* (4.6) *and Ω is uniformly generated at random with size $|\Omega|$ satisfying* (4.17)*. Then, with probability at least $1 - 1/n^4$, the following holds: for all $(X,Y) \in K_1 \cap K_2 \cap K(\delta)$, there exist $U \in \mathbb{R}^{m\times r}, \mathbb{V} \in R^{n\times r}$, such that $UV^T = M$ and*

$$\langle \nabla_X \tilde{F}(X,Y), X - U\rangle + \langle \nabla_Y \tilde{F}(X,Y), Y - V\rangle \ge \frac{p}{4}\|M - XY^T\|_F^2. \tag{4.22}$$

Note that (4.22) can be transformed to (1.20) since we will choose U, V such that $\|U - X\|_F^2 + \|V - Y\|_F^2$ is in the same order of $\|M - XY^T\|_F^2$ (see Corollary 4.3.1). As discussed after (1.20), this property is related to local strong convexity, but actually quite different. A more precise description of (4.22) might be the so-called "cost-to-go estimate" in optimization litearature; see Lemma 4.2.3.

The second lemma describes the properties of the algorithms we presented. Throughout the thesis, "under the same condition of Lemma 4.2.1" means "assume δ is defined by (4.6) and Ω is uniformly generated at random with size $|\Omega|$ satisfying (4.17), where C_0, C_d are the same numerical constants as those in Lemma 4.2.1". The proof of Lemma 4.2.2 will be given in Section 4.4.

Lemma 4.2.2 *Under the same conditions of Lemma 4.2.1, with probability at least $1 - 1/n^4$, the sequence (X_k, Y_k) generated by either of Algorithms 1-4 has the following*

properties:
(a) Each limit point of (X_k, Y_k) *is a stationary point of (P1).*
(b) $(X_k, Y_k) \in K_1 \cap K_2 \cap K(\delta),\ \forall k \geq 0.$

Intuitively, $\|X_k^{(i)}\|, \|Y_k^{(j)}\|, \|X_k\|_F, \|Y_k\|_F$ are bounded because of the regularization terms we introduced and that the objective function is decreasing, and $\|M - X_k Y_k^T\|_F$ is bounded because the objective function is decreasing (however, the intuition is not quite correct and the proof requires some extra effort). In Section 4.4 we provide some easily verifiable conditions for Property (b) to hold (see Proposition 4.4.1), so that Lemma 4.2.2 and Theorem 4.2.1 can be extended to other algorithms.

With these two lemmas, the proof of Theorem 4.2.1 is quite straightforward and presented below.

Proof of Theorem 4.2.1: Consider any limit point (X_*, Y_*) of sequence $\{(X_k, Y_k)\}$ generated by either of Algorithms 1-4. According to Property (a) of Lemma (4.2.2), (X_*, Y_*) is a stationary point of problem (P1), i.e. $\nabla_X \tilde{F}(X_*, Y_*) = 0, \nabla_Y \tilde{F}(X_*, Y_*) = 0$. According to Property (b) of Lemma 4.2.2, with probability at least $1 - 1/n^4$, $(X_k, Y_k) \in K_1 \cap K_2 \cap K(\delta)$ for all k, implying $(X_*, Y_*) \in K_1 \cap K_2 \cap K(\delta)$. Then we can apply Lemma 4.2.1 by plugging $(X, Y) = (X^*, Y^*)$ into (4.22) to conclude that with probability at least $1 - 2/n^4$, $\|M - X_* Y_*^T\|_F \leq 0$, i.e. $X_* Y_*^T = M$. $\square$

Remark: Note that $X_* Y_*^T = M$ does not necessarily imply the global optimality of (X_*, Y_*) since we have not proved $G(X_*, Y_*) = 0$. The global optimality can be proved using a different version of Lemma 4.2.1; see the discussion before Lemma 4.2.3.

The same argument can be used to show a more general result than Theorem 4.2.1, as stated in the following corollary.

Corollary 4.2.1 *Under the same conditions of Theorem 4.2.1, any algorithm satisfying Properties (a) and (b) in Lemma 4.2.2 reconstructs* M *exactly with probability at least* $1 - 2/n^4$.

4.2.2 Proof of Theorem 4.2.2

The proof of Theorem 4.2.2 applies a standard framework for first order methods: the convergence rate (or iteration complexity) can be derived from the "cost-to-go estimate" and the "sufficient decrease" condition. For instance, the linear convergence $f(\boldsymbol{x}_k) -$

$f^* \leq (1-c_1c_2)^k$ is a direct corollary of the cost-to-go estimate $\|\nabla f(\boldsymbol{x}_k)\|^2 \geq c_1[f(\boldsymbol{x}_k) - f^*]$ and the sufficient decrease condition $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq c_2\|\nabla f(\boldsymbol{x}_k)\|^2$, where f^* is the minimum value of f, and c_1, c_2 are certain constants.

For our problem, a variant of Lemma 4.2.1 can be viewed as the cost-to-go estimate; see Lemma 4.2.3 below. One difference with Lemma 4.2.1 is the following: for a stationary point (X_*, Y_*) that $\nabla\tilde{F}(X_*, Y_*) = 0$, Lemma 4.2.3 implies $\tilde{F}(X_*, Y_*) = 0$ (global optimality), but Lemma 4.2.1 implies $M = X_*Y_*^T$ (exact recovery). The relation between these two lemmas is that Lemma 4.2.3 is a direct consequence of (A.184), a slightly stronger version of Lemma 4.2.1. The main difficulty of proving Lemma 4.2.3 is the same as that of proving Lemma 4.2.1 and lies in Proposition 4.3.1 and Proposition 4.3.2; see the formal proof of Lemma 4.2.3 in Appendix A.6.

Lemma 4.2.3 *(Cost-to-go estimate) Under the same conditions of Lemma 4.2.1, with probability at least* $1 - 1/n^4$*, the following holds:*

$$\|\nabla\tilde{F}(X,Y)\|_F^2 \geq \xi\tilde{F}(X,Y),\ \forall\ (X,Y) \in K_1 \cap K_2 \cap K(\delta), \tag{4.23}$$

where $\xi = \frac{1}{C_g r^5\kappa^3} p\Sigma_{\min}$ *(here* $C_g \geq 1$ *is a numerical constant).*

The following claim shows that Algorithm 1a satisfies the sufficient decrease condition.

Claim 4.2.2 *(Sufficient decrease) For the sequence* $\boldsymbol{x}_k = (X_k, Y_k)$ *generated by Algorithm 1a (gradient descent with constant stepsize), we have*

$$\tilde{F}(\boldsymbol{x}_k) - \tilde{F}(\boldsymbol{x}_{k+1}) \geq \frac{\eta_1}{2}\|\nabla\tilde{F}(\boldsymbol{x}_k)\|_F^2, \tag{4.24}$$

where η_1 *is the stepsize bounded above by* $\bar{\eta_1}$ *defined in* (A.170).

Claim 4.2.2 is easy to prove: it is well known that for minimizing a function (possibly non-convex) with Lipschitz continuous gradient, the gradient descent method with constant step-size satisfies the sufficient decrease condition. In the proof of Claim 4.4.3, let $\lambda = 1$ in (A.171) we immediately obtain (4.24), which proves Claim 4.2.2.

The linear convergence can be easily derived from Lemma 4.2.1 and Claim 4.2.2. For completeness, we present the proof below.

Proof of Theorem 4.2.2: According to Property (b) of Lemma 4.2.2, with probability at least $1 - 1/n^4$, $(X_k, Y_k) \in K_1 \cap K_2 \cap K(\delta)$ for all k. According to Lemma 4.2.3 and

Claim 4.2.2, we have (with probability at least $1 - 2/n^4$)

$$\tilde{F}(\boldsymbol{x}_k) - \tilde{F}(\boldsymbol{x}_{k+1}) \geq \frac{\eta_1}{2}\|\nabla \tilde{F}(\boldsymbol{x}_k)\|_F^2 \geq \frac{\eta_1}{2}\xi \tilde{F}(\boldsymbol{x}_k),\ \forall k.$$

This relation can be rewritten as

$$\tilde{F}(\boldsymbol{x}_{k+1}) \leq (1 - \frac{1}{2}\eta_1 \xi)\tilde{F}(\boldsymbol{x}_k),\ \forall\ k. \tag{4.25}$$

The stepsize η_1 can be bounded as $0 < \eta_1 \leq \bar{\eta}_1 \overset{(A.167)}{\leq} \frac{1}{4\beta_T^2} = \frac{1}{4C_T r \Sigma_{\max}} \leq \frac{1}{\Sigma_{\max}}$. Since $0 < \xi = \frac{1}{C_g r^5 \kappa^3} p \Sigma_{\min} \leq \Sigma_{\min}$, we have $0 < \eta_1 \xi \leq \frac{\Sigma_{\min}}{\Sigma_{\max}} \leq 1$, which implies $0 < 1 - \frac{1}{2}\eta_1\xi < 1$. Then the relation (4.25) leads to

$$\tilde{F}(\boldsymbol{x}_k) \leq (1 - \frac{1}{2}\eta_1 \xi)^k \tilde{F}(\boldsymbol{x}_0),\ \forall\ k,$$

which finishes the proof. □

The same argument can be used to show a more general result than Theorem 4.2.2, as stated in the following corollary.

Corollary 4.2.2 *Under the same conditions of Theorem 4.2.1, any algorithm satisfying Properties (a) and (b) in Lemma 4.2.2 and the sufficient decrease condition* (4.24) *has the linear convergence property, i.e. generates a sequence* (X_k, Y_k) *that satisfies* (4.18).

4.3 Proof of Lemma 4.2.1

In Section 4.3.1, we will show that to prove Lemma 4.2.1, we only need to construct U, V to satisfy three inequalities that $\|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F$ and $\|((U-X)(V-Y)^T)\|_F$ are bounded above and $\langle \nabla_X G, X - U\rangle + \langle \nabla_Y G, Y - V\rangle$ is bounded below. In Section 4.3.2 we describe two propositions that specify the choice of U, V, and then we show that such U, V satisfy the three desired inequalities in Section 4.3.2 and subsequent subsections.

4.3.1 Preliminary analysis

Since $(X, Y) \in K(\delta)$, we have

$$d \triangleq \|M - XY^T\|_F \leq \delta \overset{(4.6)}{=} \frac{\Sigma_{\min}}{C_d r^{1.5} \kappa}. \tag{4.26}$$

To ensure (4.22) holds, we only need to ensure that the following two inequalities hold:

$$\phi_F = \langle \nabla_X F, X-U\rangle + \langle \nabla_Y F, Y-V\rangle \geq \frac{p}{4}d^2, \tag{4.27a}$$

$$\phi_G = \langle \nabla_X G, X-U\rangle + \langle \nabla_Y G, Y-V\rangle \geq 0. \tag{4.27b}$$

Define

$$a \triangleq U(Y-V)^T + (X-U)V^T, \quad b \triangleq (U-X)(V-Y)^T. \tag{4.28}$$

Then

$$XY^T - M = a+b, \;\; (X-U)Y^T + X(Y-V)^T = a+2b.$$

Using the expressions of $\nabla_X F, \nabla_Y F$ in (4.14), we bound ϕ_F as follows:

$$\begin{aligned}
\phi_F =& \langle \nabla_X F, X-U\rangle + \langle \nabla_Y F, Y-V\rangle \\
=& \langle \mathcal{P}_\Omega(XY^T-M), (X-U)Y^T + X(Y-V)^T\rangle \\
=& \langle \mathcal{P}_\Omega(a+b), \mathcal{P}_\Omega(a+2b)\rangle \\
=& \|\mathcal{P}_\Omega(a)\|_F^2 + 2\|\mathcal{P}_\Omega(b)\|_F^2 + 3\langle \mathcal{P}_\Omega(a), \mathcal{P}_\Omega(b)\rangle \\
\geq& \|\mathcal{P}_\Omega(a)\|_F^2 + 2\|\mathcal{P}_\Omega(b)\|_F^2 - 3\|\mathcal{P}_\Omega(a)\|_F\|\mathcal{P}_\Omega(b)\|_F.
\end{aligned} \tag{4.29}$$

The reason to decompose $M - XY^T$ as $a+b$ is the following. In order to bound $\|\mathcal{P}_\Omega(M-XY^T)\|_F$, we notice $E(\mathcal{P}_\Omega(M-XY^T)) = p(M-XY^T)$ and wish to prove $\|\mathcal{P}_\Omega(M-XY^T)\|_F^2 \approx O(pd^2)$. However, $\|\mathcal{P}_\Omega(A)\|_F$ could be as large as $\|A\|_F$ if the matrix A is not independent of the random subset Ω (e.g. choose A s.t. $A = \mathcal{P}_\Omega(A)$). This issue can be resolved by decomposing $XY^T - M$ as $a+b$ and bounding $\|\mathcal{P}_\Omega(a)\|_F$ and $\|\mathcal{P}_\Omega(b)\|_F$ separately. In fact, $\|\mathcal{P}_\Omega(a)\|_F$ can be bounded because a lies in a space spanned by the matrices with the same row space or column space as M, which is independent of Ω (Theorem 4.1 in [2]). $\|\mathcal{P}_\Omega(b)\|_F$ can be bounded according to a random graph lemma of [33, 74], which requires U, V, X, Y to be incoherent (i.e. have bounded row norm).

We claim that (4.27a) is implied by the following two inequalities:

$$\|\mathcal{P}_\Omega(b)\|_F = \|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F \leq \frac{1}{5}\sqrt{p}d; \tag{4.30a}$$

$$\|b\|_F = \|(U-X)(V-Y)^T\|_F \leq \frac{1}{10}d. \tag{4.30b}$$

In fact, assume (4.30a) and (4.30b) are true, we prove $\phi_F \geq pd^2/4$ as follows. By $XY^T - M = a + b$ we have

$$\|a\|_F \geq \|M - XY^T\|_F - \|b\|_F \overset{(4.30b)}{\geq} \frac{9}{10}d. \tag{4.31}$$

Recall that the SVD of M is $M = \hat{U}\Sigma\hat{V}^T$ and M satisfies the incoherence condtion (4.1). It follows from $M = UV^T = \hat{U}\Sigma\hat{V}^T$ that $M, U, \hat{U}$ have the same column space, thus there exists some matrix $B_1 \in \mathbb{R}^{r\times r}$ such that $U = \hat{U}B_1$; similarly, there exists $B_2 \in \mathbb{R}^{r\times r}$ such that $V = \hat{V}B_2$. Therefore, by the definition of a in (4.28) we have

$$a \in \mathcal{T} \triangleq \{\hat{U}W_2^T + W_1\hat{V}^T \mid W_1 \in \mathbb{R}^{m\times r}, W_2 \in \mathbb{R}^{n\times r}\}. \tag{4.32}$$

By Theorem 3.4 in [5] (or Theorem 4.1 in [2]), for $|\Omega|$ satisfying (4.17) with large enough C_0, we have that with probability at least $1 - 1/(2n^4)$, $\|\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}(a) - p\mathcal{P}_\mathcal{T}(a)\|_F \leq \frac{1}{6}p\|a\|_F$. Since $a \in \mathcal{T}$, this inequality can be simplified to

$$\|\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a) - pa\|_F \leq \frac{1}{6}p\|a\|_F. \tag{4.33}$$

Following the analysis of [2, Corollary 4.3], we have

$$\begin{aligned}\|\mathcal{P}_\Omega(a)\|_F^2 = \|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}(a)\|_F^2 = \langle a, \mathcal{P}_\mathcal{T}\mathcal{P}_\Omega^2\mathcal{P}_\mathcal{T}(a)\rangle = \langle a, \mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a)\rangle \\ = \langle a, pa\rangle + \langle a, \mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a) - pa\rangle.\end{aligned} \tag{4.34}$$

The absolute value of the second term can be bounded as

$$|\langle a, \mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a) - pa\rangle| \leq \|a\|_F\|\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a) - pa\|_F \overset{(4.33)}{\leq} \frac{1}{6}p\|a\|_F^2,$$

which implies $-\frac{1}{6}p\|a\|_F^2 \leq \langle a, \mathcal{P}_\mathcal{T}\mathcal{P}_\Omega(a) - pa\rangle \leq \frac{1}{6}p\|a\|_F^2$. Substituting into (4.34), we obtain that with probability at least $1 - 1/(2n^4)$,

$$\frac{5}{6}\|a\|_F^2 \leq \|\mathcal{P}_\Omega(a)\|_F^2 \leq \frac{7}{6}\|a\|_F^2. \tag{4.35}$$

The first inequality of the above relation implies

$$\|\mathcal{P}_\Omega(a)\|_F^2 \geq \frac{5}{6}\|a\|_F^2 \overset{(4.31)}{\geq} \frac{27}{40}pd^2. \tag{4.36}$$

According to (4.29) and the bounds (4.36) and (4.30a), we have $\phi_F/(pd^2) \geq \frac{27}{40} + 2(\frac{1}{5})^2 - \frac{3}{5}\sqrt{\frac{27}{40}} \geq \frac{1}{4}$, which proves (4.27a).

In summary, to find a factorization $M = UV^T$ such that (4.22) holds, we only need to ensure that the factorization satisfies (4.27b),(4.30a) and (4.30b). In the following three subsections, we will show that such a factorization $M = UV^T$ exists. Specifically, U, V will be defined in Table 4.7 and the three desired inequalities will be proved in Corollary 4.3.2, Proposition 4.3.3 and Claim 4.3.1 respectively.

4.3.2 Definitions of U, V and key technical results

We construct U, V according to two propositions, which will be stated in this subsection and proved in the appendix. The first proposition states that if XY^T is close to M, then there exists a factorization $M = UV^T$ such that U (resp. V) is close to X (resp. Y), and U, V are incoherent. Roughly speaking, this proposition shows the continuity of the factorization map $Z = XY^T \mapsto (X, Y)$ near a low-rank matrix M. The condition $X, Y \in K_1 \cap K_2 \cap K(\delta)$ and (4.6) implies that $d \triangleq \|M - XY^T\|_F \leq \delta = \frac{\Sigma_{\min}}{C_d r^{1.5}\kappa}$ and $\|X\|_F \leq \beta_T, \|Y\|_F \leq \beta_T$, thus for large enough C_d, the assumptions of Proposition 4.3.1 hold. Similarly, the assumptions of the other results in this subsection also hold.

Proposition 4.3.1 *Suppose M is μ-incoherent and let $\beta_T = \sqrt{C_T r \Sigma_{\max}}$. If*

$$d \triangleq \|M - XY^T\|_F \leq \frac{\Sigma_{\min}}{11r}, \tag{4.37a}$$

$$\|X\|_F \leq \beta_T, \quad \|Y\|_F \leq \beta_T, \tag{4.37b}$$

then there exists $U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$ such that

$$UV^T = M, \tag{4.38a}$$

$$\|U\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|X\|_F, \tag{4.38b}$$

$$\|U - X\|_F \leq \frac{6\beta_T}{5\Sigma_{\min}} d, \ \|V - Y\|_F \leq \frac{3\beta_T}{\Sigma_{\min}} d, \tag{4.38c}$$

$$\|U^{(i)}\|^2 \leq \frac{r\mu}{m}\beta_T^2, \ \|V^{(j)}\|^2 \leq \frac{3r\mu}{2n}\beta_T^2. \tag{4.38d}$$

The proof of Proposition 4.3.1 is given in Appendix A.3.

Remark 1: A symmetric result that switches X, U and Y, V in the above proposition holds: under the conditions of Proposition (4.3.1), there exist U, V satisfying (4.38) with U, V reversed, i.e. $UV^T = M$, $\|V\|_F(1 - \frac{d}{\Sigma_{\min}}) \leq \|Y\|_F$, $\|U - X\|_F \leq \frac{3\beta_T}{\Sigma_{\min}} d$, $\|V - Y\|_F \leq \frac{6\beta_T}{5\Sigma_{\min}} d$, and $\|U^{(i)}\|^2 \leq \frac{3r\mu}{2m}\beta_T^2$, $\|V^{(j)}\|^2 \leq \frac{r\mu}{n}\beta_T^2$.

Remark 2: To prove Theorem 4.2.1 (convergence), we only need $\|U\|_F \leq \|X\|_F$; here the slightly stronger requirement $\|U\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|X\|_F$ is for the purpose of proving Theorem 4.2.2 (linear convergence).

Remark 3: Without the incoherence assumption on M, by the same proof we can show that there still exist U, V satisfying (4.38a) and (4.38c), i.e. $M = UV^T$ and U, V are close to X, Y respectively. Such a result bears some similarity with the classical perturbation theory for singular value decomposition [75]. In particular, [75] proved that for two low-rank matrices[1] that are close, the spaces spanned by the left (resp. right) singular vectors of the two matrices are also close. Note that the singular vectors themselves may be very sensitive to perturbations and no such perturbation bounds can be established (see [100, Sec. 6]). The difference of our work with the classical perturbation theory is that we do not consider SVD of two matrices; instead, we allow one matrix to have an arbitrary factorization, and the factorization of the other matrix can be chosen accordingly. Since we do not have any restriction on the factorization XY^T (except the dimensions) and the norms of X and Y can be arbitrarily large, the distance between two corresponding factors has to be proportional to the norm of one single factor, which explains the coefficient β_T in (4.38c).

Unfortunately, Proposition 4.3.1 is not strong enough to prove $\phi_G \geq 0$ when both $\|X\|_F$ and $\|Y\|_F$ are large (see an analysis in Section 4.3.4). To resolve this issue, we need to prove the second proposition in which there is an additional assumption that both $\|X\|_F$ and $\|Y\|_F$ are large, and an additional requirement that both $\|U\|_F$ and $\|V\|_F$ are bounded (by the norms of original factors $\|X\|_F$ and $\|Y\|_F$ respectively). More specifically, the proposition states that if M is close to XY^T, and both $\|X\|_F$ and $\|Y\|_F$ are large, then there is a factorization $M = UV^T$ such that U (resp. V) is close to X (resp. Y), and $\|U\|_F \leq \|X\|_F, \|V\|_F \leq \|Y\|_F$. For the purpose of proving linear convergence, we prove a slightly stronger result that $\|V\|_F \leq (1 - d/\Sigma_{\min})\|Y\|_F$. The previous result Proposition 4.3.1 can be viewed as a perturbation analysis for an arbitrary factorization, while Proposition 4.3.2 can be viewed as an enhanced perturbation analysis for a constrained factorization. Although Proposition 4.3.2 is just a simple variant of Proposition 4.3.1, it seems to require a much more involved proof

[1] The result in [75] also covered the case of two approximately low-rank matrices, but we only consider the case of exact low-rank matrices here.

than Proposition 4.3.1. See the formal proof of Proposition 4.3.2 in Appendix A.4.

Proposition 4.3.2 *Suppose M is μ-incoherent and let $\beta_T = \sqrt{C_T r \Sigma_{\max}}$. There exists numerical constant C_d such that: if*

$$d \triangleq \|M - XY^T\|_F \leq \frac{\Sigma_{\min}}{C_d r}, \tag{4.39a}$$

$$\sqrt{\frac{2}{3}}\beta_T \leq \|X\|_F \leq \beta_T, \quad \sqrt{\frac{2}{3}}\beta_T \leq \|Y\|_F \leq \beta_T, \tag{4.39b}$$

then there exists $U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$ such that

$$UV^T = M, \tag{4.40a}$$

$$\|U\|_F \leq \|X\|_F, \ \|V\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F, \tag{4.40b}$$

$$\|U - X\|_F\|V - Y\|_F \leq 65\sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2, \quad \max\{\|U - X\|_F, \|V - Y\|_F\} \leq \frac{17}{2}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d, \tag{4.40c}$$

$$\|U^{(i)}\|^2 \leq \frac{r\mu}{m}\beta_T^2, \ \|V^{(j)}\|^2 \leq \frac{r\mu}{n}\beta_T^2. \tag{4.40d}$$

Remark: A symmetric result that switches X, U and Y, V in the above proposition still holds; the only change is that (4.40b) will become $\|U\|_F \leq (1-\frac{d}{\Sigma_{\min}})\|X\|_F, \ \|V\|_F \leq \|Y\|_F$. It is easy to prove a variant of the above proposition in which (4.40b) is changed to $\|U\|_F \leq (1 - \frac{d}{2\Sigma_{\min}})\|X\|_F, \|V\|_F \leq (1 - \frac{d}{2\Sigma_{\min}})\|Y\|_F$; in other words, the asymmetry of X, U and Y, V in (4.40b) is artificial. Nevertheless, Proposition 4.3.2 is enough for our purpose.

Throughout the proof of Lemma 4.2.1, U, V are defined in Table 4.3.2.

According to Proposition 4.3.1 and Proposition 4.3.2 (and their symmetric results), the properties of U, V defined in Tabel (4.7) are summarized in the following corollary. For simplicity, we only present the case that $\|Y\|_F \geq \|X\|_F$; in the other case that $\|Y\|_F < \|X\|_F$, a symmetric result of Corollary 4.3.1 holds.

Corollary 4.3.1 *Suppose $d \triangleq \|XY^T - M\|_F \leq \frac{\Sigma_{\min}}{C_d r}$ and $\|Y\|_F \geq \|X\|_F$, then U, V*

Table 4.7: Definition of U, V

Definition of U, V in different cases
Case 1: $\|X\|_F \le \|Y\|_F$.
Case 1.1 : $\|X\|_F < \sqrt{\frac{2}{3}}\beta_T$. Define U, V according to the symmetrical result of Proposition 4.3.1, i.e. U, V satisfy (4.38) with X, U and Y, V reversed.
Case 1.2: $\|X\|_F, \|Y\|_F \in [\sqrt{\frac{2}{3}}\beta_T, \beta_T]$. Define U, V according to Proposition 4.3.2.
Case 2: $\|Y\|_F < \|X\|_F$.
Similar to Case 1 but with the roles of X, U and Y, V reversed.

defined in Table 4.7 satisfy:

$$UV^T = M; \tag{4.41a}$$

$$\|U - X\|_F\|V - Y\|_F \le 65\sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2; \quad \max\{\|U - X\|_F, \|V - Y\|_F\} \le \frac{17}{2}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d, \tag{4.41b}$$

$$\|U^{(i)}\|^2 \le \frac{3}{2}\frac{r\mu}{m}\beta_T^2, \ \|V^{(j)}\|^2 \le \frac{3}{2}\frac{r\mu}{n}\beta_T^2; \tag{4.41c}$$

$$\|V\|_F \le (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F; \ \ \textit{if } \|X\|_F > \sqrt{\frac{2}{3}}\beta_T, \ \textit{then } \|U\|_F \le \|X\|_F. \tag{4.41d}$$

In (4.41b), we bound $\|U - X\|_F\|V - Y\|_F$ by $O(d^2)$ with a rather complicated coefficient, but to prove (4.30b) we need a bound $O(d)$ with a coefficient $1/10$. Under a slightly stronger condition on d than that of Corollary 4.3.1, which still holds for $(X, Y) \in K(\delta)$ with δ defined in (4.6), we can prove the bound (4.30b) by (4.41b).

Corollary 4.3.2 *There exists a numerical constant C_d such that if*

$$d \triangleq \|M - XY^T\|_F \le \frac{\Sigma_{\min}}{C_d r^{1.5}\kappa}, \tag{4.42}$$

then U, V defined in Table 4.7 satisfy (4.30b).

Proof of Corollary 4.3.2: According to (4.41b) , we have

$$\|U - X\|_F\|V - Y\|_F \le 65\frac{\beta_T^2}{\Sigma_{\min}^2}\sqrt{r}d^2 = 65C_T r^{1.5}\frac{\Sigma_{\max}}{\Sigma_{\min}^2}d^2 = 65C_T r^{1.5}\kappa\frac{d}{\Sigma_{\min}}d \le \frac{1}{10}d,$$

where the last inequliaty follows from (4.42) with $C_d \ge 650C_T$. □

In the next two subsections, we will use the properties in Corollary 4.3.1 to prove (4.30a) and (4.27b).

4.3.3 Upper bound on $\|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F$

The following result states that for U, V defined in Table 4.7, (4.30a) holds.

Proposition 4.3.3 *Under the same conditions as Lemma 4.2.1, with probability at least $1-1/(2n^4)$, the following is true. For any $(X,Y) \in K_1 \cap K_2 \cap K(\delta)$ and U, V defined in Table 4.7, we have*

$$\|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F^2 \leq \frac{p}{25}\|M-XY^T\|_F^2. \tag{4.43}$$

Proof of Proposition 4.3.3: We need the following random graph lemma [33, Lemma 7.1].

Lemma 4.3.1 *There exist numerical constants C_0, C_1 such that if $|\Omega| \geq C_0\sqrt{\alpha}n\log n$, then with probability at least $1-1/(2n^4)$, for all $x \in \mathbb{R}^m, y \in \mathbb{R}^n$,*

$$\sum_{(i,j)\in\Omega} x_i y_j \leq C_1 p\|x\|_1\|y\|_1 + C_1\alpha^{\frac{3}{4}}\sqrt{np}\|x\|_2\|y\|_2. \tag{4.44}$$

Let $Z = U-X, W = V-Y$ and $z_i = \|Z^{(i)}\|^2$, $w_j = \|W^{(j)}\|^2$. We have

$$\begin{aligned}\|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F^2 &= \sum_{(i,j)\in\Omega}(ZW^T)_{ij}^2 \\ &\leq \sum_{(i,j)\in\Omega}\|Z^{(i)}\|^2\|W^{(j)}\|^2 = \sum_{(i,j)\in\Omega} z_i w_j.\end{aligned} \tag{4.45}$$

Invoking Lemma 4.3.1, we have

$$\|\mathcal{P}_\Omega((U-X)(V-Y)^T)\|_F^2 \leq C_1 p\|z\|_1\|w\|_1 + C_1\alpha^{\frac{3}{4}}\sqrt{np}\|z\|_2\|w\|_2. \tag{4.46}$$

Analogous to the proof of (4.30b) in Corollary 4.3.2, we can prove that $\|U-X\|_F\|V-Y\|_F \leq d/(10\sqrt{C_1})$ for large enough C_d (in fact, $C_d \geq 650C_T\sqrt{C_1}$ suffices). Therefore, we have

$$\|z\|_1\|w\|_1 = \|Z\|_F^2\|W\|_F^2 = \|U-X\|_F^2\|V-Y\|_F^2 \leq \frac{1}{100C_1}d^2. \tag{4.47}$$

We still need to bound $\|z\|_2$ and $\|w\|_2$. We have

$$\begin{aligned}\|z\|_2 = \sqrt{\sum_i \|Z^{(i)}\|^4} &\le \sqrt{\max_i \|Z^{(i)}\|^2 \sum_j \|Z^{(j)}\|^2} \\ &\le \max_i (\|U^{(i)}\| + \|X^{(i)}\|)\|U - X\|_F \\ &\le (\sqrt{\frac{3r\mu}{2m}}\beta_T + \beta_1)\|U - X\|_F \\ &\le \sqrt{8}\sqrt{\frac{r\mu}{m}}\beta_T\|U - X\|_F.\end{aligned} \tag{4.48}$$

Here, the third inequliaty follows from the property (4.41c) in Corollary 4.3.1 and the condition $(X, Y) \in K_1$ (which implies $\|X^{(i)}\| \le \beta_1$), and the fourth inequliaty follows from the definition of β_1 in (4.5). Similarly,

$$\begin{aligned}\|w\|_2 \le& \max_j (\|V^{(j)}\| + \|Y^{(j)}\|)\|V - Y\|_F \\ \le& \sqrt{8}\sqrt{\frac{r\mu}{n}}\beta_T\|V - Y\|_F.\end{aligned} \tag{4.49}$$

Multiplying (4.48) and (4.49), we get

$$\|z\|_2\|w\|_2 \le 8\frac{r\mu}{\sqrt{mn}}\beta_T^2\|U - X\|_F\|V - Y\|_F \overset{(4.41b)}{\le}$$
$$8\frac{r\mu}{\sqrt{mn}}\beta_T^2 65\sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2 \overset{(4.5)}{=} 520C_T^2\frac{1}{\sqrt{mn}}\mu r^{3.5}\kappa^2 d^2.$$

Thus the second term in (4.46) can be bounded as

$$C_1\alpha^{\frac{3}{4}}\sqrt{np}\|z\|_2\|w\|_2 \le 520C_1C_T^2\frac{\alpha^{\frac{3}{4}}\sqrt{np}}{\sqrt{mn}}\mu r^{3.5}\kappa^2 d^2 \le \frac{3}{100}pd^2, \tag{4.50}$$

where the last inequality is equivalent to $520^2C_1^2C_T^4\alpha^{\frac{3}{2}}\mu^2 r^7\kappa^4 \le \frac{9}{100^2}|\Omega|/n$, which holds due to (4.17) with large enough numerical constant C_0. Plugging (4.47) and (4.50) into (4.46), we get $\|\mathcal{P}_\Omega((U - X)(V - Y)^T)\|_F^2 \le \frac{p}{25}d^2 = \frac{p}{25}\|M - XY^T\|_F^2$. □

4.3.4 Lower bound on ϕ_G

In this subsection, we prove the following claim.

Claim 4.3.1 *U, V defined in Table 4.7 satisfy (4.27b), i.e. $\phi_G = \langle \nabla_X G, X - U\rangle + \langle \nabla_Y G, Y - V\rangle \ge 0$.*

Proof of Claim 4.3.1:

By the expressions of $\nabla_X G, \nabla_Y G$ in (4.14), we have

$$\begin{aligned}
&\phi_G = \langle \nabla_X G, X - U\rangle + \langle \nabla_Y G, Y - V\rangle = \\
&\rho \sum_{i=1}^{m} G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})\frac{3}{\beta_1^2}\langle X^{(i)}, X^{(i)} - U^{(i)}\rangle + \rho G_0'(\frac{3\|X\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle X, X - U\rangle \\
&+\rho \sum_{j=1}^{n} G_0'(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})\frac{3}{\beta_2^2}\langle Y^{(j)}, Y^{(j)} - V^{(j)}\rangle + \rho G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle Y, Y - V\rangle,
\end{aligned} \tag{4.51}$$

where $G_0'(z) = I_{[1,\infty]}(z)2(z-1)$.

Firstly, we prove

$$h_{1i} \triangleq G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})\frac{3}{\beta_1^2}\langle X^{(i)}, X^{(i)} - U^{(i)}\rangle \geq 0, \ \forall\ i, \tag{4.52a}$$

$$h_{3j} \triangleq G_0'(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})\frac{3}{\beta_2^2}\langle Y^{(j)}, Y^{(j)} - V^{(j)}\rangle \geq 0, \ \forall\ j. \tag{4.52b}$$

We only need to prove (4.52a); the proof of (4.52b) is similar. We consider two cases.

Case 1: $\|X^{(i)}\|^2 \leq \frac{2\beta_1^2}{3}$. Note that $\frac{3\|X^{(i)}\|^2}{2\beta_1^2} \leq 1$ implies $G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) = 0$, thus $h_{1i} = 0$.

Case 2: $\|X^{(i)}\|^2 > \frac{2\beta_1^2}{3}$. By Corollary 4.3.1 and the fact that $\beta_1^2 = \beta_T^2\frac{3\mu r}{m}$, we have

$$\|U^{(i)}\|^2 \leq \frac{3r\mu}{2m}\beta_T^2 \leq \frac{2\beta_1^2}{3} < \|X^{(i)}\|^2. \tag{4.53}$$

As a result, $\langle X^{(i)}, X^{(i)}\rangle = \|X^{(i)}\|\|X^{(i)}\| > \|X^{(i)}\|\|U^{(i)}\| \geq \langle X^{(i)}, U^{(i)}\rangle$, which implies $\langle X^{(i)}, X^{(i)} - U^{(i)}\rangle \geq 0$. Combining this inequality with the fact that $G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) \geq 0$, we get $h_{1i} \geq 0$.

Secondly, we prove

$$\begin{aligned}
&h_2 + h_4 \geq 0, \\
\text{where} \quad &h_2 \triangleq G_0'(\frac{3\|X\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle X, X - U\rangle, \quad h_4 \triangleq G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle Y, Y - V\rangle.
\end{aligned} \tag{4.54}$$

Without loss of generality, we can assume $\|Y\|_F \geq \|X\|_F$, and we will apply Corollary 4.3.1 to prove (4.54). If $\|Y\|_F < \|X\|_F$, we can apply a symmetric result of Corollary 4.3.1 to prove (4.54). We further consider three cases.

Case 1: $\|X\|_F \leq \|Y\|_F \leq \sqrt{\frac{2}{3}}\beta_T$. In this case $G_0'(\frac{3\|X\|_F^2}{2\beta_T^2}) = G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2}) = 0$, which implies $h_2 = h_4 = 0$, thus (A.190) holds.

Case 2: $\|X\|_F \leq \sqrt{\frac{2}{3}}\beta_T < \|Y\|_F$. Then $G_0'(\frac{3\|X\|_F^2}{2\beta_T^2}) = 0$, which implies $h_2 = 0$. By (4.41d) in Corollary 4.3.1 we have $\|V\|_F \leq \|Y\|_F$, which implies $\langle Y, Y\rangle \geq \|Y\|_F\|V\|_F \geq \langle Y, V\rangle$, i.e. $\langle Y, Y - V\rangle \geq 0$. Combined with the nonnegativity of $G_0'(\cdot)$, we get $h_4 \geq 0$. Thus $h_2 + h_4 = h_4 \geq 0$.

Case 3: $\sqrt{\frac{2}{3}}\beta_T < \|X\|_F \leq \|Y\|_F$. By (4.41d) in Corollary 4.3.1, we have $\|U\|_F \leq \|X\|_F$ and $\|V\|_F \leq \|Y\|_F$. Similar to the argument in Case 1 we can prove $h_2 \geq 0, h_4 \geq 0$ and (4.54) follows.

In both cases, we have proved (4.54), thus (4.54) holds.

We conclude that for U, V defined in Table 4.7,

$$\phi_G \overset{(4.51)}{=} \rho\left(\sum_i h_{1i} + \sum_j h_{3j} + h_2 + h_4\right) \overset{(4.52),(4.54)}{\geq} 0,$$

which finishes the proof of Claim 4.3.1. □

Remark: Based on the above proof, we can explain why Proposition 4.3.1 is not enough to prove $\phi_G \geq 0$. Note that $h_2 = 0$ when $\|X\|_F > \sqrt{\frac{2}{3}}\beta_T$ and $h_4 = 0$ when $\|Y\|_F > \sqrt{\frac{2}{3}}\beta_T$. To prove $h_2 \geq 0, h_4 \geq 0$, it suffices to prove: (i) $\|U\|_F \leq \|X\|_F$ when $\|X\|_F > \sqrt{\frac{2}{3}}\beta_T$; (ii) $\|V\|_F \leq \|Y\|_F$ when $\|Y\|_F > \sqrt{\frac{2}{3}}\beta_T$. For the choice of U, V in Proposition 4.3.1, we have $\|U\|_F \leq \|X\|_F$, but there is no guarantee that (ii) holds. Similarly, for the choice of U, V in the symmetric result of Proposition 4.3.1, we have $\|V\|_F \leq \|Y\|_F$, but there is no guarantee that (i) holds. Thus, Proposition 4.3.1 is not enough to prove $\phi_G \geq 0$. To guarantee that (i) and (ii) hold simultaneously, we need a complementary result for the case $\|X\|_F > \sqrt{\frac{2}{3}}\beta_T, \|Y\|_F > \sqrt{\frac{2}{3}}\beta_T$. This motivates our Proposition 4.3.2.

4.4 Proof of Lemma 4.2.2

Property (a) in Lemma 4.2.2 (convergence to stationary points) is a basic requirement for many reasonable algorithms and can be proved using classical results in optimization, so the difficulty mainly lies in how to prove Property (b). We will give some easily verifiable conditions for Property (b) to hold and then show that Algorithms 1-4 satisfy these conditions. This proof framework can be used to extend Theorem 4.2.1 to many other algorithms.

The following claim states that Algorithms 1-4 satisfy Property (a). The proof of this claim is given in Appendix A.5.5.

Claim 4.4.1 *Suppose Ω satisfies* (4.19)*, then each limit point of the sequence generated by Algorithms 1-4 is a stationary point of problem (P1).*

For Property (b), we first show that the initial point (X_0, Y_0) lies in an incoherent neighborhood $(\sqrt{\frac{2}{3}}K_1)\cap(\sqrt{\frac{2}{3}}K_2)\cap K_{\delta_0}$, where cK_i denotes the set $\{(cX, cY) \mid (X, Y) \in K_i\}, i = 1, 2$. The proof of Claim 4.4.2 will be given in Appendix A.5.1. The purpose of proving $(X_0, Y_0) \in (\sqrt{\frac{2}{3}}K_1) \cap (\sqrt{\frac{2}{3}}K_2)$ rather than $(X_0, Y_0) \in K_1 \cap K_2$ is to guarantee that $G(X_0, Y_0) = 0$, where G is the regularized function defined in (4.3).

Claim 4.4.2 *Under the same condition of Lemma 4.2.1, with probability at least* $1 - 1/(2n^4)$*,* (X_0, Y_0) *given by the procedure* INITIALIZE *belongs to* $(\sqrt{\frac{2}{3}}K_1)\cap(\sqrt{\frac{2}{3}}K_2)\cap K_{\delta_0}$*, where* δ_0 *is defined by* (4.6)*, i.e.*
(a) $\|X_0^{(i)}\| \le \sqrt{\frac{2}{3}}\beta_1, i = 1, 2, \ldots, m;\ \ \|Y_0^{(j)}\| \le \sqrt{\frac{2}{3}}\beta_2, j = 1, \ldots, n;$
(b) $\|X_0\|_F \le \sqrt{\frac{2}{3}}\beta_T,\ \|Y_0\|_F \le \sqrt{\frac{2}{3}}\beta_T;$
(c) $\|M - X_0Y_0^T\|_F \le \delta_0.$

The next result provides some general conditions for (X_t, Y_t) to lie in $K_1\cap K_2\cap K(\delta)$. To simplify the notations, denote $\boldsymbol{x}_t \triangleq (X_t, Y_t)$ and

$$\boldsymbol{u}^* \triangleq (\hat{U}\Sigma^{1/2}, \hat{V}\Sigma^{1/2}),$$

where $\hat{U}\Sigma\hat{V}$ is the SVD of M. Recall that $\tilde{F}(\boldsymbol{u}^*) = 0$ (proved in the paragraph after (4.9)). We say a function $\psi(\bar{\boldsymbol{x}}, \boldsymbol{\Delta}; \lambda)$ is a convex tight upper bound of $\tilde{F}(\boldsymbol{x})$ along the direction $\boldsymbol{\Delta}$ at $\bar{\boldsymbol{x}}$ if

$$\psi(\bar{\boldsymbol{x}}, \boldsymbol{\Delta}; \lambda) \text{ is convex over } \lambda \in \mathbb{R}; \tag{4.55a}$$

$$\psi(\bar{\boldsymbol{x}}, \boldsymbol{\Delta}; \lambda) \ge \tilde{F}(\bar{\boldsymbol{x}} + \lambda\boldsymbol{\Delta}),\ \forall\ \lambda \in \mathbb{R};\quad \psi(\bar{\boldsymbol{x}}, \boldsymbol{\Delta}; 0) = \tilde{F}(\bar{\boldsymbol{x}}). \tag{4.55b}$$

For example, $\psi(\bar{\boldsymbol{x}}, \boldsymbol{\Delta}; \lambda) = \tilde{F}(\bar{\boldsymbol{x}} + \lambda\boldsymbol{\Delta})$ satisfies (4.55) for either $\boldsymbol{\Delta} = (X, 0)$ or $\boldsymbol{\Delta} = (0, Y)$, where $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ are arbitrary matrices. This definition is motivated by the block successive upper bound minimization method [82]. The proof of Proposition 4.4.1 is given in Appendix A.5.3.

Proposition 4.4.1 *Suppose the sample set Ω satisfies* (4.19) *and δ, δ_0 are defined by* (4.6). *Consider an algorithm that starts from a point $\boldsymbol{x}_0 = (X_0, Y_0)$ and generates a sequence $\{\boldsymbol{x}_t\} = \{(X_t, Y_t)\}$. Suppose $\boldsymbol{x}_0$ satisfies*

$$\boldsymbol{x}_0 \in (\sqrt{\frac{2}{3}}K_1) \cap (\sqrt{\frac{2}{3}}K_2) \cap K(\delta_0), \tag{4.56}$$

and $\{\boldsymbol{x}_t\}$ satisfies either of the following three conditions:

$$1) \qquad \tilde{F}(\boldsymbol{x}_t + \lambda \boldsymbol{\Delta}_t) \leq 2\tilde{F}(\boldsymbol{x}_0), \forall\ \lambda \in [0,1],\ \textit{where } \boldsymbol{\Delta}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t,\ \ \forall\ t; \tag{4.57a}$$

$$2) \qquad 1 = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}} \psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; \lambda),\ \textit{where } \psi \textit{ satisfies } (4.55), \boldsymbol{\Delta}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t,\ \ \forall\ t; \tag{4.57b}$$

$$3) \qquad \tilde{F}(\boldsymbol{x}_t) \leq 2\tilde{F}(\boldsymbol{x}_0), \quad d(\boldsymbol{x}_t, \boldsymbol{x}_0) \leq \frac{5}{6}\delta,\ \forall\ t. \tag{4.57c}$$

Then $\boldsymbol{x}_t = (X_t, Y_t) \in K_1 \cap K_2 \cap K(2\delta/3)$, for all $t \geq 0$.

The first condition means that $\tilde{F}$ is bounded above by $2\tilde{F}(\boldsymbol{x}_0)$ over the line segment between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ for any t. This condition holds for gradient descent or SGD with small enough stepsize (see Claim 4.4.3). The second condition means that the new point $\boldsymbol{x}_{t+1}$ is the minimum of a convex tight upper bound of the original function along the direction $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t$, and holds for BCD type methods such as Algorithm 2 and Algorithm 3 (see Claim 4.4.3). Note that the gradient descent method with exact line search stepsize does not satisfy this condition since $\tilde{F}$ is not jointly convex in the variable (X, Y). The third condition means that $\tilde{F}(\boldsymbol{x}_t)$ is bounded above and $\boldsymbol{x}_t$ is not far from $\boldsymbol{x}_0$ for any t. For standard nonlinear optimization algorithms, it is not easy to prove that $\boldsymbol{x}_t$ is not far from $\boldsymbol{x}_0$. However, as done by Algorithm 1 with restricted Armijo rule or restricted line search, we can force $d(\boldsymbol{x}_t, \boldsymbol{x}_0) \leq \frac{5}{6}\delta$ to hold when computing the new point $\boldsymbol{x}_t$.

The following claim shows that each of Algorithm 1-4 satisfies one of the three conditions in (4.57). The proof of Claim 4.4.3 is given in Appendix A.5.4.

Claim 4.4.3 *The sequence $\{\boldsymbol{x}_t\}$ generated by Algorithm 1 with either restricted Armijo rule or restricted line search satisfies* (4.57c). *The sequence $\{\boldsymbol{x}_t\}$ generated by either Algorithm 2 or Algorithm 3 satisfies* (4.57b). *Suppose the sample set Ω satisfies* (4.19), *then the sequence $\{\boldsymbol{x}_t\}$ generated by either Algorithm 1 with constant stepsize or Algorithm 4 satisfies* (4.57a).

To put things together, Claim 4.4.1 shows Algorithms 1-4 satisfy Property (a), and Proposition 4.4.1 together with Claim 4.4.2 and Claim 4.4.3 shows that Algorithms 1-4 satisfy Property (b). Therefore, we have proved Lemma 4.2.2.

References

[1] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[2] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[3] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[4] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[5] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[6] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.

[7] Amit Singer. A remark on global positioning from local distances. *Proceedings of the National Academy of Sciences*, 105(28):9507–9511, 2008.

[8] Sewoong Oh, Andrea Montanari, and Amin Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *Information Theory Workshop (ITW), 2010 IEEE*, pages 1–5. IEEE, 2010.

[9] Adel Javanmard and Andrea Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

[10] Emmanuel J Candès, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[11] Zhang Liu, Anders Hansson, and Lieven Vandenberghe. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

[12] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[13] Pei Chen and David Suter. Recovering the missing components in a large noisy low-rank matrix: Application to sfm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1051–1063, 2004.

[14] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24. ACM, 2007.

[15] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[16] Shunqiao Sun, Athina P Petropulu, and Waheed U Bajwa. Target estimation in colocated mimo radar via matrix completion. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 4144–4148. IEEE, 2013.

[17] Homa Esfahanizadeh, Farshad Lahouti, and Babak Hassibi. A matrix completion approach to linear index coding problem. *arXiv preprint arXiv:1408.3046*, 2014.

[18] A. Zare, M.R. Jovanovic, and T.T. Georgiou. Completion of partially known turbulent flow statistics. In *American Control Conference (ACC), 2014*, pages 1674–1679, June 2014.

[19] Yongxin Chen, Mihailo R Jovanovic, and Tryphon T Georgiou. State covariances and the matrix completion problem. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 1702–1707. IEEE, 2013.

[20] Fu Lin, Mihailo R Jovanovic, and Tryphon T Georgiou. An ADMM algorithm for matrix completion of partially known state covariances. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 1684–1689. IEEE, 2013.

[21] Tianxi Cai, T Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, (just-accepted), 2015.

[22] Yuxin Chen and Yuejie Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576–6601, Oct 2014.

[23] Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[24] Hulikal Keshavan et al. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

[25] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[26] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[27] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[28] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[29] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[30] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[31] Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.

[32] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[33] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[34] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *arXiv preprint arXiv:1411.8003*, 2014.

[35] Ruoyu Sun and Zhi-Quan Luo. Interference alignment using finite and dependent channel extensions: The single beam case. *IEEE Transactions on Information Theory*, 61(1):239–255, Jan 2015.

[36] Franz Király and Ryota Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. *arXiv preprint arXiv:1206.6470*, 2012.

[37] Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

[38] Franz J Király, Louis Theran, and Ryota Tomioka. The algebraic combinatorial approach for low-rank matrix completion. *arXiv preprint arXiv:1211.4116*, 2012.

[39] Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. *arXiv preprint arXiv:1402.2331*, 2014.

[40] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Major components of the gravity recommendation system. *ACM SIGKDD Explorations Newsletter*, 9(2):80–83, 2007.

[41] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[42] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[43] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[44] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[45] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[46] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[47] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[48] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[49] Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[50] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[51] Alekh Agarwal, Sahand Negahban, Martin J Wainwright, et al. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

[52] Ke Hou, Zirui Zhou, Anthony Man-Cho So, and Zhi-Quan Luo. On the linear convergence of the proximal gradient method for trace norm regularization. In *Advances in Neural Information Processing Systems*, pages 710–718, 2013.

[53] Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2007.

[54] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing (STOC)*, pages 665–674. ACM, 2013.

[55] Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 651–660. IEEE, 2014.

[56] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678, 2014.

[57] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management*, pages 337–348. Springer, 2008.

[58] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[59] Simon Funk. Netflix update: Try this at home. *http://sifter.org/ simon/journal/20061211.html*.

[60] Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, volume 2007, pages 5–8, 2007.

[61] Rainer Gemulla, Erik Nijkamp, Peter J Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77. ACM, 2011.

[62] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[63] Yong Zhuang, Wei-Sheng Chin, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel SGD for matrix factorization in shared memory systems. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 249–256. ACM, 2013.

[64] István Pilászy, Dávid Zibriczky, and Domonkos Tikk. Fast als-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 71–78. ACM, 2010.

[65] Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, and Inderjit S Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *ICDM*, pages 765–774, 2012.

[66] Emmanuel Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[67] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *arXiv preprint, http://arxiv.org/abs/0909.3304v1*, 2009.

[68] Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*, 2014.

[69] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Global convergence of stochastic gradient descent for some nonconvex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

[70] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[71] Ruoyu Sun, Zhi-Quan Luo, and Yinyu Ye. On the expected convergence of randomly permuted ADMM. *arXiv preprint arXiv:1503.06387*, 2015.

[72] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[73] Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *arXiv preprint arXiv:1410.2596*, 2014.

[74] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[75] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[76] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

[77] Ruoyu Sun and Zhi-Quan Luo. Globally optimal joint uplink base station association and power control for max-min fairness. In *2014 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 454–458. IEEE, 2014.

[78] Ruoyu Sun, Mingyi Hong, and Zhi-Quan Luo. Joint downlink base station association and power control for max-min fairness: Computation and complexity. *IEEE Journal on Selected Areas in Communications (JSAC)*, 33(6):1040–1054, June 2015.

[79] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[80] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[81] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[82] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[83] Hadi Baligh, Mingyi Hong, W Liao, Z Luo, Meisam Razaviyayn, Maziar Sanjabi, and Ruoyu Sun. Cross-layer provision of future cellular networks: A WMMSE-based approach. *Signal Processing Magazine, IEEE*, 31(6):56–68, 2014.

[84] Ruoyu Sun, Hadi Baligh, and Zhi-Quan Luo. Long-term transmit point association for coordinated multipoint transmission by stochastic optimization. In *Signal Processing Advances in Wireless Communications (SPAWC), 2013 IEEE 14th Workshop on*, pages 330–334. IEEE, 2013.

[85] Mingyi Hong, Ruoyu Sun, H. Baligh, and Zhi-Quan Luo. Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks. *IEEE Journal on Selected Areas in Communications (JSAC)*, 31(2):226–240, February 2013.

[86] Dimitri P Bertsekas. Nonlinear programming. 1999.

[87] Roger Fletcher. On the barzilai-borwein method. In *Optimization and control with applications*, pages 235–256. Springer, 2005.

[88] Yu-Hong Dai and Li-Zhi Liao. R-linear convergence of the barzilai and borwein gradient method. *IMA Journal of Numerical Analysis*, 22(1):1–10, 2002.

[89] Zhi-Quan Luo. On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks. *Neural Computation*, 3(2):226–245, June 1991.

[90] Zhi-Quan Luo and Paul Tseng. Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. *Optimization Methods and Software*, 4(2):85–101, 1994.

[91] Benjamin Recht and Christopher Ré. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. *arXiv preprint arXiv:1202.4184*, 2012.

[92] Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Ruslan R Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[93] Dimitris Achlioptas and Frank McSherry. Fast computation of low rank matrix approximations. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC)*, pages 611–618. ACM, 2001.

[94] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pages 674–682, 2014.

[95] Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. *arXiv preprint arXiv:1402.2324*, 2014.

[96] Willard I Zangwill. Non-linear programming via penalty functions. *Management science*, 13(5):344–358, 1967.

[97] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[98] Luigi Grippo and Marco Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000.

[99] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[100] Gilbert W Stewart. Perturbation theory for the singular value decomposition. 1998.

Appendix A

Proofs

A.1 Proof of Claim 4.1.1

This proof is quite straightforward and we mainly use the triangular inequalities and the boundedness of the considered region $\Gamma(\beta_0)$. In this proof, $f'(x)$ denotes the derivative of a function f at x.

Since $(X, Y), (U, V)$ belong to $\Gamma(\beta_0)$, we have

$$\|X\|_F \leq \beta_0, \|Y\|_F \leq \beta_0, \|U\|_F \leq \beta_0, \|V\|_F \leq \beta_0. \tag{A.1}$$

We first prove

$$\|\nabla F(X, Y) - \nabla F(U, V)\|_F \leq 4\beta_0^2 \|(X, Y) - (U, V)\|_F. \tag{A.2}$$

By the triangular inequality, we have

$$\begin{aligned} \|\nabla_X F(X, Y) - \nabla_X F(U, V)\|_F \leq \|\nabla_X F(X, Y) - \nabla_X F(U, Y)\|_F \\ + \|\nabla_X F(U, Y) - \nabla_X F(U, V)\|_F. \end{aligned} \tag{A.3}$$

The first term of (A.3) can be bounded as follows

$$\begin{aligned}\|\nabla_X F(X,Y) - \nabla_X F(U,Y)\|_F &= \|\mathcal{P}_\Omega(XY^T - M)Y - \mathcal{P}_\Omega(UY^T - M)Y\|_F \\ &\le \|\mathcal{P}_\Omega(XY^T - M) - \mathcal{P}_\Omega(UY^T - M)\|_F\|Y\|_F \\ &= \|\mathcal{P}_\Omega[(X-U)Y^T]\|_F\|Y\|_F \\ &\le \|(X-U)Y^T\|_F\|Y\|_F \\ &\le \|X-U\|_F\|Y\|_F^2 \\ &\le \|X-U\|_F\beta_0^2.\end{aligned}$$

The second term of (A.3) can be bounded as

$$\begin{aligned}&\|\nabla_X F(U,Y) - \nabla_X F(U,V)\|_F \\ &= \|\mathcal{P}_\Omega(UY^T - M)Y - \mathcal{P}_\Omega(UV^T - M)V\|_F \\ &\le \|\mathcal{P}_\Omega(M)(V-Y)\|_F + \|\mathcal{P}_\Omega(UY^T)Y - \mathcal{P}_\Omega(UV^T)V\|_F \\ &\le \|\mathcal{P}_\Omega(M)(V-Y)\|_F + \|\mathcal{P}_\Omega(UY^T)Y - \mathcal{P}_\Omega(UY^T)V\|_F + \|\mathcal{P}_\Omega(UY^T)V - \mathcal{P}_\Omega(UV^T)V\|_F \\ &\le \|\mathcal{P}_\Omega(M)\|_F\|V-Y\|_F + \|\mathcal{P}_\Omega(UY^T)\|_F\|Y-V\|_F + \|\mathcal{P}_\Omega[U(Y-V)^T]\|_F\|V\|_F \\ &\le \|M\|_F\|V-Y\|_F + \|U\|_F\|Y\|_F\|Y-V\|_F + \|U\|_F\|Y-V\|_F\|V\|_F \\ &\le 3\beta_0^2\|Y-V\|_F,\end{aligned}$$

where the last inequliaty follows from (A.1) and the fact that $\|M\|_F \le \sqrt{r}\Sigma_{\max} \overset{(4.5)}{=} \frac{1}{C_T\sqrt{r}}\beta_T^2 \le \beta_T^2 \le \beta_0^2$ (here the second last inequality follows from the fact that the numerical constant $C_T \ge 1$, and the last inequality follows from the assumption of Claim 4.1.1).

Plugging the above two bounds into (A.3), we obtain

$$\|\nabla_X F(X,Y) - \nabla_X F(U,V)\|_F \le \beta_0^2(\|X-U\|_F + 3\|Y-V\|_F).$$

Similarly, we have

$$\|\nabla_Y F(X,Y) - \nabla_Y F(U,V)\|_F \le \beta_0^2(3\|X-U\|_F + \|Y-V\|_F).$$

Combining the above two relations, we have (denote $\omega_1 \triangleq \|X-U\|_F, \omega_2 \triangleq \|Y-V\|_F$)

$$\begin{aligned}
&\|\nabla F(X,Y) - \nabla F(U,V)\|_F \\
&= \sqrt{\|\nabla_X F(X,Y) - \nabla_X F(U,V)\|_F^2 + \|\nabla_Y F(X,Y) - \nabla_Y F(U,V)\|_F^2} \\
&\le \beta_0^2 \sqrt{(\omega_1 + 3\omega_2)^2 + (3\omega_1 + \omega_2)^2} \\
&\le 4\beta_0^2 \sqrt{\omega_1^2 + \omega_2^2} \\
&= 4\beta_0^2 \|(X,Y) - (U,V)\|_F,
\end{aligned}$$

which proves (A.2).

Next we prove

$$\|\nabla G(X,Y) - \nabla G(U,V)\|_F \le 54\rho \frac{\beta_0^2}{\beta_1^4} \|(X,Y) - (U,V)\|_F. \tag{A.4}$$

Denote

$$G_{1i}(X) \triangleq G_0\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right), \quad G_2(X) \triangleq G_0\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right), \tag{A.5}$$

then we have

$$\nabla G_{1i}(X) = G_0'\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right)\frac{3\bar{X}^{(i)}}{\beta_1^2}, \quad \nabla G_2(X) = G_0'\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right)\frac{3X}{\beta_T^2}, \tag{A.6}$$

where $G_0'(z) = I_{[1,\infty]}(z)2(z-1)$ and $\bar{X}^{(i)}$ denotes a matrix with the i-th row being $X^{(i)}$ and the other rows being zero. Obviously $G_{1i}(X)$ is a matrix with all but the i-th row being zero. Recall that

$$G(X,Y) = \rho \sum_i G_{1i}(X) + \rho G_2(X) + f_0(Y),$$

where $f_0(Y)$ is a certain function of Y which we can ignore for now. Then we have

$$\begin{aligned}
\nabla_X G(X,Y) &= \rho \sum_i \nabla G_{1i}(X) + \rho \nabla G_2(X) \\
&= \rho \sum_{i=1}^m G_0'\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right)\frac{3\bar{X}^{(i)}}{\beta_1^2} + \rho G_0'\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right)\frac{3X}{\beta_T^2},
\end{aligned} \tag{A.7}$$

and, similarly,

$$\nabla_X G(U,V) = \rho \sum_i \nabla G_{1i}(U) + \rho G_2(U).$$

Therefore, we have

$$\begin{aligned}
&\|\nabla_X G(X,Y) - \nabla_X G(U,V)\|_F \\
&= \|\rho \sum_i [\nabla G_{1i}(X) - \nabla G_{1i}(U)] + \rho[\nabla G_2(X) - \nabla G_2(U)]\|_F \\
&\leq \|\rho \sum_i [\nabla G_{1i}(X) - \nabla G_{1i}(U)]\|_F + \rho\|\nabla G_2(X) - \nabla G_2(U)\|_F \\
&= \rho\sqrt{\sum_i \|\nabla G_{1i}(X) - \nabla G_{1i}(U)\|_F^2} + \rho\|\nabla G_2(X) - \nabla G_2(U)\|_F,
\end{aligned} \tag{A.8}$$

where the last equality is due to the fact that each $\nabla G_{1i}(X) - \nabla G_{1i}(U)$ is a matrix with all but the i-th row being zero. Denote

$$z_1 \triangleq \frac{3\|X\|_F^2}{2\beta_T^2}, z_2 \triangleq \frac{3\|U\|_F^2}{2\beta_T^2}. \tag{A.9}$$

Then by (A.9), (A.6) and the triangle inequality we have

$$\begin{aligned}
\frac{\beta_T^2}{3}\|\nabla G_2(X) - \nabla G_2(U)\|_F &= \|G_0'(z_1)X - G_0'(z_2)U\|_F \\
&\leq |G_0'(z_1)|\|X - U\|_F + |G_0'(z_1) - G_0'(z_2)|\|U\|_F.
\end{aligned} \tag{A.10}$$

By the definitions of z_1, z_2 in (A.9) and using $\|X\|_F \leq \beta_0, \|Y\|_F \leq \beta_0$, we have

$$\begin{aligned}
|z_1 - z_2| = \frac{3}{2\beta_T^2}(\|X\|_F^2 - \|U\|_F^2) &= \frac{3}{2\beta_T^2}(\|X\|_F + \|U\|_F)(\|X\|_F - \|U\|_F) \\
&\leq \frac{3\beta_0}{\beta_T^2}\|X - U\|_F.
\end{aligned} \tag{A.11}$$

According to (A.1) and the definitions of z_1, z_2 in (A.9), we have

$$\max\{z_1, z_2\} \leq \frac{3}{2}\frac{\beta_0^2}{\beta_T^2}. \tag{A.12}$$

We can bound the first and second order derivative of G_0 as follows:

$$G_0'(z) = I_{[1,\infty]}(z)2(z-1) \leq 3\frac{\beta_0^2}{\beta_T^2}, \ \ \forall z \in [0, \frac{3}{2}\frac{\beta_0^2}{\beta_T^2}], \tag{A.13}$$

$$G_0''(z) = 2I_{[1,\infty]}(z) \leq 2, \ \ \forall z \in [0, \infty). \tag{A.14}$$

By the mean value theorem and (A.14), we have

$$|G_0'(z_1) - G_0'(z_2)| \leq 2|z_1 - z_2| \overset{\text{(A.11)}}{\leq} \frac{6\beta_0}{\beta_T^2}\|X - U\|_F. \tag{A.15}$$

Plugging (A.13) (with $z = z_1$) and (A.15) into (A.10), we obtain

$$\frac{\beta_T^2}{3}\|\nabla G_2(X) - \nabla G_2(U)\|_F \le 3\frac{\beta_0^2}{\beta_T^2}\|X-U\|_F + \frac{6\beta_0}{\beta_T^2}\|X-U\|_F\|U\|_F \le 9\frac{\beta_0^2}{\beta_T^2}\|X-U\|_F$$

$$\Longrightarrow \qquad \|\nabla G_2(X) - \nabla G_2(U)\|_F \le 27\frac{\beta_0^2}{\beta_T^4}\|X-U\|_F. \tag{A.16}$$

Since $\|X^{(i)}\|_F \le \|X\|_F \le \beta_0, \|U^{(i)}\| \le \|U\|_F \le \beta_0$, by an argument analogous to that for (A.16), we can prove

$$\|\nabla G_{1i}(X) - \nabla G_{1i}(U)\|_F \le 27\frac{\beta_0^2}{\beta_1^4}\|X^{(i)} - U^{(i)}\|, \ \ \forall\, i,$$

which further implies

$$\sqrt{\sum_i \|\nabla G_{1i}(X) - \nabla G_{1i}(U)\|^2} \le 27\frac{\beta_0^2}{\beta_1^4}\sqrt{\sum_i \|X^{(i)} - U^{(i)}\|^2} = 27\frac{\beta_0^2}{\beta_1^4}\|X-U\|_F. \tag{A.17}$$

Plugging (A.16) and (A.17) into (A.8), we obtain

$$\|\nabla_X G(X,Y) - \nabla_X G(U,V)\|_F \le 54\rho\frac{\beta_0^2}{\beta_1^4}\|X-U\|_F.$$

Similarly, we can prove

$$\|\nabla_Y G(X,Y) - \nabla_Y G(U,V)\|_F \le 54\rho\frac{\beta_0^2}{\beta_2^4}\|Y-V\|_F \le 54\rho\frac{\beta_0^2}{\beta_1^4}\|Y-V\|_F,$$

where the last inequality is due to $\beta_1 = \beta_T\sqrt{\frac{3\mu r}{m}} \le \beta_T\sqrt{\frac{3\mu r}{n}} = \beta_2$. Combining the above two relations yields (A.4).

Finally, we combine (A.2) and (A.4) to obtain

$$\begin{aligned}\|\nabla\tilde{F}(X,Y) - \nabla\tilde{F}(U,V)\|_F &\le \|\nabla F(X,Y) - \nabla F(U,V)\|_F + \|\nabla G(X,Y) - \nabla G(U,V)\|_F \\ &\le \left(4\beta_0^2 + 54\rho\frac{\beta_0^2}{\beta_1^4}\right)\|(X,Y) - (U,V)\|_F,\end{aligned}$$

which finishes the proof of Claim 4.1.1. □

Remark: If we further assume that the norm of each $X^{(i)}$ (resp. $Y^{(j)}$) is bounded by $O(\beta_1)$ (resp. $O(\beta_2)$), the Lipschitz constant can be improved to $4\beta_0^2 + 54\rho\frac{\beta_0^2}{\beta_T^4}$.

A.2 Solving the Subproblem of Algorithm 3

The subproblem of Algorithm 3 for the row vector $X^{(i)}$ is

$$\min_{X^{(i)}} \tilde{F}(X_k^{(1)}, \ldots, X_k^{(i-1)}, X^{(i)}, X_{k-1}^{(i+1)} \ldots, X_{k-1}^{(m)}, Y_{k-1}) + \frac{\lambda_0}{2}\|X^{(i)} - X_{k-1}^{(i)}\|^2.$$

For simplicity, denote $X^{(i)} = x_i$, $X_{k-1}^{(i)} = \bar{x}_i$, $X_k^{(j)} = x_j, 1 \le j \le i-1$, $X_{k-1}^{(j)} = x_j, i+1 \le j \le m$, and $Y_{k-1}^{(j)} = y_j, 1 \le j \le n$. Then the above problem becomes

$$\min_{x_i} \tilde{F}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_m, y_1, \ldots, y_n) + \frac{\lambda_0}{2}\|x_i - \bar{x}_i\|^2.$$

The optimal solution x_i^* to this subproblem satisfies the equation $\nabla_{x_i}\tilde{F} = 0$, i.e.

$$Ax_i - b + g(\|x_i\|)x_i = 0, \tag{A.18}$$

where $A = \sum_{j\in\Omega_i^x} y_j y_j^T + \lambda_0 I$ is a symmetric PD (positive definite) matrix, $b = \sum_{j\in\Omega_i^x} M_{ij}y_j + \lambda_0\bar{x}_i$, and g is a function defined as

$$g(z) = \rho\frac{3}{\beta_1^2}G_0'(\frac{3z^2}{2\beta_1^2}) + \rho\frac{3}{\beta_T^2}G_0'(\frac{3(z^2+\xi_i)}{2\beta_T^2}),$$

in which $\xi_i = \sum_{j\neq i}\|x_j\|^2$ is a constant. Note that g has the following properties: a) $g(z) = 0$ when $z^2 \le \min\{\frac{2\beta_1^2}{3}, \frac{2\beta_T^2}{3} - \xi_i\}$; b) g is an increasing function in $[0, \infty)$. The equation (A.18) is equivalent to

$$x_i = (A + g(\|x_i\|)I)^{-1}b. \tag{A.19}$$

Suppose the eigendecomposition of A is $B\Lambda B^T$ and let $\Phi = B^T bb^T B$, then (A.19) implies

$$\begin{aligned}
\|x_i\|^2 =& \|(A + g(\|x_i\|)I)^{-1}b\|^2 = \mathrm{Tr}((A + g(\|x_i\|)I)^{-2}bb^T) \\
&= \mathrm{Tr}((\Lambda + g(\|x_i\|)I)^{-2}\Phi) = \sum_{k=1}^{r}\frac{\Phi_{kk}}{(\Lambda_{kk} + g(\|x_i\|))^2}, \\
\Longrightarrow \qquad & 1 = \frac{1}{\|x_i\|^2}\sum_{k=1}^{r}\frac{\Phi_{kk}}{(\Lambda_{kk} + g(\|x_i\|))^2},
\end{aligned} \tag{A.20}$$

where Z_{kk} denotes the (k,k)-th entry of matrix Z. Since A and Φ are PSD (positive semidefinite) matrices, we have $\Phi_{kk} \ge 0, \Lambda_{kk} \ge 0$. The righthand side of (A.20) is a

decreasing function of $\|x_i\|$, thus the equation (A.20) can be solved via a simple bisection procedure. After obtaining the norm of the optimal solution $z^* = \|x_i^*\|$, the optimal solution x_i^* can be obtained by (A.19), i.e.

$$x_i^* = (A + g(z^*)I)^{-1}b. \tag{A.21}$$

Similarly, the subproblem for $Y^{(j)}$ can also be solved by a bisection procedure.

A.3 Proof of Proposition 4.3.1

A.3.1 Matrix norm inequalities

We first prove some basic inequalities related to the matrix norms. These simple results will be used in the proof of Propositions 4.3.1 and 4.3.2.

Proposition A.3.1 *If $A, B \in \mathbb{R}^{n_1 \times n_2}$, then*

$$\|A - B\|_2 \geq \sigma_{\min}(A) - \sigma_{\min}(B). \tag{A.22}$$

Proof: $\sigma_{\min}(A) = \min_{\|v\|=1} \|Av\| \leq \min_{\|v\|=1}(\|Bv\| + \|(A-B)v\|) \leq \min_{\|v\|=1} \|Bv\| + \|A - B\| = \sigma_{\min}(B) + \|A - B\|$.

Proposition A.3.2 *For any $A \in \mathbb{R}^{n_1 \times n_2}, B \in \mathbb{R}^{n_2 \times n_3}$, we have*

$$\sigma_{\min}(AB) \leq \sigma_{\min}(A)\|B\|_2. \tag{A.23}$$

Proof: $\sigma_{\min}(AB) = \min_{v \in \mathbb{R}^{n_1 \times 1}, \|v\|=1} \|v^T AB\| \leq \min_{v \in \mathbb{R}^{n_1 \times 1}, \|v\|=1} \|v^T A\| \|B\|_2 = \sigma_{\min}(A)\|B\|_2$.

Proposition A.3.3 *Suppose $A, B \in \mathbb{R}^{n_1 \times n_2}$ and $c_i A^{(i)} = B^{(i)}$, where $c_i \in \mathbb{R}$ and $|c_i| \leq 1$, for $i = 1, \ldots, n_1$ (recall that $Z^{(i)}$ denotes the i-th row of Z). Then*

$$\|B\|_2 \leq \|A\|_2.$$

Proof: For simplicity, denote $a_i \triangleq (A^{(i)})^T, b_i \triangleq (B^{(i)})^T$. Then

$$\|B\|_2^2 = \max_{\|v\|=1} \|Bv\|^2 = \max_{\|v\|=1} \sum_i (b_i^T v)^2 = \max_{\|v\|=1} \sum_i c_i^2 (a_i^T v)^2 \leq \max_{\|v\|=1} \sum_i (a_i^T v)^2 = \|A\|_2^2.$$

Corollary A.3.1 *Suppose $B \in \mathbb{R}^{n_1 \times n_2}$ is a submatrix of $A \in \mathbb{R}^{m_1 \times m_2}$, then*

$$\|B\|_2 \leq \|A\|_2. \tag{A.24}$$

Proof: By Proposition A.3.3, we have

$$\|(X_1, X_2)\|_2 \geq \|(X_1, 0)\|_2 = \|X_1\|_2.$$

Without loss of generality, suppose $A = \begin{bmatrix} B & B_1 \\ B_2 & B_3 \end{bmatrix}$. Applying the above inequality twice, we get

$$\|A\|_2 \geq \|(B, B_1)\|_2 \geq \|B\|_2.$$

Proposition A.3.4 *For any $A \in \mathbb{R}^{n_1 \times n_2}, B \in \mathbb{R}^{n_2 \times n_3}$, we have*

$$\|AB\|_F \leq \|A\|_2 \|B\|_F, \tag{A.25a}$$

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2. \tag{A.25b}$$

Further, if $n_1 \geq n_2$, then

$$\sigma_{\min}(A)\|B\|_F \leq \|AB\|_F, \tag{A.26a}$$

$$\sigma_{\min}(A)\|B\|_2 \leq \|AB\|_2. \tag{A.26b}$$

Proof: Assume the SVD of A is $A_1 D A_2$, where $A_1 \in \mathbb{R}^{n_1 \times n_1}, A_2 \in \mathbb{R}^{n_2 \times n_2}$ are orthonormal matrices and $D \in \mathbb{R}^{n_1 \times n_2}$ has nonzero entries $D_{ii}, i = 1, \ldots, \min\{n_1, n_2\}$. Note that

$$\sigma_{\min}(A) \leq D_{ii} \leq \|A\|_2, \forall\ i.$$

Let $B' = A_2 B$ and suppose the i-th row of B' is $b_i, i = 1, \ldots, n_2$, then

$$\|AB\|_F^2 = \|DA_2B\|_F^2 = \|DB'\|_F^2 = \sum_{i=1}^{\min\{n_1, n_2\}} D_{ii}^2 \|b_i\|^2. \tag{A.27}$$

The the RHS (right hand side) can be bounded from above as

$$\sum_{i=1}^{\min\{n_1,n_2\}} D_{ii}^2\|b_i\|^2 \leq \|A\|_2^2 \sum_{i=1}^{\min\{n_1,n_2\}} \|b_i\|^2 \leq \|A\|_2^2 \sum_{i=1}^{n_2} b_i^2 = \|A\|_2^2\|B'\|_F^2 = \|A\|_2^2\|B\|_F^2.$$

Combining the above relation and (A.27) leads to (A.25a).

If $n_1 \geq n_2$, then $\min\{n_1, n_2\} = n_2$, and the RHS of (A.27) can be bounded from below as

$$\sum_{i=1}^{\min\{n_1,n_2\}} D_{ii}^2 \|b_i\|^2 = \sum_{i=1}^{n_2} D_{ii}^2 \|b_i\|^2 \geq \sigma_{\min}(A)^2 \sum_{i=1}^{n_2} \|b_i\|^2$$
$$= \sigma_{\min}(A)^2 \|B'\|_F^2 = \sigma_{\min}(A)^2 \|B\|_F^2.$$

Combining the above relation and (A.27) leads to (A.26a).

Next we prove the inequalities related to the spectral norm. We have

$$\|AB\|_2 = \|DA_2B\|_2 = \|DB'\|_2 = \max_{\|v\| \leq 1, v \in \mathbb{R}^{n_1 \times 1}} \|v^T DB'\|. \tag{A.28}$$

Note that $\{v^T D \mid \|v\| \leq 1, v \in \mathbb{R}^{n_1 \times 1}\} \subseteq \{u^T \mid u \in \mathbb{R}^{n_2 \times 1}, \|u\| \leq \|A\|_2\}$, thus the RHS of (A.28) can be bounded from above as

$$\max_{\|v\| \leq 1, v \in \mathbb{R}^{n_1 \times 1}} \|v^T DB'\| \leq \max_{u \in \mathbb{R}^{n_2 \times 1}, \|u\| \leq \|A\|_2} \|u^T B'\| = \|A\|_2 \|B'\|_2 = \|A\|_2 \|B\|_2.$$

Combining the above relation and (A.28) leads to (A.25b).

If $n_1 \geq n_2$, then $\{u^T \mid u \in \mathbb{R}^{n_2 \times 1}, \|u\| \leq \sigma_{\min}(A)\} \subseteq \{v^T D \mid \|v\| \leq 1, v \in \mathbb{R}^{n_1 \times 1}\}$ (in fact, for any $\|u\| \leq \sigma_{\min}(A)$, let $v_i = u_i / D_{ii}, i = 1, \ldots, n_2$ and $v_i = 0, n_2 < i \leq n_1$, where v_i denotes the i-th entry of v, then $v^T D = u^T$ and $\|v\| \leq 1$). Thus the RHS of (A.28) can be bounded from below as

$$\max_{\|v\| \leq 1, v \in \mathbb{R}^{n_1 \times 1}} \|v^T DB'\| \geq \max_{u \in \mathbb{R}^{n_2 \times 1}, \|u\| \leq \sigma_{\min}(A)} \|u^T B'\| = \sigma_{\min}(A) \|B'\|_2 = \sigma_{\min}(A) \|B\|_2.$$

Combining the above relation and (A.28) leads to (A.26b). □

A.3.2 Proof of Proposition 4.3.1

Let M, X, Y satisfy the condition (4.37). First, we specify the choice of U, V. Suppose the SVD of M is $M = \hat{U}\Sigma\hat{V} = Q_1 \tilde{\Sigma} Q_2^T$, where $Q_1 \in \mathcal{R}^{m \times m}, Q_2 \in \mathcal{R}^{n \times n}$ are unitary matrices, and $\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$. Suppose $Q_1 = (Q_{11}, Q_{12})$, $Q_2 = (Q_{21}, Q_{22})$, where $Q_{11} = \hat{U} \in \mathcal{R}^{m \times r}, Q_{21} = \hat{V} \in \mathcal{R}^{n \times r}$ are incoherent matrices, and $Q_{12} \in \mathbb{R}^{m \times (m-r)}, Q_{22} \in \mathbb{R}^{n \times (n-r)}$. Let us write X, Y as

$$X = Q_1 \begin{pmatrix} X_1' \\ X_2' \end{pmatrix}, \quad Y = Q_2 \begin{pmatrix} Y_1' \\ Y_2' \end{pmatrix}, \tag{A.29}$$

where $X_1', Y_1' \in \mathbb{R}^{r\times r}, X_2' \in \mathbb{R}^{(m-r)\times r}, Y_2' \in \mathbb{R}^{(n-r)\times r}$. Define

$$U \triangleq Q_1 \begin{pmatrix} U_1' \\ 0 \end{pmatrix}, \quad V \triangleq Q_2 \begin{pmatrix} V_1' \\ 0 \end{pmatrix}, \tag{A.30}$$

where

$$U_1' = (1-\bar{\eta})X_1', \;\; V_1' = \frac{1}{1-\bar{\eta}}\Sigma(X_1')^{-T},$$

in which

$$\bar{\eta} \triangleq \frac{d}{\Sigma_{\min}} \leq \frac{1}{11}.$$

The definition of V_1' is valid since X_1' is invertible (otherwise, $\mathrm{rank}(X_1'(Y_1')^T) \leq \mathrm{rank}(X_1') \leq r-1$, thus $d \geq \|\Sigma - X_1'(Y_1')^T\|_F \overset{(A.22)}{\geq} \Sigma_{\min} - \sigma_{\min}(X_1'(Y_1')^T) = \Sigma_{\min}$, which contradicts (4.37a).By this definition, we have

$$U_1'(V_1')^T = (1-\bar{\eta})X_1'(V_1')^T = \Sigma. \tag{A.31}$$

Now, we prove that U, V defined in (A.30) satisfy the requirement (4.38). The requirement (4.38a) $UV^T = M$ follows from (A.31) and (A.30). The requirement (4.38b) $\|U\|_F \leq (1-\frac{d}{\Sigma_{\min}})\|X\|_F$ can be proved as follows:

$$\|U\|_F = \|U_1'\|_F = (1-\frac{d}{\Sigma_{\min}})\|X_1'\|_F \leq (1-\frac{d}{\Sigma_{\min}})\|X\|_F.$$

As a side remark, the following variant of the requirement (4.38b) also holds:

$$\|U\|_2 \leq (1-\frac{d}{\Sigma_{\min}})\|X\|_2. \tag{A.32}$$

In fact, $\|U\|_2 = \|U_1'\|_2 = (1-\frac{d}{\Sigma_{\min}})\|X_1'\|_2 \overset{(A.24)}{\leq} (1-\frac{d}{\Sigma_{\min}})\left\|\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}\right\|_2 = (1-\frac{d}{\Sigma_{\min}})\|X\|_2$.

To prove the requirement (4.38c), we first provide the bounds on $\|X_2'\|_F, \|V_1' -$

$Y_1'\|_F, \|Y_2'\|_F$. Note that

$$\begin{aligned} d^2 =& \|M - XY^T\|_F^2 \\ =& \left\| \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} - Q_1^T XY^T Q_2 \right\|_F^2 \\ \overset{(A.29)}{=}& \left\| \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} X_1'(Y_1')^T & X_1'(Y_2')^T \\ X_2'(Y_1')^T & X_2'(Y_2')^T \end{pmatrix} \right\|_F^2 \\ =& \|\Sigma - X_1'(Y_1')^T\|_F^2 + \|X_1'(Y_2')^T\|_F^2 + \|X_2'(Y_1')^T\|_F^2 + \|X_2'(Y_2')^T\|_F^2. \\ \overset{(A.31)}{=}& \|X_1'((1-\bar{\eta})V_1' - Y_1')^T\|_F^2 + \|X_1'(Y_2')^T\|_F^2 + \|X_2'(Y_1')^T\|_F^2 + \|X_2'(Y_2')^T\|_F^2. \end{aligned} \tag{A.33}$$

Intuitively, since $\|X_1'\|_F, \|Y_1'\|_F$ are $O(1)$, we can upper bound $\|(1-\bar{\eta})V_1' - Y_1'\|_F$, $\|Y_2'\|_F$, $\|X_2'\|_F$ as $O(d)$. More rigorously, it follows from (A.33) that $d \geq \|X_1'((1-\bar{\eta})V_1' - Y_1')^T\|_F \overset{(A.26a)}{\geq} \sigma_{\min}(X_1')\|(1-\bar{\eta})V_1' - Y_1'\|_F$ and, similarly, $d \geq \sigma_{\min}(X_1')\|(Y_2')^T\|_F$, $d \geq \sigma_{\min}(Y_1')\|(X_2')^T\|_F$. These three inequalities imply

$$\begin{aligned} \|(1-\bar{\eta})V_1' - Y_1'\|_F \leq \frac{d}{\sigma_{\min}(X_1')}, \\ \|Y_2'\|_F \leq \frac{d}{\sigma_{\min}(X_1')}, \quad \|X_2'\|_F \leq \frac{d}{\sigma_{\min}(Y_1')}. \end{aligned} \tag{A.34}$$

We can lower bound $\sigma_{\min}(X_1')$ and $\sigma_{\min}(Y_1')$ as

$$\sigma_{\min}(X_1') \geq \frac{10\Sigma_{\min}}{11\beta_T}, \quad \sigma_{\min}(Y_1') \geq \frac{10\Sigma_{\min}}{11\beta_T}. \tag{A.35}$$

To prove (A.35), notice that (A.33) implies that $d \geq \|\Sigma - X_1'(Y_1')^T\|_F \geq \|\Sigma - X_1'(Y_1')^T\|_2 \overset{(A.22)}{\geq} \Sigma_{\min} - \sigma_{\min}(X_1'(Y_1')^T)$, which further implies

$$\sigma_{\min}(X_1'(Y_1')^T) \geq \Sigma_{\min} - d \geq \frac{10}{11}\Sigma_{\min}.$$

According to Proposition A.3.2, we have $\sigma_{\min}(X_1'(Y_1')^T) \leq \sigma_{\min}(X_1')\|Y_1'\|_2$. Combining this inequality with the above relation, we get $\sigma_{\min}(X_1')\|Y_1'\|_2 \geq \sigma_{\min}(X_1'(Y_1')^T) \geq 5\Sigma_{\min}/6$, which further implies

$$\sigma_{\min}(X_1') \geq \frac{10\Sigma_{\min}}{11\|Y_1'\|_2}. \tag{A.36}$$

Similarly, we have

$$\sigma_{\min}(Y_1') \geq \frac{10\Sigma_{\min}}{11\|X_1'\|_2}. \tag{A.37}$$

Plugging $\|Y_1'\|_2 \le \|Y_1'\|_F \le \|Y\|_F \le \beta_T$ and similarly $\|X_1'\|_2 \le \beta_T$ into (A.36) and (A.37), we obtain (A.35).

Combining (A.35) and (A.34), we obtain

$$\max\{\|(1-\bar{\eta})V_1' - Y_1'\|_F, \|X_2'\|_F, \|Y_2'\|_F\} \le \frac{11}{10}\frac{d}{\Sigma_{\min}}\beta_T \le \frac{1}{10}\beta_T. \tag{A.38}$$

We can bound the norm of V_1' as

$$\begin{aligned}\|V_1'\|_F = \frac{1}{1-\bar{\eta}}\|(1-\bar{\eta})V_1'\|_F &\le \frac{1}{1-\bar{\eta}}(\|(1-\bar{\eta})V_1' - Y_1'\|_F + \|Y_1'\|_F) \\ &\overset{(A.38)}{\le} \frac{11}{10}\left(\frac{1}{10}\beta_T + \beta_T\right) \le \left(\frac{11}{10}\right)^2 \beta_T.\end{aligned} \tag{A.39}$$

Combining this relation with (A.38), we have

$$\|V_1' - Y_1'\|_F \le \|(1-\bar{\eta})V_1' - Y_1'\|_F + \bar{\eta}\|V_1'\|_F \le \frac{11}{10}\frac{d}{\Sigma_{\min}}\beta_T + \bar{\eta}\left(\frac{11}{10}\right)^2\beta_T \le \frac{7\beta_T}{3\Sigma_{\min}}d.$$

From (A.38) and the above relation we obtain

$$\begin{aligned}\|U - X\|_F &= \|X_2'\|_F \le \frac{11\beta_T}{10\Sigma_{\min}}d \le \frac{6\beta_T}{5\Sigma_{\min}}d, \\ \|V - Y\|_F &= \sqrt{\|V_1' - Y_1'\|_F^2 + \|Y_2'\|_F^2} \le \sqrt{\left(\frac{7}{3}\right)^2 + \left(\frac{11}{10}\right)^2}\frac{\beta_T}{\Sigma_{\min}}d \le \frac{3\beta_T}{\Sigma_{\min}}d,\end{aligned}$$

which finishes the proof of the requirement (4.38c).

As a side remark, the requirement (4.38c) can be slightly improved to

$$\|U - X\|_F \le \frac{6\|Y\|_2}{5\Sigma_{\min}}d, \quad \|V - Y\|_F \le \frac{3\|X\|_2}{\Sigma_{\min}}d. \tag{A.40}$$

In fact, plugging $\|X_1'\|_2 \overset{(A.24)}{\le} \|\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}\|_2 = \|X\|_2$ and similarly $\|Y_1'\|_2 \le \|Y\|_2$ into (A.36) and (A.37), we obtain $\sigma_{\min}(X_1') \ge \frac{5\Sigma_{\min}}{6\|Y\|_2}$, $\sigma_{\min}(Y_1') \ge \frac{5\Sigma_{\min}}{6\|X\|_2}$. Combining with (A.34), we obtain (A.40). This inequality will be used in the proof of Claim 4.4.2 in Appendix A.5.1.

At last, we prove the requirement (4.38d). By the definitions of U, V in (A.30), we have

$$\begin{aligned}U &= (Q_{11}, Q_{12})\begin{pmatrix} U_1' \\ 0 \end{pmatrix} = Q_{11}U_1', \\ V &= (Q_{21}, Q_{22})\begin{pmatrix} V_1' \\ 0 \end{pmatrix} = Q_{21}V_1'.\end{aligned} \tag{A.41}$$

The assumption that M is μ-incoherent implies

$$\|Q_{11}^{(i)}\|^2 = \|\hat{U}^{(i)}\|^2 \leq \frac{r\mu}{m}, \ \ \|Q_{21}^{(i)}\|^2 = \|\hat{V}^{(j)}\|^2 \leq \frac{r\mu}{n}, \ \ \forall\, i, j.$$

Notice the following fact: for any matrix $A \in \mathbb{R}^{K\times r}, B \in \mathbb{R}^{r\times r}$, where $K \in \{m, n\}$, we have

$$\|(AB)^{(i)}\|^2 = \|A^{(i)}B\|^2 \leq \|A^{(i)}\|^2\|B\|_F^2.$$

Therefore, we have (using the fact $\|U_1'\|_F \leq \|X_1'\|_F \leq \|X\|_F \leq \beta_T$ and (A.39))

$$\begin{aligned} \|U^{(i)}\|^2 = \|(Q_{11}U_1')^{(i)}\|^2 &\leq \|Q_{11}^{(i)}\|^2\|U_1'\|_F^2 \leq \frac{r\mu}{m}\beta_T^2; \\ \|V^{(j)}\|^2 = \|(Q_{21}V_1')^{(j)}\|^2 \leq \frac{r\mu}{n}\|V_1'\|_F^2 &\overset{(A.39)}{\leq} \left(\frac{11}{10}\right)^4 \frac{r\mu}{n}\beta_T^2 \leq \frac{3}{2}\frac{r\mu}{n}\beta_T^2, \end{aligned} \tag{A.42}$$

which finishes the proof the requirement (4.38d).

A.4 Proof of Proposition 4.3.2

A.4.1 Transformation to a simpler problem

We first transform the problem to a simpler problem that only involves $r \times r$ matrices. In particular, we will show that to prove Proposition 4.3.2 we only need to prove Proposition A.4.1.

Similar to the proof of Proposition 4.3.1, we use $Q_1 \in \mathbb{R}^{m\times m}, Q_2 \in \mathbb{R}^{n\times n}$ to denote the SVD factors of M (Q_1 and Q_2 are unitary matrices), and write X, Y as

$$X = Q_1 \begin{pmatrix} X_1' \\ X_2' \end{pmatrix}, \quad Y = Q_2 \begin{pmatrix} Y_1' \\ Y_2' \end{pmatrix}.$$

Define

$$U = Q_1 \begin{pmatrix} U_1' \\ 0 \end{pmatrix}, \quad V = Q_2 \begin{pmatrix} V_1' \\ 0 \end{pmatrix}, \tag{A.43}$$

where $U_1' \in \mathbb{R}^{r\times r}$ and $V_1' \in \mathbb{R}^{r\times r}$ are to be determined.

We can convert the conditions on U, V to the conditions on U_1', V_1'. As proved in Appendix A.3 (combining (A.34) and (A.35)),

$$\|X_2'\|_F \leq \frac{6\beta_T}{5\Sigma_{\min}}d, \ \ \|Y_2'\|_F \leq \frac{6\beta_T}{5\Sigma_{\min}}d. \tag{A.44}$$

Obviously, the condition (4.39a) implies the following condition on X_1', Y_1':

$$d' \triangleq \|\Sigma - (X_1')(Y_1')^T\| \leq \frac{\Sigma_{\min}}{C_d r}. \tag{A.45}$$

Using (A.44) and the facts $\|X\|_F = \sqrt{\|X_1'\|_F^2 + \|X_2'\|_F^2}$ and $\|Y\|_F = \sqrt{\|Y_1'\|_F^2 + \|Y_2'\|_F^2}$, the condition (4.39b) implies the following condition on X_1', Y_1':

$$\sqrt{\frac{3}{5}}\beta_T \leq \|X_1'\|_F \leq \beta_T, \ \ \sqrt{\frac{3}{5}}\beta_T \leq \|Y_1'\|_F \leq \beta_T. \tag{A.46}$$

We have the following proposition.

Proposition A.4.1 *There exist numerical constants C_d, C_T such that: if $X_1', Y_1' \in \mathbb{R}^{r\times r}$ satisfy* (A.45) *and* (A.46)*, where $\beta_T = \sqrt{C_T r \Sigma_{\max}}$, then there exist $U_1' \in \mathbb{R}^{r\times r}, V_1' \in \mathbb{R}^{r\times r}$ such that*

$$U_1'(V_1')^T = \Sigma, \tag{A.47a}$$

$$\|U_1'\|_F \leq \|X_1'\|_F, \ \|V_1'\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|Y_1'\|_F, \tag{A.47b}$$

$$\|U_1' - X_1'\|_F\|V_1' - Y_1'\|_F \leq 63\sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2, \ \max\{\|U_1' - X_1'\|_F, \|V_1' - Y_1'\|_F\} \leq \frac{58}{7}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d. \tag{A.47c}$$

We claim that Proposition A.4.1 implies Proposition 4.3.2. Since we have already proved that the conditions of Proposition 4.3.2 imply the conditions of Proposition A.4.1, we only need to prove that the conclusion of Proposition A.4.1 implies the conclusion of Proposition 4.3.2. In other words, we only need to show that if U_1', V_1' satisfy (A.47), then they satisfy the requirements (4.40).

The requirement (4.40a) $UV^T = M$ follows directly from (A.47a) and the definition of U, V in (A.43). The requirement (4.40b) can be proved as $\|V\|_F = \|V_1'\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|Y_1'\|_F \leq (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F$ and $\|U\|_F = \|U_1'\|_F \leq \|X\|_F$. Analogous to (A.42), the requirement (4.40d) can be proved as $\|V^{(j)}\|^2 = \|(Q_{21}V_1')^{(j)}\|^2 \leq \frac{r\mu}{n}\|V_1'\|_F^2 \leq \frac{r\mu}{n}\beta_T^2$ and, similarly, $\|U^{(i)}\|^2 \leq \frac{r\mu}{m}\beta_T^2$. At last, we prove the requirement (4.40c). The first relation

in (4.40c) can be proved as

$$
\begin{aligned}
&\|U-X\|_F\|V-Y\|_F \\
&= \sqrt{\|U_1'-X_1'\|_F^2+\|X_2'\|_F^2}\sqrt{\|V_1'-Y_1'\|_F^2+\|Y_2'\|_F^2} \\
&= \sqrt{\|U_1'-X_1'\|_F^2\|V_1'-Y_1'\|_F^2+\|X_2'\|_F^2\|V_1'-Y_1'\|_F^2+\|U_1'-X_1'\|_F^2\|Y_2'\|_F^2+\|X_2'\|_F^2\|Y_2'\|_F^2} \\
&\overset{\text{(A.44),(A.47c)}}{\leq} \sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2\sqrt{63^2+(\frac{6}{5})^2(\frac{58}{7})^2+(\frac{58}{7})^2(\frac{6}{5})^2+(\frac{6}{5})^4}, \\
&< 65\sqrt{r}\frac{\beta_T^2}{\Sigma_{\min}^2}d^2,
\end{aligned}
$$

where in the second last inequality we also use the fact $d' \leq d$. The second relation in (4.40c) can be proved by

$$
\begin{aligned}
\|U-X\|_F &= \sqrt{\|U_1'-X_1'\|_F^2+\|X_2'\|_F^2} \\
&\overset{\text{(A.44),(A.47c)}}{\leq} \sqrt{(\frac{6}{5})^2+(\frac{58}{7})^2}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d \leq \frac{17}{2}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d
\end{aligned}
$$

and a similar inequality for $\|V-Y\|_F$.

A.4.2 Preliminary analysis for the proof of Proposition A.4.1

To simplify the notations, from now on, we use X, Y, U, V, d to replace $X_1', Y_1', U_1', V_1', d'$ in Proposition (A.4.1).

Before presenting the formal proof, we analyze the problem through two simple examples. We denote the i-th row of X, Y as x_i, y_i, respectively.

In the first example (see Figure A.1), we set $r = 2$, $\Sigma = I$ (which implies $\Sigma_{\min} = \Sigma_{\max} = 1$), $d = 1/(C_d r)$ and

$$
\begin{aligned}
X &= \mathrm{Diag}\,(x_{11}, x_{22}) = \mathrm{Diag}\left(C, \frac{1-d/\sqrt{2}}{C}\right), \\
Y &= \mathrm{Diag}\,(y_{11}, y_{22}) = \mathrm{Diag}\left(\frac{1-d/\sqrt{2}}{C}, C\right),
\end{aligned}
\tag{A.48}
$$

where $C > 1$ is to be determined, and $\mathrm{Diag}(w_1, w_2)$ denotes a 2×2 diagonal matrix with diagonal entries w_1, w_2. In this setting $\beta_T = \sqrt{rC_T\Sigma_{\max}} = \sqrt{2C_T}$ is a large constant. Condition (A.45) holds since $\|XY^T - \Sigma\|_F = \|(1-d/\sqrt{2})I - I\|_F = d = 1/(C_d r)$. Note

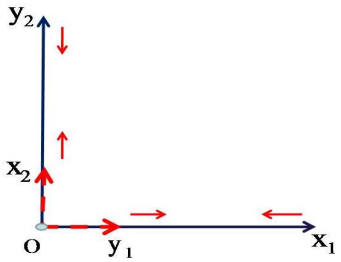


Figure A.1: Illustration of the first example. $X = (x_1^T, x_2^T) = \text{Diag}\,(x_{11}, x_{22})$, $Y = (y_1^T, y_2^T) = \text{Diag}\,(y_{11}, y_{22})$, where $x_{11} = y_{22} \gg x_{22} = y_{11}$ and $x_{11}y_{11} = x_{22}y_{22} = 1 - d/\sqrt{2}$. We use the following operation to define U, V: shrink x_1 and extend x_2 to obtain U, while keeping the norm invariant (i.e. $\|U\|_F = \|X\|_F$); shrink y_2 and extend y_1 to obtain V, while keeping the norm invariant (i.e. $\|V\|_F = \|Y\|_F$). We can prove that there exists an operation such that $u_{ii}v_{ii} = 1 > x_{ii}y_{ii}, i = 1, 2$.

that $\|X\|_F = \|Y\|_F = \sqrt{C^2 + \frac{(1-d/\sqrt{2})^2}{C^2}} \approx C$, thus there exists $C \in [\sqrt{3/5}\beta_T, \beta_T]$ so that (A.46) holds.

How should we define $U = \text{Diag}(u_{11}, u_{22}), V = \text{Diag}(v_{11}, v_{22})$ so that (A.47) holds? Due to the "symmetry" of X and Y in this example (by symmetry we mean $x_{11} = y_{22}, x_{22} = y_{11}$), we choose U, V such that $u_{11} = v_{22}, u_{22} = v_{11}$. Then the requirements (A.47a) and (A.47b) reduce to:

$$\begin{aligned} u_{11}u_{22} = 1 = \frac{x_{11}x_{22}}{1 - d/\sqrt{2}}, \\ u_{11}^2 + u_{22}^2 \le x_{11}^2 + x_{22}^2. \end{aligned} \tag{A.49}$$

It can be easily shown that there exist u_{11}, u_{22} satisfying (A.49). In fact, define $R = \|X\|_F = \sqrt{x_{11}^2 + x_{22}^2}$ and let a point (w_1, w_2) move along the circle $\{(w_1, w_2) \mid w_1^2 + w_2^2 = R^2\}$ from (x_{11}, x_{22}) to $(R/\sqrt{2}, R/\sqrt{2})$. During this process, the norm of (w_1, w_2) does not change and the product w_1w_2 monotonically increases from $x_{11}x_{22}$ to $R^2/2$. Therefore, there exist u_{11}, u_{22} satisfying (A.49) as long as $R^2/2 > x_{11}x_{22}/(1 - d/\sqrt{2})$. This inequality is equivalent to $(1-d/\sqrt{2})(x_{11}^2+x_{22}^2)/2 > x_{11}x_{22}$, which can be simplified to $(1-d/\sqrt{2})(x_{11}-x_{22})^2 > \sqrt{2}dx_{11}x_{22} = \sqrt{2}d(1-d/\sqrt{2})$, or equivalently, $(x_{11}-x_{22})^2 > \sqrt{2}d$. The last inequality holds when $x_{11} - x_{22} = C - (1 - d/\sqrt{2})/C$ is large enough (i.e. C is large enough).

To summarize, we will increase the small entry x_{22} (resp. y_{11}) and decrease the large entry x_{11} (resp. y_{22}) to obtain a more balanced diagonal matrix U (resp. V), which has the same norm as X (resp. Y). The percentage of increase in the small entry

x_{22} (resp. y_{11}) will be much larger than the percentage of decrease in the large entry x_{11} (resp. y_{22}), thus the products $x_{22}y_{22}$ and $x_{11}y_{11}$ will increase; in other words, the product UV^T of the more balanced matrices U, V will have larger entries than XY^T.

Note that the above idea of shrinking/extending works when there is a large imbalance in the lengths of the rows of X, Y, regardless of whether X, Y are diagonal matrices or not. By the assumption that $\|X\|_F$ and $\|Y\|_F$ are large, we know that there must be a row of X (resp. Y) that has large norm (here "large" means much larger than $1/\sqrt{r}$); however, it is possible that all rows of X and Y have large norm and there is no imbalance in terms of the lengths of the rows. See below for such an example.

In the second example (see Figure A.2), we still set $r = 2$, $\Sigma = I$, $d = 1/(C_d r)$. Suppose $X = (x_1^T, x_2^T)$, $Y = (y_1^T, y_2^T)$. We define $x_1 = (C, 0), x_2 = (-C\sin\alpha, C\cos\alpha)$ and $y_1 = (C\cos\alpha, C\sin\alpha), y_2 = (0, C)$, where C is a large constant, and $\alpha \in (0, \pi/2)$ is chosen so that

$$C^2\cos\alpha = 1 - d/\sqrt{2}. \tag{A.50}$$

When C is large, $\alpha \approx \arccos(1/C^2)$ is also large (i.e. close to $\pi/2$). Condition (A.45) holds since $\|XY^T - \Sigma\|_F = \|C^2\cos\alpha I - I\|_F = \|(1 - d/\sqrt{2})I - I\|_F = d = 1/(C_d r)$. Note that $\|X\|_F = \|Y\|_F = \sqrt{2}C$, so we can choose $C = \beta_T/\sqrt{2} = \sqrt{2C_T}/\sqrt{2} = \sqrt{C_T}$ so that (A.46) holds.

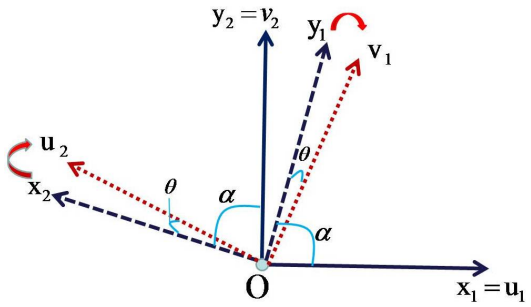


Figure A.2: Illustration of the second example. $X = (x_1^T, x_2^T)$, $Y = (y_1^T, y_2^T)$, where $x_1 = (C, 0), x_2 = (-C\sin\alpha, C\cos\alpha)$ and $y_1 = (C\cos\alpha, C\sin\alpha), y_2 = (0, C)$, where C is a large constant. Choose α so that $C^2\cos\alpha = 1 - d/\sqrt{2}$. We use the following operation to define $U = (u_1^T, u_2^T)$, $V = (v_1^T, v_2^T)$: rotate y_1 (resp. x_2) by angle θ to obtain v_1(resp. u_2), and let $u_2 = x_2, v_1 = y_1$. Here the angle of rotation θ is chosen so that $\langle u_1, v_1\rangle = \langle u_2, v_2\rangle = 1$.

How should we choose $U = (u_1^T, u_2^T)$, $V = (v_1^T, v_2^T)$ so that (A.47) holds? The idea for the first example no longer works since it requires that the difference of $\|x_1\|$ and $\|x_2\|$

(resp. $\|y_1\|$ and $\|y_2\|$) is large; however, in this example, $\|x_1\| - \|x_2\| = \|y_1\| - \|y_2\| = 0$. The key idea for this example is to use rotation. Rotating a vector does not change the norm, so requirement (A.46) will not be violated if u_i (resp. v_i) is obtained by rotating x_i(resp. y_i). For simplicity, we rotate y_1, x_2 to obtain v_1, u_2 respectively and let $u_1 = x_1, v_2 = y_2$ (see Figure A.2). Note that y_1 and x_2 should be rotated by the same angle as v_1 should be orthogonal to u_2 (since the off-diagonal entries of UV^T are zero). To increase the inner product $\langle x_i, y_i \rangle$ from $1 - d/\sqrt{2}$ to 1, we need to decrease the angle of x_i and y_i, thus y_1 (resp. x_2) should be rotated towards x_1(resp. y_2). Finally, let us specify the angle of rotation $\theta \triangleq \angle(y_1, v_1) = \angle(x_2, u_2)$. The requirement $\langle u_1, v_1 \rangle = 1$ is equivalent to $1 = \|u_1\|\|v_1\| \cos\angle(u_1, v_1) = \|x_1\|\|y_1\| \cos(\alpha - \theta)$, which can be rewritten as

$$1 = C^2 \cos(\alpha - \theta). \tag{A.51}$$

The right-hand side of (A.51) is an increasing function of θ, ranging from $C^2 \cos(\alpha) \overset{(A.50)}{=} 1 - d/\sqrt{2}$ to C^2 for $\theta \in [0, \alpha]$. Since 1 lies in the range $[1 - d/\sqrt{2}, C^2]$, there exists a unique θ so that (A.51) holds. One can further verify the requirement (A.47c), i.e. the difference of X (resp. Y) and U (resp. V) is small. As a rough summary, we rotate x_i, y_i to obtain u_i, v_i when the angle of x_i and y_i is large. This operation does not change the norm and can increase the inner product $\langle x_i, y_i \rangle$ to the desired amount (1 in this case).

In the above two examples, we have used two different operations: one is based on shrinking/extending, and the other is based on rotation. As we mentioned before, the first operation cannot deal with the second example; also, it is obvious that the second operation cannot deal with the first example (the angle between x_i and y_i is zero, so rotation only decreases the inner product). Therefore, both operations are necessary.

Are these two operations sufficient? Fortunately, the answer is yes for the case that XY^T is diagonal and $\langle x_i, y_i \rangle \leq \Sigma_i$ (we need extra effort to reduce the general problem to this case). When all the angles between x_i and y_i are smaller than a constant $\bar{\alpha}$, there must be an imbalance in the lengths of x_i, y_i's (to illustrate this, if all $\|x_i\| = \|y_i\|$, then $\|x_i\|^2 = \|x_i\|\|y_i\| \approx \Sigma_i / \cos\angle(x_i, y_i) \leq \Sigma_i / \cos(\bar{\alpha})$, which implies $\|X\|_F^2 \lesssim r\Sigma_{\max} / \cos(\bar{\alpha}) \ll \frac{3}{5} C_T r \Sigma_{\max} = \frac{3}{5}\beta_T^2$ for large enough C_T, a contradiction to (A.45)). Thus we can use the first operation (i.e. shrinking/extending the vectors x_i, y_i's) to obtain the desired U, V. When all the angles between x_i and y_i are larger than a constant $\bar{\alpha}$, we can use the second operation (i.e. rotating the vectors x_i, y_i's) to obtain

the desired U, V. In general, some angles may be larger than $\bar{\alpha}$ and others may be smaller, then a natural solution is to use the two operations *simultaneously*: use the first operation for the pairs (x_i, y_i) with small angles and the second operation for those with large angles.

We had a proof using the two operations simultaneously, but the bounds on $\|U - X\|_F, \|V - Y\|_F$ have large exponent of r. In the following subsection, we present a different proof that does not use the two operations simultaneously, but only use one of the two operations. The basic proof framework is summarized as follows. We first define $\hat{Y}$ so that $X\hat{Y} = \Sigma$; in other words, we try to satisfy the requirement (A.47a) first. Then we try to modify $\hat{Y}$ to satisfy the requirement (A.47b). In particular, we need to reduce the norm of $\hat{Y}$ and keep the norm of X unchanged, while maintaining the relation $X\hat{Y}^T = \Sigma$. We consider two cases: in Case 1, "most" angles between X and $\hat{Y}$ are smaller than $\bar{\alpha}$, and using the first operation (shrinking/extending) can obtain the desired U, V; in Case 2, "most" angles between X and $\hat{Y}$ are larger than $\bar{\alpha}$, and using the second operation (rotation) can obtain the desired U, V (see (A.59) for a precise definition of Case 1 and Case 2). The difference of this proof framework and the previous one is the following. In our previous proof framework, we need to take into account every pair x_i, y_i so that its inner product is modified to Σ_i, thus two operations have to be applied simultaneously. In contrast, in this new proof framework, $\langle x_i, \hat{y}_i \rangle$ is already Σ_i, and we only need to worry about the "overall" requirement that $\|\hat{Y}\|_F$ should be reduced, thus dealing only with the pairs with small angles (or only with the pairs with large angles) is enough to satisfy the requirement.

Finally, we would like to mention that when Σ is an identity matrix, the proof can be rather simple. In fact, in this case one can assume X to be diagonal by proper orthonormal transformation, and then assume Y to be diagonal since the off-diagonal entries are small. By just using the first operation (scaling of the diagonal entries), we can construct the desired U, V and the proof is similar to that in Appendix A.4.3. However, when Σ is not a diagonal matrix, it seems that the second operation has to be used and the proof becomes more involved.

A.4.3 Proof of Proposition A.4.1

As mentioned earlier, to simplify the notations, we use X, Y, U, V, d to replace $X_1', Y_1', U_1', V_1', d'$ in Proposition (A.4.1). Throughout the proof, we choose

$$C_T = 20, \tag{A.52}$$

and $C_d = 108$, which implies

$$\frac{d}{\Sigma_{\min}} \leq \frac{1}{108r}. \tag{A.53}$$

There are two "hard" requirements on U, V: (A.47a) and (A.47b). Our construction of U, V can be viewed as a two-step approach, whereby we satisfy one requirement in each step. In Step 1, we construct

$$\hat{Y} = \Sigma(\Sigma + D)^{-T}Y, \quad \text{where } D \triangleq XY^T - \Sigma,$$

then

$$X\hat{Y}^T = (XY^T)(XY^T)^{-1}\Sigma = \Sigma,$$

i.e. the first requirement is satisfied. Since the new $\hat{Y}$ may have higher norm than $\|Y\|_F$, in Step 2 we modify $X, \hat{Y}$ to U, V so that the product does not change, and $\|V\|_F \leq \|Y\|_F, \|U\|_F \leq \|X\|_F$.

Claim A.4.1 *Let $\hat{Y} = \Sigma(\Sigma + D)^{-T}Y$, then*

$$\eta \triangleq 1 - \frac{\|Y\|_F}{\|\hat{Y}\|_F} \leq \frac{d}{\Sigma_{\min}}, \tag{A.54a}$$

$$\|Y - \hat{Y}\|_F \leq \frac{d}{\Sigma_{\min} - d}\|Y\|_F. \tag{A.54b}$$

Proof of Claim A.4.1: By the definition of $\hat{Y}$ we have $Y = (\Sigma + D)^T\Sigma^{-1}\hat{Y}$, then we have

$$\begin{aligned} \|Y - \hat{Y}\|_F &= \|(\Sigma + D)^T\Sigma^{-1}\hat{Y} - \hat{Y}\|_F = \|D^T\Sigma^{-1}\hat{Y}\|_F \\ &\leq \|D^T\Sigma^{-1}\|_F\|\hat{Y}\|_F \leq \|D^T\|_F\Sigma_{\min}^{-1}\|\hat{Y}\|_F = \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F. \end{aligned} \tag{A.55}$$

Using the triangular inequality and (A.55), we have

$$\begin{aligned} & \|\hat{Y}\|_F \leq \|Y - \hat{Y}\|_F + \|Y\|_F \leq \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F + \|Y\|_F, \\ \implies \quad & \|Y\|_F \geq (1 - \frac{d}{\Sigma_{\min}})\|\hat{Y}\|_F. \end{aligned} \tag{A.56}$$

The first desired inequality (A.54a) follows immediately from (A.56), and the second desired inequality (A.54b) is proved by combining (A.56) and (A.55). □

Combining (A.54a) and (A.53), we obtain

$$\eta \leq \frac{1}{108r}. \tag{A.57}$$

If $\eta \leq 0$, i.e. $\|\hat{Y}\|_F \leq \|Y\|_F$, then $U = X, V = \hat{Y}$ already satisfy (A.47). From now on, we assume $\eta > 0$, i.e. $\|\hat{Y}\|_F > \|Y\|_F$. Denote $x_i^T, \hat{y}_i^T, u_i^T, v_i^T$ as the *i-th* row of $X, \hat{Y}, U, V$, respectively. Denote $\alpha_i \triangleq \angle(x_i, \hat{y}_i)$, i.e. the angle between the two vectors x_i and $\hat{y}_i$. Since $\langle x_i, \hat{y}_i \rangle = \Sigma_i > 0$, we have $\alpha_i \in [0, \frac{\pi}{2})$. Without loss of generality, assume

$$\alpha_1, \ldots, \alpha_s > \frac{3}{8}\pi, \quad \alpha_{s+1}, \ldots, \alpha_r \leq \frac{3}{8}\pi, \tag{A.58}$$

where $s \in \{0, 1, \ldots, r\}$. We consider three cases and construct U, V that satisfy the desired properties in the subsequent three subsections.

$$\text{Case 1}: \quad \sum_{i=s+1}^{r} \|\hat{y}_i\|^2 \geq \frac{2}{3}\|\hat{Y}\|_F^2, \quad \sum_{i=s+1}^{r} \|x_i\|^2 \geq \frac{2}{3}\|X\|_F^2. \tag{A.59a}$$

$$\text{Case 2a}: \quad \sum_{i=1}^{s} \|\hat{y}_i\|^2 > \frac{1}{3}\|\hat{Y}\|_F^2. \tag{A.59b}$$

$$\text{Case 2b}: \quad \sum_{i=1}^{s} \|x_i\|^2 > \frac{1}{3}\|X\|_F^2. \tag{A.59c}$$

Proof of Case 1

Without loss of generality, assume

$$\|x_{s+1}\| \leq \|x_{s+2}\| \leq \cdots \leq \|x_r\|. \tag{A.60}$$

Let K be the smallest integer in $\{s+1, s+2, \ldots, r\}$ so that

$$\sum_{i=s+1}^{K} \|\hat{y}_i\|^2 \geq 2 \sum_{j=K+1}^{r} \|\hat{y}_j\|^2. \tag{A.61}$$

By this definition of K, we have

$$\sum_{i=s+1}^{K-1} \|\hat{y}_i\|^2 < 2 \sum_{j=K}^{r} \|\hat{y}_j\|^2. \tag{A.62}$$

Table A.1: Operation 1

Operation 1: Shrinking and Extending
Input: $x_k, \hat{y}_k, k = 1, \ldots, r$.
Output: $u_k, v_k, k = 1, \ldots, r$.
Procedure:
(i) For each $j \leq s$, keep $x_j, \hat{y}_j$ unchanged, i.e.
$$u_j \triangleq x_j,\ v_j \triangleq \hat{y}_j,\ \ j = 1, \ldots, s. \tag{A.63}$$
(ii) For each $i \in \{s+1, \ldots, K\}$, extend x_i to obtain u_i and shrink $\hat{y}_i$ to obtain v_i. For each $i \geq K+1$, shrink x_i to obtain u_i and extend $\hat{y}_i$ to obtain v_i. More specifically,
$$u_i \triangleq \frac{x_i}{1-\epsilon_i}, v_i \triangleq \hat{y}_i(1-\epsilon_i),\ \text{where } \epsilon_i = \begin{cases} 7\bar{\eta} & i \leq K, \\ -4.5\bar{\eta} & i \geq K+1, \end{cases} \quad i = s+1, s+2, \ldots, r, \tag{A.64}$$
in which
$$\bar{\eta} \triangleq \frac{d}{\Sigma_{\min}} \geq \eta. \tag{A.65}$$

We will shrink and extend $x_i, \hat{y}_i$ to obtain U, V. The precise definition of $U = (u_1, u_2, \ldots, u_r)^T, V = (v_1, \ldots, v_r)^T$ is given in Table A.1.

We will show that such U, V satisfy the requirements (A.47). The requirement (A.47a) follows directly from the definition of U, V and the fact $X\hat{Y}^T = \Sigma$.

We then prove the requirement (A.47c). We can bound $\|U - X\|_F$ as

$$\begin{aligned} \|U - X\|_F &= \sqrt{\sum_{i>s} \|\frac{1}{1-\epsilon_i} x_i - x_i\|^2} = \sqrt{\sum_{i>s} \left(\frac{\epsilon_i}{1-\epsilon_i}\right)^2 \|x_i\|^2} \\ &\leq \frac{7\bar{\eta}}{1-7\bar{\eta}} \sqrt{\sum_{i>s} \|x_i\|^2} \leq \frac{7\bar{\eta}}{1-7\bar{\eta}} \|X\|_F \leq \frac{15}{2} \bar{\eta} \beta_T. \end{aligned} \tag{A.66}$$

The bound of $\|V - \hat{Y}\|_F$ is given as

$$\|V - \hat{Y}\|_F = \sqrt{\sum_{i>s} \|(1-\epsilon_i)\hat{y}_i - \hat{y}_i\|^2} \leq \sqrt{\sum_{i>s} \epsilon_i^2 \|\hat{y}_i\|^2} \leq 7\bar{\eta} \|\hat{Y}\|_F.$$

Combining with the bound (A.55), we can bound $\|V - Y\|_F$ as

$$\begin{aligned}\|V - Y\|_F \le \|V - \hat{Y}\|_F + \|\hat{Y} - Y\|_F &\le 7\bar{\eta}\|\hat{Y}\|_F + \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F \\ &= 8\bar{\eta}\|\hat{Y}\|_F \overset{\text{(A.54a)}}{\le} \frac{8\bar{\eta}}{1-\bar{\eta}}\|Y\|_F \le \frac{58}{7}\bar{\eta}\beta_T.\end{aligned} \tag{A.67}$$

The first part of the requirement (A.47c) now follows by multiplying (A.66) and (A.67), and the second part of the requirement (A.47c) follows directly from (A.66) and (A.67).

At last, we prove that U, V satisfy the requirement (A.47b). Let

$$S_1 \triangleq \sum_{i=s+1}^{K} \|\hat{y}_i\|^2,\ S_2 \triangleq \sum_{j=K+1}^{r} \|\hat{y}_j\|^2,\ S_3 \triangleq \sum_{k=1}^{s} \|\hat{y}_k\|^2,$$

then (A.61) and (A.59a) imply

$$S_2 \le S_1/2,\ \ S_3 \le (S_1 + S_2)/2 \le 3S_1/4. \tag{A.68}$$

Since $\bar{\eta} = d/\Sigma_{\min} \ge \eta$, we have $(1-\eta)^2(1-\bar{\eta})^2 \ge (1-2\eta)(1-2\bar{\eta}) \ge (1-2\bar{\eta})^2$. Then

$$\begin{aligned}
&(1-\eta)^2(1-\bar{\eta})^2\|\hat{Y}\|_F^2 - \|V\|_F^2 \\
&\ge (1-2\bar{\eta})^2\|\hat{Y}\|_F^2 - \|V\|_F^2 \\
&= \sum_{i\ge s+1}((1-2\bar{\eta})^2\|\hat{y}_i\|^2 - \|v_i\|^2) + \sum_{k\le s}((1-2\bar{\eta})^2\|\hat{y}_k\|^2 - \|v_k\|^2) \\
&= \sum_{i\ge s+1}((1-2\bar{\eta})^2\|\hat{y}_i\|^2 - (1-\epsilon_i)^2\|\hat{y}_i\|^2) + \sum_{k\le s}((1-2\bar{\eta})^2\|\hat{y}_k\|^2 - \|\hat{y}_k\|^2) \\
&= \sum_{i\ge s+1}(\epsilon_i - 2\bar{\eta})(2-\epsilon_i-2\bar{\eta})\|\hat{y}_i\|^2 - \sum_{k\le s}4\bar{\eta}(1-\bar{\eta})\|\hat{y}_k\|^2 \\
&\overset{\text{(A.64)}}{=} \sum_{i=s+1}^{K} 5\bar{\eta}(2-5\bar{\eta}-2\bar{\eta})\|\hat{y}_i\|^2 - \sum_{j=K+1}^{r} 6.5\bar{\eta}(2+4.5\bar{\eta}-\bar{\eta})\|\hat{y}_j\|^2 - \sum_{k\le s}4\bar{\eta}(1-\bar{\eta})\|\hat{y}_k\|^2 \\
&= 5\bar{\eta}(2-7\bar{\eta})S_1 - 6.5\bar{\eta}(2+2.5\bar{\eta})S_2 - 4\bar{\eta}(1-\bar{\eta})S_3 \\
&\overset{\text{(A.68)}}{\ge} 5\bar{\eta}(2-7\bar{\eta})S_1 - 6.5\bar{\eta}(2+2.5\bar{\eta})\frac{1}{2}S_1 - 4\bar{\eta}(1-\bar{\eta})\frac{3}{4}S_1 \\
&\ge (0.5 - 41\bar{\eta})\bar{\eta}S_1 \ge 0,
\end{aligned} \tag{A.69}$$

where the last inequliaty follows from (A.53). Note that $(1-\eta)\|\hat{Y}\|_F = \|Y\|_F$, thus (A.69) implies

$$\|V\|_F \le (1-\eta)(1-\bar{\eta})\|\hat{Y}\|_F = (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F,$$

which proves the second part of (A.47b).

We then prove the first part of (A.47b), i.e. $\|U\|_F \le \|X\|_F$. Let

$$T_1 \triangleq \sum_{i=s+1}^{K} \|x_i\|^2, \;\; T_2 \triangleq \sum_{j=K}^{r} \|x_j\|^2.$$

We claim that

$$T_2 \ge 2T_1. \tag{A.70}$$

We prove (A.70) by contradiction. Assume the contrary that $T_2 < 2T_1$, then $\frac{1}{3}(T_2+T_1) < T_1$, i.e.

$$\frac{1}{3}\sum_{k=s+1}^{r} \|x_k\|^2 < \sum_{i=s+1}^{K} \|x_i\|^2 \overset{(A.60)}{\le} (K-s)\|x_K\|^2. \tag{A.71}$$

Plugging the second inequality of (A.59a), i.e. $\sum_{k=s+1}^{r} \|x_k\|^2 \ge \frac{2}{3}\|X\|_F^2$, into the above relation, we obtain

$$\|X\|_F^2 \le \frac{9}{2}(K-s)\|x_K\|^2 \le \frac{9}{2}K\|x_K\|^2. \tag{A.72}$$

When $j \in \{K, K+1, \ldots, r\}$, we have

$$\Sigma_{\max} \ge \Sigma_j = \langle x_j, \hat{y}_j \rangle = \|x_j\|\|\hat{y}_j\| \cos(\alpha_j) \overset{(A.60),(A.58)}{\ge} \|x_K\|\|\hat{y}_j\| \cos(3\pi/8).$$

which implies

$$\|\hat{y}_j\| \le \omega, \text{ where } \omega \triangleq \frac{1}{\cos(3\pi/8)} \frac{\Sigma_{\max}}{\|x_K\|}, \;\; j = K, K+1, \ldots, s. \tag{A.73}$$

Therefore,

$$\|\hat{Y}\|_F^2 \overset{(A.59a)}{\le} \frac{3}{2}\sum_{j=s+1}^{r} \|\hat{y}_j\|^2 \overset{(A.62)}{\le} \frac{9}{2}\sum_{j=K}^{r} \|\hat{y}_j\|^2 \overset{(A.73)}{\le} \frac{9}{2}(r-K+1)\omega^2. \tag{A.74}$$

Combining (A.72) and (A.74), and using $K(r-K+1) \le \frac{1}{4}(r+1)^2 \le r^2$, we get

$$\|X\|_F^2\|\hat{Y}\|_F^2 \le \frac{81}{4}r^2\|x_K\|^2\omega^2 \overset{(A.73)}{=} \frac{81}{4}r^2\|x_K\|^2 \frac{1}{\cos(3\pi/8)^2}\frac{\Sigma_{\max}^2}{\|x_K\|^2} < 140r^2\Sigma_{\max}^2. \tag{A.75}$$

According to (A.46), we have $\|X\|_F^2\|\hat{Y}\|_F^2 \ge \|X\|_F^2\|Y\|_F^2 \ge (\frac{3}{5})^2\beta_T^4 = \frac{9}{25}C_T^2 r^2 \Sigma_{\max}^2$; combining with (A.75), we get $140 > \frac{9}{25}C_T^2$, which implies $C_T^2 < 389$. This contradicts the definition (A.52) that $C_T = 20$, thus (A.70) is proved.

Now we are ready to prove the first part of (A.47b) as follows:

$$\begin{aligned}
\|X\|_F^2 - \|U\|_F^2 &= \sum_{i\geq s+1}(\|x_i\|^2 - \|u_i\|^2) + \sum_{k\leq s}(\|x_k\|^2 - \|u_k\|^2) \\
&= \sum_{i\geq s+1}(\|x_i\|^2 - \frac{1}{(1-\epsilon_i)^2}\|x_i\|^2) + 0 \\
&= \sum_{i\geq s+1}\frac{\epsilon_i(\epsilon_i-2)}{(1-\epsilon_i)^2}\|x_i\|^2 \\
&= \sum_{K<j\leq r}\frac{4.5\bar{\eta}(4.5\bar{\eta}+2)}{(1+4.5\bar{\eta})^2}\|x_j\|^2 - \sum_{s+1\leq i\leq K}\frac{7\bar{\eta}(2-7\bar{\eta})}{(1-7\bar{\eta})^2}\|x_i\|^2 \\
&\overset{(A.70)}{\geq} T_2\bar{\eta}\left[\frac{4.5(4.5\bar{\eta}+2)}{(1+4.5\bar{\eta})^2} - \frac{1}{2}\frac{7(2-7\bar{\eta})}{(1-7\bar{\eta})^2}\right] \\
&\geq T_2\bar{\eta}\left[\frac{9}{(1+4.5\bar{\eta})^2} - \frac{7}{(1-7\bar{\eta})^2}\right] \geq 0,
\end{aligned}$$

where the last inequality is because $\frac{(1-7\bar{\eta})^2}{(1+4.5\bar{\eta})^2} > 0.79 > \frac{7}{9}$ when $\bar{\eta} \leq 1/(108r) < 1/100$. Thus the first part of (A.47b) is proved.

Proof of Case 2a

Denote

$$X^0 = X, Y^0 = \hat{Y}, x_k^0 = x_k, y_k^0 = \hat{y}_k, \alpha_k^0 = \alpha_k,\ k = 1, \ldots, r. \tag{A.76}$$

We will define $X^i = (x_1^i, \ldots, x_r^i)^T, Y^i = (y_1^i, \ldots, y_r^i)^T$ recursively. In specific, at the i-th iteration, we will adjust X^{i-1}, Y^{i-1} to X^i, Y^i so that $\|X^i\|_F \leq \|X^{i-1}\|_F, \|Y^i\|_F < \|Y^{i-1}\|_F$ while keeping the first requirement satisfied, i.e. $X^i(Y^i)^T = \Sigma$. The angle α_k^i is defined accordingly, i.e. $\alpha_k^i \triangleq \langle x_k^i, y_k^i\rangle$.

To adjust X^{i-1}, Y^{i-1} to X^i, Y^i, we will define an operation that consists of rotation and shrinking. The basic idea is the following: since the angle between x_i^{i-1} and y_i^{i-1} is large, we can rotate x_i^{i-1} to x_i^i and shrink y_i^{i-1} to y_i^i to keep the inner product invariant, i.e. $\langle x_i^{i-1}, y_i^{i-1}\rangle = \langle x_i^i, y_i^i\rangle$. However, rotating x_i^{i-1} may destroy the orthogonal relationship between x_i^{i-1} and $y_j^{i-1}, \forall j \neq i$, thus we further rotate and shrink y_j^{i-1} to y_j^i for all $j \neq i$ so that y_j^i is orthogonal to the new vector x_i^i. Fortunately, we can prove that using such an operation we still have $\langle x_j^{i-1}, y_j^i\rangle = \Sigma_j, \forall j \neq i$.

A complete description of this operation is given in Table A.2. Without loss of generality, we can make the assumption (A.77). In fact, if (A.77) does not hold, we can

switch i and $m_i \triangleq \arg\min_{k\in\{i,i+1,\ldots,s\}} \alpha_k^{i-1}$ and then apply Operation 2.

Table A.2: Operation 2 that defines X^i, Y^i, where $i \in \{1,\ldots,s\}$

Operation 2: Rotation and Shrinking
Input: $x_k^{i-1}, y_k^{i-1}, \alpha_k^{i-1} \triangleq \angle(x_k^{i-1}, y_k^{i-1}), k = 1,\ldots,r$ and D_i.
Output: $x_k^i, y_k^i, k = 1,\ldots,r$ and $\alpha_k^i \triangleq \angle(x_k^i, y_k^i)$.
Procedure:
(1) Rotate x_i^{i-1} in $\text{span}\{x_i^{i-1}, y_i^{i-1}\}$ to get x_i^i, such that $\langle x_i^i, y_i^{i-1}\rangle = \Sigma_i + D_i$.
(2) Shrink y_i^{i-1} to get y_i^i such that $\langle x_i^i, y_i^i\rangle = \Sigma_i$.
(3) For all $j \neq i$, find y_j^i in $\text{span}\{y_j^{i-1}, y_i^{i-1}\} = \text{span}_{k\neq i,j}\{x_k^{i-1}\}^\perp$ such that $y_j^i \perp x_i^i$ and $\langle x_j^{i-1}, y_j^i\rangle = \langle x_j^{i-1}, y_j^{i-1}\rangle$.
(4) Define $x_j^i \triangleq x_j^{i-1}, \forall j \neq i$.

We will prove that Operation 2 is valid (for D_i that is small enough), i.e. X^i, Y^i defined in Operation 2 indeed exist. The properties of X^i, Y^i obtained by Operation 2 are summarized in the following claim, which will be proved in Appendix A.4.4.

Claim A.4.2 *Consider $i \in \{1,2,\ldots,s\}$. Suppose*

$$\alpha_i^{i-1} \leq \alpha_j^{i-1},\ \forall\ j \in \{i+1, i+2,\ldots,s\}, \tag{A.77}$$

and $D_i > 0$ satisfies

$$\frac{D_i}{\Sigma_i} \leq \frac{1}{12r}, \tag{A.78}$$

then $X^i = (x_1^i,\ldots,x_r^i)^T, Y^i = (y_1^i,\ldots,y_r^i)^T$ described in Operation 2 exist and satisfy the following properties:

$$X^i(Y^i)^T = \Sigma, \tag{A.79a}$$

$$\|x_k^i\| = \|x_k^{i-1}\|, \forall k,\ \|Y^i - Y^{i-1}\|_F^2 \leq \frac{4}{5}\frac{D_i}{\Sigma_i}(\|Y^{i-1}\|_F^2 - \|Y^i\|_F^2), \tag{A.79b}$$

$$\|X^i - X^{i-1}\|_F = \|x_i^i - x_i^{i-1}\| \leq \frac{1}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|x_i^{i-1}\|,\ \ \|Y^i - Y^{i-1}\|_F \leq \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|Y^{i-1}\|_F, \tag{A.79c}$$

$$\alpha_l^i \geq \alpha_l^{i-1} - \frac{1}{r}\frac{\pi}{24} \geq \frac{1}{3}\pi,\ l = i, i+1,\ldots,s. \tag{A.79d}$$

$$\|y_k^{i-1}\| \geq \|y_k^i\| \geq \|y_k^{i-1}\| - \frac{1}{10r}\|y_k^{i-1}\|,\ k = 1,2,\ldots,s. \tag{A.79e}$$

$$\|Y^{i-1}\|_F^2 - \|Y^i\|_F^2 \geq \frac{5}{3}\frac{D_i}{\Sigma_i}\|y_i^i\|^2. \tag{A.79f}$$

We continue to prove Proposition A.4.1 using Claim A.4.2. Given any $D_1, \ldots, D_s$ that satisfy (A.78), we can apply a sequence of Operation 2 for $i = 1, 2, \ldots, s$ to define two sequences of matrices $Y^1, \ldots, Y^s$ and $X^1, \ldots, X^s$. Since $Y^1, \ldots, Y^s$ depend on $D_1, \ldots, D_s$, thus we can use $Y^s(D_1, \ldots, D_s)$ to denote the obtained Y^s by applying Operation 2 for $D_1, \ldots, D_s$. Obviously $Y^s(0, \ldots, 0) = Y^0$. We can also view $\|Y^s\|_F^2$ as a function of $D_1, \ldots, D_s$, denoted as

$$f(D_1, \ldots, D_s) \triangleq \|Y^s(D_1, \ldots, D_s)\|_F^2. \tag{A.80}$$

It can be easily seen that f is a continuous function with respect to $D_1, \ldots, D_s$.

Define[1]

$$\bar{\eta} \triangleq \frac{d}{\Sigma_{\min}} \overset{\text{(A.54a)}}{\geq} \eta, \quad \bar{D}_i \triangleq 9\bar{\eta}\Sigma_i,\ i = 1, \ldots, s. \tag{A.81}$$

We prove that

$$f(\bar{D}_1, \ldots, \bar{D}_s) \leq (1 - 4\bar{\eta})\|\hat{Y}\|_F^2. \tag{A.82}$$

Suppose $\bar{X}^i, \bar{Y}^i, i = 1, \ldots, s$ are recursively defined by Operation 2 for the choices of $D_i = \bar{D}_i$ and denote $\bar{X}^0 = X, \bar{Y}^0 = \hat{Y}$. Since

$$\bar{\eta} = d/\Sigma_{\min} \overset{\text{(A.53)}}{\leq} 1/(108r),$$

we know that $D_i = \bar{D}_i, i = 1, \ldots, s$ as defined in (A.81) satisfy the condition (A.78), thus the property (A.79) holds for $\bar{X}^i, \bar{Y}^i$. Suppose the k-th row of $\bar{Y}^i$ is $(\bar{y}_k^i)^T$, $k = 1, \ldots, r$. By (A.79f) and the fact $\hat{Y} = \bar{Y}^0$, we have

$$\|\hat{Y}\|_F^2 - f(\bar{D}_1, \ldots, \bar{D}_s) = \|\bar{Y}^0\|_F^2 - \|\bar{Y}^s\|_F^2 = \sum_{i=1}^{s}(\|\bar{Y}^{i-1}\|_F^2 - \|\bar{Y}^i\|_F^2) \geq \sum_{i=1}^{s} \frac{5}{3}\frac{\bar{D}_i}{\Sigma_i}\|\bar{y}_i^i\|^2. \tag{A.83}$$

We can bound $\|\bar{y}_i^i\|$ according to (A.79e) as

$$\|\bar{y}_i^i\| \geq \|\bar{y}_i^{i-1}\| - \frac{1}{10r}\|\bar{y}_i^{i-1}\| \geq \|\bar{y}_i^{i-1}\| - \frac{1}{10r}\|\bar{y}_i^0\| \geq \cdots \geq \|\bar{y}_i^0\| - \frac{i}{10r}\|\bar{y}_i^0\| \geq \frac{9}{10}\|\bar{y}_i^0\|.$$

[1] In the first version of the paper, we define $\bar{D}_i \triangleq \frac{9}{2}\eta\Sigma_i \leq \frac{9}{2}\bar{\eta}\Sigma_i \leq 9\frac{d}{\Sigma_{\min}}\Sigma_i$, which is enough for proving Theorem 4.2.1. Here we use a slightly different definition of $\bar{D}_i$ for the purpose of proving Theorem 4.2.2 (linear convergence of the algorithm.)

Plugging into (A.83), we get

$$\|\hat{Y}\|_F^2 - f(\bar{D}_1, \ldots, \bar{D}_s) \geq \sum_{i=1}^{s} \frac{5}{3} \frac{\bar{D}_i}{\Sigma_i} (\frac{9}{10})^2 \|\bar{y}_i^0\|^2$$

$$\overset{\text{(A.81)}}{=} 15 \frac{81}{100} \bar{\eta} \sum_{i=1}^{s} \|\hat{y}_i\|^2 \overset{\text{(A.59b)}}{>} 12 \bar{\eta} \frac{1}{3} \|\hat{Y}\|_F^2 = 4\bar{\eta} \|\hat{Y}\|_F^2,$$

which immediately leads to (A.82).

Combining (A.82) and the fact $f(0, \ldots, 0) = \|Y^0\|_F^2 = \|\hat{Y}\|_F^2$, we have

$$f(0, \ldots, 0) = \|\hat{Y}\|_F^2 > (1 - 4\bar{\eta}) \|\hat{Y}\|_F^2 = f(\bar{D}_1, \ldots, \bar{D}_s).$$

Since f is continuous (in the proof of Claim A.4.2 in Appendix A.4.4, all new vectors depend continuously on D_i), and notice that $1 - 4\bar{\eta} < (1 - \bar{\eta})^4 \leq (1 - \bar{\eta})^2 (1 - \eta)^2 \leq 1$, there must exist

$$0 \leq D_i \leq \bar{D}_i = 9\bar{\eta}\Sigma_i, \;\; i = 1, \ldots, s \tag{A.84}$$

such that

$$f(D_1, \ldots, D_s) = (1 - \bar{\eta})^2 (1 - \eta)^2 \|\hat{Y}\|_F^2. \tag{A.85}$$

Suppose $X^i, Y^i, i = 1, \ldots, s$ are recursively defined by Operation 2 for these choices of D_i, where Y^s is the simplified notation for $Y^s(D_1, \ldots, D_s)$. Define

$$V \triangleq Y^s, \;\; U \triangleq X^s, \tag{A.86}$$

By this definition of V and (A.80), the relation (A.85) can be rewritten as

$$\|V\|_F^2 = (1 - \bar{\eta})^2 (1 - \eta)^2 \|\hat{Y}\|_F^2. \tag{A.87}$$

We show that U, V defined by (A.86) satisfy the requirements (A.47). The requirement (A.47a) follows by the property (A.79a) for $i = s$. The requirement (A.47b) is proved as follows. Combining (A.87) with (A.54a) leads to

$$\|V\|_F = (1 - \bar{\eta})(1 - \eta) \|\hat{Y}\|_F = (1 - \bar{\eta}) \|Y\|_F = (1 - \frac{d}{\Sigma_{\min}}) \|Y\|_F. \tag{A.88}$$

According to the property (A.79b), we have $\|X^i\|_F = \|X^{i-1}\|_F, i = 1, \ldots, s$. Thus $\|X^s\|_F = \|X^{s-1}\|_F = \cdots = \|X^0\|_F = \|X\|_F$, which implies

$$\|U\|_F = \|X\|_F. \tag{A.89}$$

Combining (A.89) and (A.88) leads to the requirement (A.47b) .

It remains to show that U, V satisfy the requirement (A.47c). By the property (A.79b), we have $\|x_k^{i-1}\| = \|x_k^i\|, \forall 1 \le k \le r, 1 \le i \le s$, which implies

$$\|x_k^i\| = \|x_k^0\| = \|x_k\|, \ \ \forall 1 \le k \le r, 1 \le i \le s. \tag{A.90}$$

Note that X^i differs from X^{i-1} only in the i-th row (according to (A.79c)), thus

$$\begin{aligned} \|U - X\|_F = \|X^s - X^0\|_F = \sqrt{\sum_{i=1}^{s} \|x_i^i - x_i^{i-1}\|^2} &\overset{\text{(A.79c)}}{\le} \frac{1}{\sqrt{3}} \frac{D_i}{\Sigma_i} \sqrt{\sum_{i=1}^{s} \|x_i^{i-1}\|^2} \\ &\overset{\text{(A.90)}}{=} \frac{1}{\sqrt{3}} \frac{D_i}{\Sigma_i} \sqrt{\sum_{i=1}^{s} \|x_i\|^2} \le \frac{1}{\sqrt{3}} \frac{D_i}{\Sigma_i} \|X\|_F \overset{\text{(A.84)}}{\le} 3\sqrt{3}\bar{\eta}\|X\|_F. \end{aligned} \tag{A.91}$$

Plugging $\bar{\eta} = d/\Sigma_{\min}$ and $\|X\|_F \le \beta_T$ into the above inequality, we get

$$\|U - X\|_F \le 3\sqrt{3}\frac{\beta_T}{\Sigma_{\min}} d. \tag{A.92}$$

We then bound $\|V - \hat{Y}\|_F^2$ as

$$\begin{aligned} \|V - \hat{Y}\|_F^2 &= \|Y^s - Y^0\|_F^2 \\ &\le s \sum_{i=1}^{s} \|Y^i - Y^{i-1}\|_F^2 \overset{\text{(A.79b)}}{\le} s\frac{4}{5}\frac{D_i}{\Sigma_i} \sum_{i=1}^{s} (\|Y^{i-1}\|_F^2 - \|Y^i\|_F^2) \\ &= s\frac{4}{5}\frac{D_i}{\Sigma_i}(\|Y^0\|_F^2 - \|Y^s\|_F^2) = s\frac{4}{5}\frac{D_i}{\Sigma_i}(\|\hat{Y}\|_F^2 - \|V\|_F^2) \\ &\overset{\text{(A.84)}}{\le} \frac{36}{5} s\bar{\eta}(\|\hat{Y}\|_F^2 - \|V\|_F^2) \overset{\text{(A.87)}}{\le} \frac{36}{5} s\bar{\eta}(2\eta + 2\bar{\eta})\|\hat{Y}\|_F^2 \le \frac{144}{5} r\bar{\eta}^2 \|\hat{Y}\|_F^2, \end{aligned}$$

which leads to

$$\|V - \hat{Y}\|_F \le \frac{12}{\sqrt{5}}\bar{\eta}\sqrt{r}\|\hat{Y}\|_F. \tag{A.93}$$

Then we can bound $\|V - Y\|_F$ as

$$\begin{aligned} \|V - Y\|_F &\le \|V - \hat{Y}\|_F + \|Y - \hat{Y}\| \\ &\overset{\text{(A.93),(A.55)}}{\le} \frac{12}{\sqrt{5}}\bar{\eta}\sqrt{r}\|\hat{Y}\|_F + \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F \\ &= (\frac{12}{\sqrt{5}} + 1)\frac{d}{\Sigma_{\min}}\sqrt{r}\|\hat{Y}\|_F \\ &\overset{\text{(A.54a)}}{=} (\frac{12}{\sqrt{5}} + 1)\frac{d}{\Sigma_{\min}}\sqrt{r}\|Y\|_F \frac{1}{1-\eta} < \frac{13d}{2\Sigma_{\min}}\sqrt{r}\|Y\|_F \le \frac{13\beta_T}{2\Sigma_{\min}}\sqrt{r}d, \end{aligned} \tag{A.94}$$

where the second last inequality is due to $(\frac{12}{\sqrt{5}}+1)/(1-\eta) \overset{(A.53)}{\leq} (\frac{12}{\sqrt{5}}+1)/(1-\frac{1}{108}) < 6.5$. The first part of the requirement (A.47c) now follows by multiplying (A.92) and (A.94), and the second part of the requirement (A.47c) follows directly from (A.92) and (A.94).

Proof of Case 2b

Similar to Case 2a, denote

$$X^0 = X, Y^0 = \hat{Y}, x_k^0 = x_k, y_k^0 = \hat{y}_k, \alpha_k^0 = \alpha_k.$$

By a symmetric argument to that for Case 2a (switch the role of $U, X^j, j = 0, \ldots, s$ and $V, Y^j, j = 0, \ldots, s$), we can prove that there exist $\bar{U}, \bar{V}$ that satisfy properties analogous to (A.47a), (A.88), (A.89), (A.91) and (A.93), i.e.

$$\bar{U}\bar{V}^T = \Sigma, \tag{A.95a}$$

$$\|\bar{U}\|_F = (1-\eta)(1-\bar{\eta})\|X^0\|_F, \ \|\bar{V}\|_F = \|Y^0\|_F, \tag{A.95b}$$

$$\|\bar{V} - Y^0\|_F \leq 3\sqrt{3}\bar{\eta}\|Y^0\|_F, \quad \|\bar{U} - X^0\|_F \leq \frac{12}{\sqrt{5}}\bar{\eta}\sqrt{r}\|X^0\|_F. \tag{A.95c}$$

We will show that the following U, V satisfy the requirements (A.47):

$$U \triangleq \frac{\bar{U}}{(1-\eta)(1-\bar{\eta})}, \ V \triangleq \bar{V}(1-\eta)(1-\bar{\eta}). \tag{A.96}$$

The requirement (A.47a) follows directly from (A.95a) and (A.96). According to (A.95b), (A.96) and the facts $X^0 = X$, $\|Y^0\|_F = \|\hat{Y}\|_F = \|Y\|_F/(1-\eta)$, we have $\|U\|_F = \frac{\|\bar{U}\|_F}{(1-\eta)(1-\bar{\eta})} = \|X^0\|_F = \|X\|_F$, $\|V\|_F = \|\bar{V}\|_F(1-\eta)(1-\bar{\eta}) = \|Y^0\|_F(1-\eta)(1-\bar{\eta}) = \|Y\|_F(1-\bar{\eta})$, thus the requirement (A.47b) is proved.

It remains to prove the requirement (A.47c). We bound $\|U - X\|_F$ as

$$\begin{aligned} \|U - X\|_F &\leq \|U - \bar{U}\|_F + \|\bar{U} - X\|_F \overset{(A.96)}{\leq} 2\bar{\eta}\|U\|_F + \|\bar{U} - X^0\|_F \\ &\overset{(A.47b),(A.95c)}{\leq} 2\bar{\eta}\|X\|_F + \frac{12}{\sqrt{5}}\bar{\eta}\sqrt{r}\|X^0\|_F \leq \frac{15}{2}\bar{\eta}\sqrt{r}\|X\|_F \leq \frac{15}{2}\frac{\beta_T}{\Sigma_{\min}}\sqrt{r}d. \end{aligned} \tag{A.97}$$

Using the fact $\hat{Y} = Y^0$, we bound $\|V - Y\|_F$ as

$$\begin{aligned}
\|V - Y\|_F &\le \|V - \bar{V}\|_F + \|\bar{V} - \hat{Y}\|_F + \|\hat{Y} - Y\|_F \\
&\overset{(A.96),(A.55)}{\le} 2\bar{\eta}\|\bar{V}\|_F + \|\bar{V} - Y^0\|_F + \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F \\
&\overset{(A.95b),(A.95c)}{\le} 2\bar{\eta}\|\hat{Y}\|_F + 3\sqrt{3}\bar{\eta}\|Y^0\|_F + \frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F \qquad \text{(A.98)} \\
&= (3 + 3\sqrt{3})\frac{d}{\Sigma_{\min}}\|\hat{Y}\|_F \\
&\overset{(A.54a)}{=} \frac{3 + 3\sqrt{3}}{1 - \eta}\frac{d}{\Sigma_{\min}}\|Y\|_F \le \frac{58\beta_T}{7\Sigma_{\min}}d.
\end{aligned}$$

The first part of the requirement (A.47c) now follows by multiplying (A.97) and (A.98), and the second part follows directly from (A.97) and (A.98).

A.4.4 Proof of Claim (A.4.2)

Suppose Claim (A.4.2) holds for $1, 2, \ldots, i-1$, we prove Claim (A.4.2) for i. By the property (A.79a) and (A.79d) of Claim A.4.2 for $i-1$, we have

$$X^{i-1}(Y^{i-1})^T = \Sigma. \qquad \text{(A.99a)}$$

$$\alpha_i^{i-1} \ge \alpha_i^{[0]} - \frac{i-1}{r}\frac{1}{24}\pi \ge \frac{3}{8}\pi - \frac{1}{24}\pi + \frac{1}{24r}\pi = \frac{1}{3}\pi + \frac{1}{24r}\pi \ge \frac{1}{3}\pi. \qquad \text{(A.99b)}$$

To simplify the notations, throughout the proof of Claim A.4.2, we denote X^{i-1}, Y^{i-1} as X, Y and denote X^i, Y^i as X', Y'. The notations $\alpha_k^{i-1}, \alpha_k^i$ are changed accordingly to α_k, α_k'. Then (A.99a) and (A.99b) become

$$XY^T = \Sigma, \qquad \text{(A.100a)}$$

$$\alpha_i \ge \frac{1}{3}\pi + \frac{1}{24r}\pi \ge \frac{1}{3}\pi. \qquad \text{(A.100b)}$$

We need to prove that X', Y' exist and satisfy the properties in Claim (A.4.2), i.e.

(with the simplification of notations)

$$X'(Y')^T = \Sigma. \tag{A.101a}$$

$$\|x'_k\| = \|x_k\|, \forall k,\ \|Y' - Y\|_F^2 \leq \frac{4}{5}\frac{D_i}{\Sigma_i}(\|Y\|_F^2 - \|Y'\|_F^2). \tag{A.101b}$$

$$\|X' - X\|_F = \|x'_i - x_i\| \leq \frac{1}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|x_i\|,\ \|Y' - Y\|_F \leq \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|Y\|_F. \tag{A.101c}$$

$$\alpha'_l \geq \alpha_l - \frac{1}{r}\frac{\pi}{24} \geq \frac{1}{3}\pi,\ l = i, i+1, \ldots, s. \tag{A.101d}$$

$$\|y_k\| \geq \|y'_k\| \geq \|y_k\| - \frac{1}{10r}\|y_k\|,\ \ k = 1, 2, \ldots, s. \tag{A.101e}$$

$$\|Y\|_F^2 - \|Y'\|_F^2 \geq \frac{5}{3}\frac{D_i}{\Sigma_i}\|y_i\|^2. \tag{A.101f}$$

Ideas of the proof of Claim (A.4.2)

Before presenting the formal proof, we briefly describe its idea. The goal of Operation 2 is to reduce the norm of Y while keeping $\langle X, Y\rangle$ and $\|X\|_F$ invariant, by rotating and shrinking $x_i,\ y_k, k = 1, \ldots, K$ (note that $x_j, \forall j \neq i$, do no change). We first rotate x_i and shrink y_i at the same time so that the new inner product $\langle x'_i, y'_i\rangle$ equals the previous one $\langle x_i, y_i\rangle$ (this step can be viewed as a combination of two steps: first rotate x_i to increase the inner product, then shrink y_i to reduce the inner product). In order to preserve the orthogonality of X and Y, we need to rotate $y_j, \forall j \neq i$, so that the new y'_j is orthogonal to x'_i.

Although the above procedure is simple, there are two questions to be answered. The first question is: will the inner product $\langle x_j, y_j\rangle$ increase as we rotate y_j, for all $j \neq i$? If yes, we could first rotate and then shrink y_j to obtain y'_j so that the new inner product $\langle x_j, y'_j\rangle$ equals $\langle x_j, y_j\rangle$, which achieves the goal of Operation 2. By resorting to the geometry (in a rigourous way) we are able to provide an affirmative answer to the above question. To gain an intuition why this is possible, we use Figure A.3 to illustrate. Consider the case $i = 2$ and rotate x_2 towards y_2 to obtain x'_2, then y_1 has to be rotated so that y'_1 is orthogonal to x'_2. It is clear from this figure that the angle between y_1 and x_1 also decreases, or equivalently, the inner product $\langle x_1, y_1\rangle$ also increases. One might ask whether we have utilized additional assumptions on the relative positions of x_i, y_i's. In fact, we do not utilize additional assumptions; what we implicitly utilize is the fact that $\langle x_i, y_i\rangle > 0, \forall i$ (see Figure A.4, Figure A.5 and the paragraph after (A.108) for

detailed explanations).

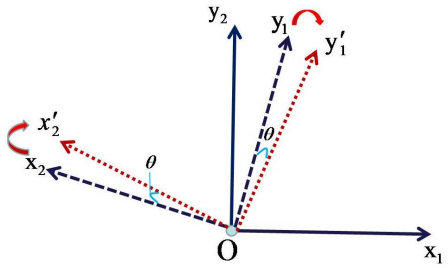


Figure A.3: Rotation of y_1, y_2 increases both $\langle x_i, y_i\rangle, i = 1, 2$

The second question is: will the angle $\alpha'_j = \angle(x_j, y'_j)$ still be larger than, say, $\frac{1}{3}\pi$, for all $j > i$? If yes, then we can apply Operation 2 repeatedly for all $i = 1, 2, \ldots, s$. To provide an affirmative answer, we should guarantee that each angle decreases at most $\frac{1}{s}(\frac{3}{8}\pi - \frac{1}{3}\pi) = \frac{1}{24s}\pi$, i.e. $\angle(x_j, y'_j) \geq \angle(x_j, y_j) - \frac{1}{24s}\pi, \forall\ i < j \leq s$. Unlike the first question which can be answered by reading Figure A.4 and Figure A.5, this question cannot be answered by just reading figures. We make some algebraic computation to obtain the following result: under the assumption that α_i is no less than α_j, during Operation 2 the amount of decrease in α_j is upper bounded by the amount of decrease in α_i, which can be further bounded above by $\frac{1}{24s}\pi$. This result explains why our proof requires the assumption $\alpha_i \geq \alpha_j, \forall\ i < j \leq s$, i.e. (A.77).

Formal proof of Claim (A.4.2)

We first show how to define x'_i and y'_i. Note that

$$\|x_i\|\|y_i\| = \frac{\langle x_i, y_i\rangle}{\cos\alpha_i} \geq \frac{\Sigma_i}{\cos(\frac{\pi}{3})} = 2\Sigma_i. \tag{A.102}$$

Since (A.102) implies $\frac{\Sigma_i + D_i}{\|x_i\|\|y_i\|} \leq \frac{2\Sigma_i}{\|x_i\|\|y_i\|} \leq 1$, we can define

$$\alpha'_i \triangleq \arccos(\frac{\Sigma_i + D_i}{\|x_i\|\|y_i\|}) \in [0, \frac{\pi}{2}].$$

There is a unique x'_i in the plane $\text{span}\{x_i, y_i\}$ which satisfies

$$\|x'_i\| = \|x_i\| \tag{A.103}$$

and $\angle(x'_i, y_i) = \alpha'_i$. By the definition of α'_i above, we have

$$\langle x'_i, y_i\rangle = \Sigma_i + D_i.$$

The existence of x_i' is proved. We define

$$y_i' \triangleq \frac{\Sigma_i}{\Sigma_i + D_i} y_i, \tag{A.104}$$

then

$$\langle x_i', y_i' \rangle = \frac{\Sigma_i}{\Sigma_i + D_i} \langle x_i', y_i \rangle = \Sigma_i. \tag{A.105}$$

The existence of y_i' is also proved.

Since $0 < \langle x_i, y_i \rangle = \Sigma_i < \langle x_i', y_i \rangle$, we have $\frac{\pi}{2} > \alpha_i > \alpha_i' > 0$, thus we can define

$$\theta \triangleq \alpha_i - \alpha_i' = \angle(x_i', x_i) \in (0, \alpha_i). \tag{A.106}$$

Fix any $j \neq i$, we then show how to define y_j'. Define

$$A_i \triangleq \text{span}_{j \neq i}\{x_j\} \perp y_i, \;\; B_i \triangleq \text{span}_{j \neq i}\{y_j\} \perp x_i, \;\; T_i \triangleq A_i \cap B_i.$$

Let $\overrightarrow{OY_j} = y_j$, $K_j \triangleq \mathcal{P}_{A_i}(Y_j), H_j \triangleq \mathcal{P}_{T_i}(Y_j)$. Then $\angle Y_j H_j K_j = \min\{\angle(x_i, y_i), \pi - \angle(x_i, y_i)\} = \angle(x_i, y_i) = \alpha_i$. Since $\alpha_i > \theta$, there exists a unique point Y_j' in the line segment $Y_j K_j$ such that

$$\angle Y_j H_j Y_j' = \theta. \tag{A.107}$$

Since $K_j = \mathcal{P}_{A_i}(Y_j)$ and $x_k \in A_i, \forall k \neq i$, we have $\overrightarrow{Y_j K_j} \perp x_k, \forall k \neq i$, thus

$$\overrightarrow{Y_j Y_j'} \perp x_k, \;\; \forall k \neq i. \tag{A.108}$$

See Figure A.4 and Figure A.5 for the geometrical interpretation; note that T_i in general is not a line but a $r-2$ dimensional space. The righthand side subfigures represents the 2 dimensional subspace $T_i^\perp$; since $\text{span}\{H_j Y_j, H_j K_j\} = T_i^\perp = \text{span}\{x_i, y_i\}$, we can draw x_i, y_i, y_i' as the vectors starting from H_j and lying in the plane $H_j Y_j K_j = T_i^\perp$ in the figures. Figure A.4 and Figure A.5 differ in the relative position of x_i and K_j: x_i and K_j lie in the same side of line $H_j Y_j$ in Figure A.4 but in different sides in Figure A.5. Given the positions of x_i and H_j, Y_j, K_j, the position of y_i is determined since $y_i \perp \overrightarrow{H_j K_j}$ and $\angle(x_i, y_i) < \frac{\pi}{2}$.

In both figures, we have

$$\angle(\overrightarrow{H_j Y_j'}, x_i') = \angle(\overrightarrow{H_j Y_j}, x_i) - \angle(x_i', x_i) + \angle Y_j H_j Y_j' \overset{\text{(A.106),(A.107)}}{=\!=\!=} \frac{\pi}{2} - \theta + \theta = \frac{\pi}{2},$$

$$\Longrightarrow \quad \overrightarrow{H_j Y_j'} \perp x_i'. \tag{A.109}$$

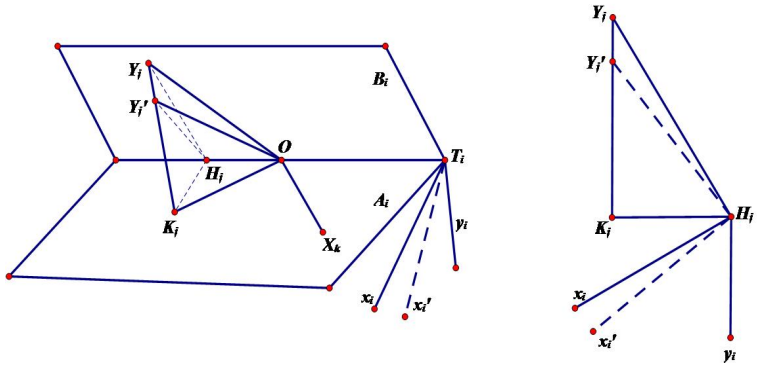


Figure A.4: Left: Space A_i, B_i, T_i, vectors x_i, y_i, x_i', x_k and some points related to y_j. Right: Some points and vectors in plane $H_jY_jK_j = T_i^{\perp} = \text{span}\{x_i, y_i\}$. This figure shows the first possibility: x_i and K_j lie in the same side of line H_jY_j.

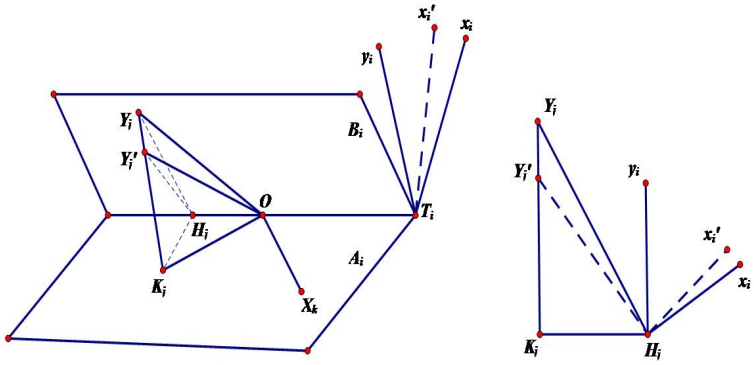


Figure A.5: Same objects as in Figure A.4, but for the second possibility: x_i and K_j lie in different sides of line H_jY_j.

Now we are ready to define y_j' and establish its properties. Define

$$y_j' \triangleq \overrightarrow{OY_j'}. \tag{A.110}$$

Since Y_j' lies in the line segment K_jY_j and $\angle Y_jK_jO = \pi/2$, we have

$$\|y_j'\| \leq \|y_j\|. \tag{A.111}$$

We also have

$$y_j' = y_j + \overrightarrow{Y_jY_j'} \in \text{span}\{y_j, y_i\} \perp x_k,\ \forall k \neq i, j. \tag{A.112}$$

According to the fact $\overrightarrow{OH_j} \perp x_i'$ and (A.109), we have

$$y_j' = \overrightarrow{OH_j} + \overrightarrow{H_jY_j'} \perp x_i'. \tag{A.113}$$

Let $k = j$ in (A.108), we obtain

$$0 = \langle \overrightarrow{Y_j Y_j'}, x_j \rangle = \langle y_j' - y_j, x_j \rangle = 0 \Longrightarrow \langle x_j, y_j' \rangle = \langle x_j, y_j \rangle. \tag{A.114}$$

We have shown that y_j' defined in (A.110) satisfies (A.112), (A.113) and (A.114), thus the existence of y_j' in Operation 2 is proved.

Having defined x_i', y_i' and $y_j', \forall j \neq i$, we further define

$$x_j' \triangleq x_j, \forall j \neq i, \tag{A.115}$$

which completes the definition of X', Y'. In the rest, we prove that X', Y' satisfy the desired property (A.101).

The property (A.101a) can be directly proved by the definitions of X', Y'. In specific, according to (A.105), (A.114) and the definition (A.115), we have $\langle x_k', y_k' \rangle = \Sigma_k, \forall k$. According to the definitions (A.115), (A.104) and the fact $y_i \perp x_j, \forall j \neq i$, we have $y_i' \perp x_j', \forall j \neq i$. Together with (A.112) and (A.113), we obtain $\langle x_k', y_l' \rangle = 0, \forall k \neq l$. Thus $X'(Y')^T = \Sigma$.

Next, we prove the property (A.101d). We first prove

$$\alpha_i' - \alpha_i = \theta \leq \frac{1}{r} \frac{\pi}{24}. \tag{A.116}$$

Define $h_i \triangleq x_i' - x_i$, then

$$\|h_i\| = 2\|x_i\| \sin(\frac{\theta}{2}). \tag{A.117}$$

From $\langle x_i', y_i \rangle = \Sigma_i + D_i = \langle x_i, y_i \rangle + D_i$, we obtain $\langle h_i, y_i \rangle = D_i$. Note that $\langle h_i, y_i \rangle = \|h_i\| \|y_i\| \cos(\angle(h_i, y_i))$ and $\angle(h_i, y_i) = \frac{\pi}{2} - \alpha_i + \frac{\theta}{2}$, thus

$$\|h_i\| = \frac{D_i}{\|y_i\| \sin(\alpha_i - \frac{\theta}{2})}. \tag{A.118}$$

According to (A.117) and (A.118), we have

$$\frac{D_i}{\|x_i\| \|y_i\|} = 2 \sin(\alpha_i - \frac{\theta}{2}) \sin(\frac{\theta}{2}) \geq 2 \sin(\frac{\alpha_i}{2}) \sin(\frac{\theta}{2}) \geq 2 \sin(\frac{\pi}{6}) \sin(\frac{\theta}{2}) = \sin(\frac{\theta}{2}) \geq \frac{\theta}{\pi},$$

where the last equality follows from the fact that $\frac{\sin(t)}{t}$ is decreasing in $t \in (0, \frac{\pi}{2}]$. Note that $\frac{D_i}{\|x_i\| \|y_i\|}$ can be upper bounded as

$$\frac{D_i}{\|x_i\| \|y_i\|} \overset{\text{(A.102)}}{\leq} \frac{D_i}{2\Sigma_i} \overset{\text{(A.78)}}{\leq} \frac{1}{24r}.$$

Combining the above two relations, we get (A.116).

To prove

$$\alpha_j - \alpha'_j \leq \frac{\pi}{24r}, \forall j \in \{i+1, \ldots, s\}, \tag{A.119}$$

we only need to prove

$$\theta_j \triangleq \alpha_j - \alpha'_j \leq \theta, \ \ \forall j \in \{i+1, \ldots, s\} \tag{A.120}$$

and then use (A.116). The equality (A.114) implies that

$$\|x_j\|\|y_j\| \cos(\alpha_j) = \|x_j\|\|y'_j\| \cos(\alpha'_j),$$

which leads to

$$\frac{\cos(\alpha_j)}{\cos(\alpha_j - \theta_j)} = \frac{\cos(\alpha_j)}{\cos(\alpha'_j)} = \frac{\|y'_j\|}{\|y_j\|}.$$

For any two points P_1, P_2, we use $|P_1P_2|$ to denote the length of the line segment P_1P_2. Since $\overrightarrow{OH_j}$ is orthogonal to plane $H_jK_jY_j$, we have

$$\frac{\|y'_j\|^2}{\|y_j\|^2} = \frac{|OH_j|^2 + |H_jY'_j|^2}{|OH_j|^2 + |H_jY_j|^2} \geq \frac{|H_jY'_j|^2}{|H_jY_j|^2},$$

where the last inequality follows from the fact that $|H_jY'_j| \leq |H_jY_j|$. Since $\angle Y_jH_jK_j = \alpha_i, \angle Y'_jH_jK_j = \alpha'_i$ and $\angle Y_jK_jH_j = \frac{\pi}{2}$, we have

$$\frac{|H_jY'_j|}{|H_jY_j|} = \frac{\sin \angle Y'_jY_jH_j}{\sin \angle Y_jY'_jH_j} = \frac{\sin(\pi/2 - \alpha_i)}{\sin(\pi/2 + \alpha'_i)} = \frac{\cos(\alpha_i)}{\cos(\alpha'_i)}.$$

According to the assumption (A.77) and $i < j \leq s$, we have $0 \leq \alpha_i \leq \alpha_j \leq \frac{\pi}{2}$. Since $\cos(x)/\cos(x-\theta)$ is decreasing in $[0, \frac{\pi}{2}]$, we can get

$$\frac{\cos(\alpha_i)}{\cos(\alpha'_i)} = \frac{\cos(\alpha_i)}{\cos(\alpha_i - \theta)} \geq \frac{\cos(\alpha_j)}{\cos(\alpha_j - \theta)}.$$

Combining the above four relations, we get

$$\frac{\cos(\alpha_j)}{\cos(\alpha_j - \theta_j)} \geq \frac{\cos(\alpha_j)}{\cos(\alpha_j - \theta)},$$

which implies $\cos(\alpha_j - \theta) \geq \cos(\alpha_j - \theta_j)$ that immediately leads to (A.120). Thus we have proved (A.119), which combined with (A.116) establishes the property (A.101d).

Then we prove the property (A.101c). Since $x'_j = x_j, \forall j \neq i$, we have $\|X' - X\|_F = \|x'_i - x_i\|$, which can be bounded as

$$\begin{aligned}\|x'_i - x_i\| = \|h_i\| &\overset{(A.118)}{=} \frac{D_i}{\|y_i\| \sin(\alpha_i - \frac{\theta}{2})} \leq \frac{\|x_i\| D_i}{\|x_i\| \|y_i\| \sin(\frac{\pi}{3})} \\ &\overset{(A.102)}{\leq} \frac{\|x_i\| D_i}{2\Sigma_i \sin(\frac{\pi}{3})} < \frac{1}{\sqrt{3}} \frac{\|x_i\|}{\Sigma_i} D_i,\end{aligned}$$

where the first inequality is due to

$$\alpha_i - \theta/2 \geq \alpha_i - \theta \overset{(A.100b)}{\geq} \pi/3 + \pi/24 - \theta \overset{(A.116)}{\geq} \pi/3. \tag{A.121}$$

Thus the first part of (A.101c) is proved.

According to (A.117) and (A.118), we have

$$2\sin(\frac{\theta}{2}) = \frac{D_i}{\|x_i\| \|y_i\| \sin(\alpha_i - \frac{\theta}{2})} \tag{A.122}$$

Now we upper bound $\|y'_j - y_j\|$ as

$$\begin{aligned}\|y'_j - y_j\| = |Y'_j Y_j| &= \frac{\sin(\theta)}{\cos(\alpha_i - \theta)} |H_j Y_j| \\ &= 2\sin(\frac{\theta}{2}) \cos(\frac{\theta}{2}) \frac{1}{\cos(\alpha_i - \theta)} |H_j Y_j| \\ &\overset{(A.122)}{=} \frac{D_i}{\|x_i\| \|y_i\| \sin(\alpha_i - \frac{\theta}{2})} \cos(\frac{\theta}{2}) \frac{1}{\cos(\alpha_i - \theta)} |H_j Y_j| \\ &\leq \frac{D_i}{\|x_i\| \|y_i\| \sin(\alpha_i - \frac{\theta}{2})} \frac{1}{\cos(\alpha_i)} |H_j Y_j| \\ &\overset{(A.121)}{\leq} \frac{D_i}{\sin(\frac{\pi}{3}) \langle x_i, y_i \rangle} |H_j Y_j| \\ &\leq \frac{2}{\sqrt{3}} \frac{D_i}{\Sigma_i} |H_j Y_j|,\end{aligned} \tag{A.123}$$

where the last inequality is due to the fact $\langle x_i, y_i \rangle = \Sigma_i$. Using $|H_j Y_j| \leq \|y_j\|$, we obtain

$$\|y'_j - y_j\| \leq \frac{2}{\sqrt{3}} \frac{D_i}{\Sigma_i} \|y_j\|. \tag{A.124}$$

According to the definition (A.104), we have

$$\|y_i - y'_i\| = (1 - \frac{\Sigma_i}{\Sigma_i + D_i}) \|y_i\| = \frac{D_i}{\Sigma_i + D_i} \|y_i\| \leq \frac{D_i}{\Sigma_i} \|y_i\|. \tag{A.125}$$

According to (A.124) (which holds for any $j \in \{1,\ldots,r\}\backslash\{i\}$) and (A.125), we get

$$\|Y - Y'\|_F = \sqrt{\sum_{k=1}^{r} \|y_k - y'_k\|^2} \le \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\sqrt{\sum_{k=1}^{r} \|y_k\|^2} = \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|Y\|_F,$$

which proves the second part of (A.101c).

The property (A.101e) can be proved as follows. By the definition (A.104), we have $\|y'_i\| \le \|y_i\|$, which combined with (A.111) (for all $j \ne i$) leads to

$$\|y'_k\| \le \|y_k\|, \;\; k = 1,\ldots,s.$$

According to (A.124) (for all $j \ne i$) and (A.125), we have $\|y'_k - y_k\| \le \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|y_k\|, \forall k$, which implies

$$\|y'_k\| \ge \|y_k\| - \|y'_k - y_k\| \ge \|y_k\| - \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\|y_k\| \overset{(A.78)}{\ge} \|y_k\| - \frac{1}{10r}\|y_k\|, \; \forall k.$$

Combining the above two relations we obtain the property (A.101e).

The property (A.101f) can be easily proved by (A.104). In fact, we have

$$\begin{aligned} \|y_i\|^2 - \|y'_i\|^2 &= (\|y_i\| - \|y'_i\|)(\|y_i\| + \|y'_i\|) \\ &\ge 2\|y'_i\|(\|y_i\| - \|y'_i\|) \overset{(A.104)}{=} 2\|y'_i\|(\frac{\Sigma_i + D_i}{\Sigma_i} - 1)\|y'_i\| \\ &= 2\frac{D_i}{\Sigma_i}\|y'_i\|^2 \ge 2\frac{D_i}{\Sigma_i}(\frac{11}{12})^2\|y_i\|^2 \ge \frac{5}{3}\frac{D_i}{\Sigma_i}\|y_i\|^2. \end{aligned} \tag{A.126}$$

where the second last inequliaty follows from $\|y'_i\| \ge \|y_i\| - \|y_i - y'_i\| \overset{(A.125)}{\ge} \|y_i\| - D_i\|y_i\|/\Sigma_i \overset{(A.78)}{\ge} 11\|y_i\|/12$. According to (A.111) (for all $j \ne i$), we have $\|Y\|_F^2 - \|Y'\|_F^2 \ge \|y_i\|^2 - \|y'_i\|^2$, which combined with (A.126) leads to the property (A.101f).

At last, we prove the property (A.101b). The first part $\|X'\|_F = \|X\|_F$ follows from (A.103) and (A.115), thus it remains to prove the second part. Denote $\varphi_j \triangleq \angle Y_j O Y'_j, \beta_j \triangleq \angle Y_j O K_j$ as shown in Figure A.6. Pick a point Z_j in the line segment OY_j so that $|OZ_j| = |OY'_j|$, then $|Y_j Z_j| = \|y_j\| - \|y'_j\|$. Thus we have

$$\frac{\|y_j - y'_j\|}{\|y_j\| - \|y'_j\|} = \frac{|Y_j Y'_j|}{|Y_j Z_j|} = \frac{\sin(\angle Y_j Z_j Y'_j)}{\sin(\angle Y_j Y'_j Z_j)} = \frac{\sin(\pi/2 - \varphi_j/2)}{\sin(\beta_j - \varphi_j/2)} \le \frac{1}{\sin(\beta_j - \varphi_j)}. \tag{A.127}$$

Figure A.6: Illustration for the proof of the property (A.101b)

In order to bound $1/\sin(\beta_j - \varphi_j)$ [2] , we use the following bound:

$$\frac{\sin\beta_j}{\sin(\beta_j-\varphi_j)} = \frac{|Y_jK_j|}{\|y_j\|}\frac{\|y'_j\|}{|Y'_jK_j|} \le \frac{|Y_jK_j|}{|Y'_jK_j|} = \frac{\tan\alpha_i}{\tan(\alpha_i-\theta)}.$$

Then we have

$$\frac{\sin\beta_j}{\sin(\beta_j-\varphi_j)}\frac{\sin(\alpha_i-\theta)}{\sin(\alpha_i)} = \frac{\cos(\alpha_i-\theta)}{\cos(\alpha_i)} = \frac{\cos\alpha_i\cos\theta+\sin\alpha_i\sin\theta}{\cos(\alpha_i)} \le \frac{\sin(\theta)}{\cos(\alpha_i)}+1. \tag{A.128}$$

According to (A.122) and the fact $\cos(\alpha_i) = \langle x_i, y_i\rangle/(\|x_i\|\|y_i\|) = \Sigma_i/(\|x_i\|\|y_i\|)$, we have

$$\frac{\sin(\theta)}{\cos(\alpha_i)} \le \frac{2\sin(\theta/2)}{\cos(\alpha_i)} = \frac{D_i}{\|x_i\|\|y_i\|\sin(\alpha_i-\theta/2)}\frac{\|x_i\|\|y_i\|}{\Sigma_i} = \frac{D_i}{\Sigma_i}\frac{1}{\sin(\alpha_i-\theta/2)}$$
$$\overset{(A.78),(A.121)}{\le} \frac{1}{12}\frac{1}{\sin(\pi/3)} = \frac{1}{6\sqrt{3}}.$$

Plugging the above relation into (A.128), we obtain

$$\frac{\sin\beta_j}{\sin(\beta_j-\varphi_j)}\frac{\sin(\alpha_i-\theta)}{\sin(\alpha_i)} \le \frac{6\sqrt{3}+1}{6\sqrt{3}}. \tag{A.129}$$

[2] The part from (A.127) to (A.129) can be replaced by a simpler bound $\sin(\beta_j-\varphi_j) \ge \sin(\beta_j/2) \ge \sin(\beta_j)/2$ and we can still obtain a similar bound as (A.131); however, by using this simpler yet looser bound, the constant coefficient 7/8 will be replaced by a larger constant.

Combining (A.127) and (A.123), we obtain

$$\begin{aligned}\frac{\|y_j - y_j'\|}{\|y_j\| - \|y_j'\|}\frac{\|y_j - y_j'\|}{\|y_j\|} &\le \frac{1}{\sin(\beta_j - \varphi_j)}\frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\frac{|H_jY_j|}{\|y_j\|} \\ &\overset{(A.129)}{\le} \frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i}\frac{6\sqrt{3}+1}{6\sqrt{3}}\frac{|H_jY_j|}{\|y_j\|}\frac{\sin(\alpha_i)}{\sin(\beta_j)}\frac{1}{\sin(\alpha_i - \theta)} \\ &= \frac{6\sqrt{3}+1}{9}\frac{D_i}{\Sigma_i}\frac{1}{\sin(\alpha_i - \theta)} \overset{(A.121)}{\le} \frac{6\sqrt{3}+1}{9}\frac{2}{\sqrt{3}}\frac{D_i}{\Sigma_i} \le \frac{3}{2}\frac{D_i}{\Sigma_i},\end{aligned} \tag{A.130}$$

where the last equality is due to $|H_jY_j|\sin(\alpha_i) = |Y_jK_j| = \|y_j\|\sin(\beta_j)$.

According to (A.124) and (A.78), we obtain that $\|y_j - y_j'\| \le \frac{2}{\sqrt{3}}\frac{1}{12}\|y_j\| \le \frac{1}{8}\|y_j\|$, which further implies $\|y_j'\| + \|y_j\| \ge 2\|y_j\| - \|y_j - y_j'\| \ge \frac{15}{8}\|y_j\|$. Then by (A.130) we have

$$\begin{aligned}\|y_j - y_j'\|^2 &\le \frac{5\sqrt{3}+1}{6}\frac{D_i}{\Sigma_i}(\|y_j\| - \|y_j'\|)\|y_j\| \\ \le \frac{3}{2}\frac{D_i}{\Sigma_i}(\|y_j\| - \|y_j'\|)(\|y_j'\| + \|y_j\|)\frac{8}{15} &= \frac{4}{5}\frac{D_i}{\Sigma_i}(\|y_j\|^2 - \|y_j'\|^2).\end{aligned} \tag{A.131}$$

According to the definition (A.104), we have

$$\frac{\|y_i\|^2 - \|y_i'\|^2}{\|y_i - y_i'\|^2} = \frac{1 - (\Sigma_i)^2/(\Sigma_i + D_i)^2}{[1 - \Sigma_i/(\Sigma_i + D_i)]^2} = \frac{(\Sigma_i + D_i)^2 - \Sigma_i^2}{D_i^2} = \frac{D_i^2 + 2D_i\Sigma_i}{D_i^2} \ge 2\frac{\Sigma_i}{D_i},$$

which implies

$$\|y_i - y_i'\|^2 \le \frac{1}{2}\frac{D_i}{\Sigma_i}(\|y_i\|^2 - \|y_i'\|^2). \tag{A.132}$$

Summing up (A.131) for $j \in \{1, \ldots, r\}\backslash\{i\}$ and (A.132), we obtain

$$\|Y - Y'\|_F^2 \le \frac{4}{5}\frac{D_i}{\Sigma_i}(\|Y\|_F^2 - \|Y'\|_F^2),$$

which proves the second part of (A.101b).

A.5 Proofs of the results in Section 4.4

A.5.1 Proof of Claim 4.4.2

The proof of this claim consists of two parts: first, by a classical result we have that M_0, the best rank-r approximation of $\frac{1}{p}\mathcal{P}_\Omega(M)$, is close to M; second, show that the scaling does not change the closeness.

We first present the following result.

Lemma A.5.1 *Assume M is a rank r matrix of dimension $m \times n$ with $m \geq n$, and denote $M_{\max} = \|M\|_\infty$ as the maximum magnitude of the entries of M. Suppose each entry of M is included in Ω with probability $p \geq C_0 \frac{\log(m+n)}{m}$, and M_0 is the best rank-r approximation of $\frac{1}{p}\mathcal{P}_\Omega(M)$. Then with probability larger than $1 - 1/(2n^4)$,*

$$\frac{1}{mnM_{\max}^2}\|M - M_0\|_F^2 \leq C_2 \frac{\alpha^{\frac{3}{2}} r}{pm}, \tag{A.133}$$

for some numerical constant C_2.

Remark: Lemma A.5.1 can be found in [33]. The original version [33, Theorem 1.1] holds for $M_0 = \mathrm{P}_r(\mathrm{T_r}(\mathcal{P}_\Omega(M))/p)$, where $\mathrm{T_r}(\cdot)$ denotes a trimming operator which sets to zero all rows and columns that have too many observed entries, and $\mathrm{P}_r(\cdot)$ denotes the best rank-r approximation. By standard Chernoff bound one can show that none of the rows and columns have too many observed entries with high probability, thus the conclusion of [33, Theorem 1.1] holds for $M_0 = \mathrm{P}_r(\mathcal{P}_\Omega(M))/p$. The key to establish Lemma A.5.1 is a bound on $\|M - \frac{1}{p}\mathcal{P}_\Omega(M)\|_2$, which can be simply proved by matrix concentration inequalities; see [24, Remark 6.1.2], [2, Theorem 6.3] or [5, Theorem 3.5]. The proof of [33, Theorem 1.1] is more complicated than applying matrix concentration inequalities since it holds for a weaker condition $|\Omega| \geq O(n)$.

Note that $\hat{X}_0, \hat{Y}_0$ defined in Table 4.1 satisfy

$$\hat{X}_0 \hat{Y}_0^{\,T} = \mathrm{P}_r(\mathcal{P}_\Omega(M)/p) = M_0. \tag{A.134}$$

Recall that the SVD of M is $M = \hat{U}\Sigma\hat{V}$, where $\hat{U}, \hat{V}$ satisfies (4.1). We have

$$\begin{aligned} |M_{ij}| &= \sum_{k=1}^{r} |\hat{U}_{ik}\hat{V}_{jk}\Sigma_k| \leq \Sigma_{\max} \sum_{k=1}^{r} |\hat{U}_{ik}\hat{V}_{jk}| \\ &\leq \Sigma_{\max} \sqrt{\sum_{k=1}^{r} \hat{U}_{ik}^2} \sqrt{\sum_{k=1}^{r} \hat{V}_{jk}^2} \overset{(4.1)}{\leq} \Sigma_{\max} \frac{\mu r}{\sqrt{mn}}, \ \forall\ i, j. \end{aligned} \tag{A.135}$$

The above relation implies $M_{\max} \leq \Sigma_{\max} \frac{\mu r}{\sqrt{mn}}$. Plugging this inequality and $p = |\Omega|/(mn)$ into (A.133), we get

$$\|M - M_0\|_F^2 \leq C_2 \frac{mn\alpha^{\frac{3}{2}} r}{pm} \Sigma_{\max}^2 \frac{\mu^2 r^2}{mn} = C_2 n \frac{\alpha^{\frac{3}{2}} r^3 \kappa^2 \mu^2}{|\Omega|} \Sigma_{\min}^2. \tag{A.136}$$

Plugging (A.134) and the assumption (4.17) into (A.136), we get

$$\hat{\delta}_0 \triangleq \|M - \hat{X}_0 \hat{Y}_0^T\|_F \leq \sqrt{\frac{C_2}{C_0}} \frac{\Sigma_{\min}}{r^{1.5}\kappa^2}. \tag{A.137}$$

The property (a), i.e. $(X_0, Y_0) \in (\sqrt{2/3}K_1)$ follows directly from the definitions of X_0 and Y_0 in (4.13). We then prove the property (b), i.e. $(X_0, Y_0) \in (\sqrt{2/3}K_2)$. By (A.137) we have $\|M - M_0\|_F \leq \Sigma_{\min}/5 \leq \Sigma_{\max}/5$ for large enough C_0. This inequality combined with $\|M - M_0\|_F \geq \|M - M_0\|_2 \geq \|M_0\|_2 - \Sigma_{\max}$ yields

$$\|M_0\|_2 \leq \frac{6}{5}\Sigma_{\max}. \tag{A.138}$$

By the definitions of $\hat{X}_0, \hat{Y}_0$ (i.e. $\hat{X}_0 = \bar{X}_0 D_0^{\frac{1}{2}}$, $\hat{Y}_0 = \bar{Y}_0 D_0^{\frac{1}{2}}$, where $\bar{X}_0 D_0 \bar{Y}_0^T$ is the SVD of M_0), we have

$$\|\hat{X}_0\|_2 = \|\hat{Y}_0\|_2 = \sqrt{\|M_0\|_2} \overset{\text{(A.138)}}{\leq} \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}. \tag{A.139}$$

Then we have

$$\|\hat{X}_0\|_F^2 \leq r\|\hat{X}_0\|_2^2 \leq \frac{6}{5} r\Sigma_{\max} \overset{(4.5)}{<} \frac{2}{3}\beta_T^2, \tag{A.140}$$

where the last inequality follows from $C_T > 9/5$. By the definition of X_0 in (4.13), we have $\|X_0\|_F^2 \leq \|\hat{X}_0\|_F^2 \leq \frac{2}{3}\beta_T^2$. Similarly, we can prove $\|Y_0\|_F^2 \leq \frac{2}{3}\beta_T^2$. Thus the property (b) is proved.

Next we prove the property (c), i.e. $\|M - X_0 Y_0^T\|_F \leq \delta_0$. Since $\hat{X}_0, \hat{Y}_0$ satisfy $\max\{\|\hat{X}_0\|_F, \|\hat{Y}_0\|_F\} \leq \beta_T$ (due to (A.140) and the analogous inequality for $\hat{Y}_0$) and (A.137), it follows from Proposition 4.3.1 that there exist U_0, V_0 such that

$$U_0 V_0^T = M; \tag{A.141a}$$

$$\|U_0\|_2 \leq \|X_0\|_2; \tag{A.141b}$$

$$\|U_0 - \hat{X}_0\|_F \leq \frac{6\|\hat{Y}_0\|_2}{5\Sigma_{\min}}\hat{\delta}_0, \ \ \|V_0 - \hat{Y}_0\|_F \leq \frac{3\|\hat{X}_0\|_2}{\Sigma_{\min}}\hat{\delta}_0; \tag{A.141c}$$

$$\|U_0^{(i)}\|^2 \leq \frac{r\mu}{m}\beta_T^2, \ \ \|V_0^{(j)}\|^2 \leq \frac{3r\mu}{2n}\beta_T^2. \tag{A.141d}$$

Note that the above inequalities (A.141b) and (A.141c) are not due to (4.38b) and (4.38c) of Proposition 4.3.1, but stronger results (A.32) and (A.40) established during the proof of Proposition 4.3.1.

Note that

$$\begin{aligned}\|M - X_0Y_0^T\|_F &= \|U_0(V_0 - Y_0)^T + (U_0 - X_0)Y_0^T\|_F \\ &\le \|U_0(V_0 - Y_0)^T\|_F + \|(U_0 - X_0)Y_0^T\|_F \\ &\le \|U_0\|_2\|V_0 - Y_0\|_F + \|U_0 - X_0\|_F\|Y_0\|_2,\end{aligned} \tag{A.142}$$

where the last inequality follows from Proposition A.3.4. Since $X_0^{(i)}$ and $\hat{X}_0^{(i)}$ has the same direction and $\|X_0^{(i)}\| \le \|\hat{X}_0^{(i)}\|$, by Proposition A.3.3 we have

$$\|X_0\|_2 \le \|\hat{X}_0\|_2 \le \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}. \tag{A.143}$$

Combining (A.141b) and (A.143), we get

$$\|U_0\|_2 \le \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}. \tag{A.144}$$

Similar to (A.143), we have

$$\|Y_0\|_2 \le \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}. \tag{A.145}$$

It remains to bound $\|V_0 - Y_0\|_F$ and $\|U_0 - X_0\|_F$. Let us prove the following inequality:

$$\|U_0^{(i)} - X_0^{(i)}\| \le \|U_0^{(i)} - \hat{X}_0^{(i)}\|, \ \ \forall\ i. \tag{A.146}$$

If $\|\hat{X}_0^{(i)}\| \le \sqrt{\frac{2}{3}}\beta_1$, then (A.146) becomes equality since $\hat{X}_0^{(i)} = X_0^{(i)}$. Thus we only need to consider the case $\|\hat{X}_0^{(i)}\| > \sqrt{\frac{2}{3}}\beta_1$. In this case by the definition of X_0 in (4.13) we have $\|X_0^{(i)}\| = \sqrt{\frac{2}{3}}\beta_1$. From (A.141d), we get

$$\|U_0^{(i)}\|^2 < \frac{3}{2}\frac{r\mu}{m}\beta_T^2 \le \frac{2}{3}\beta_1^2 < \|\hat{X}_0^{(i)}\|^2. \tag{A.147}$$

For simplicity, denote $u \triangleq U_0^{(i)}, x \triangleq X_0^{(i)}, \tau \triangleq \frac{\|\hat{X}_0^{(i)}\|}{\sqrt{2/3}\beta_1} = \frac{\|\hat{X}_0^{(i)}\|}{\|x\|} > 1$. Then (A.147) becomes $\|u\| \le \|x\|$ and (A.146) becomes $\|u - x\| \le \|u - \tau x\|$. The latter can be transformed as follows:

$$\begin{aligned}\|u - x\| \le \|u - \tau x\| &\iff \|x\|^2 - 2\langle u, x\rangle \le \tau^2\|x\|^2 - 2\tau\langle u, x\rangle \\ &\iff 2(\tau - 1)\langle u, x\rangle \le (\tau^2 - 1)\|x\|^2 \\ &\iff 2\langle u, x\rangle \le (\tau + 1)\|x\|^2.\end{aligned} \tag{A.148}$$

Since $\langle u, x\rangle \leq \|u\|\|x\| \leq \|x\|^2$ (here we use $\|u\| \leq \|x\|$ which is equivalent to (A.147)) and $2 < \tau+1$, the last inequality of (A.148) holds, which implies that $\|u-x\| \leq \|u-\tau x\|$ holds and, consequently, (A.146) holds.

An immediate consequence of (A.146) is

$$\|U_0 - X_0\|_F \leq \|U_0 - \hat{X}_0\|_F \overset{\text{(A.141c)}}{\leq} \frac{5\|\hat{Y}_0\|_2}{4\Sigma_{\min}}\hat{\delta}_0 \overset{\text{(A.139)}}{\leq} \frac{5}{4}\sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}\frac{\hat{\delta}_0}{\Sigma_{\min}}. \tag{A.149}$$

Similarly, we have

$$\|V_0 - Y_0\|_F \overset{\text{(A.141c)}}{\leq} 3\sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}\frac{\hat{\delta}_0}{\Sigma_{\min}}. \tag{A.150}$$

Plugging (A.144), (A.145), (A.149) and (A.150) into (A.142), we get

$$\begin{aligned}\|M - X_0Y_0^T\|_F \leq& \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}\frac{5}{4}\sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}\frac{\hat{\delta}_0}{\Sigma_{\min}} + \sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}3\sqrt{\frac{6}{5}}\sqrt{\Sigma_{\max}}\frac{\hat{\delta}_0}{\Sigma_{\min}}\\ =&(\frac{3}{2}+\frac{18}{5})\kappa\hat{\delta}_0\\ \overset{\text{(A.137)}}{\leq}& \frac{51}{10}\sqrt{\frac{C_2}{C_0}}\frac{\Sigma_{\min}}{r^{1.5}\kappa}\\ \overset{(4.6)}{\leq}& \delta_0,\end{aligned}$$

where the last inequality holds for $C_d \geq \frac{5}{153}\sqrt{\frac{C_0}{C_2}}$. Therefore property (c) is proved.

A.5.2 Proof of Claim 4.2.1

Denote $d \triangleq \|M - XY^T\|_F$. Let $a = U(V-Y)^T + (U-X)V^T$, $b = (U-X)(V-Y)$, where U, V are defined with the properties in Corollary 4.3.1.

According to (4.36) we have $\|\mathcal{P}_\Omega(a)\|_F^2 \geq \frac{27}{40}pd^2$. According to (4.30a), we have $\|\mathcal{P}_\Omega(b)\|_F \leq \frac{1}{5}\sqrt{p}d$. Therefore, $\|\mathcal{P}_\Omega(M - XY^T)\|_F = \|\mathcal{P}_\Omega(a-b)\|_F \geq \|\mathcal{P}_\Omega(a)\|_F - \|\mathcal{P}_\Omega(b)\|_F \geq \sqrt{\frac{27}{40}}\sqrt{p}d - \frac{1}{5}\sqrt{p}d \geq \frac{3}{5}\sqrt{p}d \geq \frac{1}{\sqrt{3}}\sqrt{p}d$.

According to (4.30b), we have $\|b\|_F \leq \frac{1}{10}d$. According to (4.35) (which is a corollary of [2, Theorem 4.1]), we have $\|\mathcal{P}_\Omega(a)\|_F^2 \leq \frac{7}{6}p\|a\|_F^2 \leq \frac{7}{6}p(\|M-XY^T\|_F+\|b\|_F)^2 \leq \frac{7}{6}p(1+\frac{1}{10})^2d^2 \leq \frac{17}{12}pd^2$. Thus, $\|\mathcal{P}_\Omega(a-b)\|_F \leq \|\mathcal{P}_\Omega(a)\|_F + \|\mathcal{P}_\Omega(b)\|_F \leq (\sqrt{\frac{17}{12}}+\frac{1}{5})\sqrt{p}d \leq \sqrt{2p}d$. □

A.5.3 Proof of Proposition 4.4.1

We first provide a general condition for $(X, Y) \in K_1 \cap K_2$ (i.e. incoherent and bounded) based on the function value $\tilde{F}(X, Y)$.

Proposition A.5.1 *Suppose the sample set Ω satisfies* (4.19) *and $\rho = 2p\delta_0^2/G_0(3/2)$, where δ_0 is defined in* (4.6). *Suppose (X_0, Y_0) satisfies* (4.56) *and*

$$\tilde{F}(X, Y) \leq 2\tilde{F}(X_0, Y_0). \tag{A.151}$$

Then $(X, Y) \in K_1 \cap K_2$.

Proof of Proposition A.5.1: We prove by contradiction. Assume the contrary that $(X, Y) \notin K_1 \cap K_2$. By the definition of K_1, K_2 in (4.20), we have either $\|X^{(i)}\|^2 > \beta_1^2$ for some i, $\|Y^{(j)}\|^2 > \beta_2^2$ for some j, $\|X\|_F^2 > \beta_T^2$ or $\|Y\|_F^2 > \beta_T^2$. Hence at least one term of $G(X, Y) = \rho \sum_{i=1}^m G_0(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) + \rho \sum_{j=1}^n G_0(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2}) + \rho G_0(\frac{3\|X\|_F^2}{2\beta_T^2}) + \rho G_0(\frac{3\|Y\|_F^2}{2\beta_T^2})$ is larger than $G_0(\frac{3}{2})$. In addition, all the other terms in the expression of $G(X, Y)$ are nonnegative, thus we have $G(X, Y) > \rho G_0(\frac{3}{2})$. Therefore,

$$\tilde{F}(X, Y) \geq G(X, Y) > \rho G_0(\frac{3}{2}) = 2p\delta_0^2. \tag{A.152}$$

We have

$$\tilde{F}(X_0, Y_0) = \frac{1}{2}\|\mathcal{P}_\Omega(M - X_0 Y_0^T)\|_F^2 \leq p\|M - X_0 Y_0^T\|_F^2 \leq p\delta_0^2, \tag{A.153}$$

where the first equality is due to $G(X_0, Y_0) = 0$ which follows from $(X_0, Y_0) \in (\sqrt{\frac{2}{3}}K_1) \cap (\sqrt{\frac{2}{3}}K_2)$, the second inequality follows from (4.19) and the fact $(X_0, Y_0) \in (\sqrt{\frac{2}{3}}K_1) \cap (\sqrt{\frac{2}{3}}K_2) \cap K(\delta_0) \subseteq K_1 \cap K_2 \cap K(\delta)$, and the last inequality is due to $(X_0, Y_0) \in K(\delta_0)$. Combining (A.152) and (A.153), we get

$$\tilde{F}(X, Y) > 2\tilde{F}(X_0, Y_0),$$

which contradicts (A.151). □

We can prove that (4.57) implies

$$\tilde{F}(\boldsymbol{x}_i) \leq 2\tilde{F}(\boldsymbol{x}_0), \ \forall\ i. \tag{A.154}$$

In fact, when (4.57c) holds, as the first inequality in (4.57c) the above relation also holds. When (4.57a) holds, let $\lambda = 0$ in (4.57a) we get (A.154). When (4.57b) holds, we have

$$\psi(\boldsymbol{x}_i, \boldsymbol{\Delta}_i; 1) \overset{(4.57b)}{\leq} \psi(\boldsymbol{x}_i, \boldsymbol{\Delta}_i; 0) \overset{(4.55b)}{=} \tilde{F}(\boldsymbol{x}_i), \tag{A.155}$$

which implies $\tilde{F}(\boldsymbol{x}_{i+1}) = \tilde{F}(\boldsymbol{x}_i + \boldsymbol{\Delta}_i) \overset{(4.55b)}{\leq} \psi(\boldsymbol{x}_i, \boldsymbol{\Delta}_i; 1) \leq \tilde{F}(\boldsymbol{x}_i)$. This relation holds for any i, thus $\tilde{F}(\boldsymbol{x}_{i+1}) \leq \tilde{F}(\boldsymbol{x}_i) \leq \cdots \leq \tilde{F}(\boldsymbol{x}_0) \leq 2\tilde{F}(\boldsymbol{x}_0)$.

Since (4.57) implies implies $\tilde{F}(\boldsymbol{x}_t) \leq 2\tilde{F}(\boldsymbol{x}_0)$ (see (A.154)), by Proposition A.5.1 we have $\boldsymbol{x}_t \in K_1 \cap K_2$. The rest of the proof is devoted to establish

$$\boldsymbol{x}_t \in K(\frac{2}{3}\delta), \ \ \forall\ t. \tag{A.156}$$

Define the distance of $\boldsymbol{x} = (X, Y)$ and $\boldsymbol{u} = (U, V)$ as

$$d(\boldsymbol{x}, \boldsymbol{u}) = \|XY^T - UV^T\|_F,$$

then $(X_t, Y_t) \in K(\delta) \Longleftrightarrow \|X_t Y_t^T - M\|_F \leq \delta$ can be expressed as

$$d(\boldsymbol{x}_t, \boldsymbol{u}^*) \leq \delta.$$

We first prove the following result:

Lemma A.5.2 *If $\tilde{F}(\boldsymbol{x}) \leq 2\tilde{F}(\boldsymbol{x}_0)$, then $d(\boldsymbol{u}^*, \boldsymbol{x}) \notin [\frac{2}{3}\delta, \delta]$.*

Proof of Lemma A.5.2: We prove by contradiction. Assume the contrary that

$$d(\boldsymbol{u}^*, \boldsymbol{x}) \in [\frac{2}{3}\delta, \delta]. \tag{A.157}$$

Since $\boldsymbol{x}_0$ satisfies (4.56), according to the proof of Proposition A.5.1 we have (A.153), i.e.

$$\tilde{F}(\boldsymbol{x}_0) \leq p\delta_0^2. \tag{A.158}$$

According to Proposition A.5.1 and the assumption $\tilde{F}(\boldsymbol{x}) \leq 2\tilde{F}(\boldsymbol{x}_0)$, we have $\boldsymbol{x} \in K_1 \cap K_2$. Together with (A.157) we get $\boldsymbol{x} \in K_1 \cap K_2 \cap K(\delta)$. Then we have

$$\tilde{F}(\boldsymbol{x}) \geq \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|^2 \overset{(4.19)}{\geq} \frac{1}{6}p\|M - XY^T\|^2 = \frac{1}{6}pd(\boldsymbol{u}^*, \boldsymbol{x})^2. \tag{A.159}$$

Plugging $d(\boldsymbol{u}^*, \boldsymbol{x})^2 \geq (\frac{2}{3})^2\delta^2 \overset{(4.6)}{=} 16\delta_0^2 \overset{(A.158)}{\geq} 16\tilde{F}(\boldsymbol{x}_0)/p$ into (A.159), we get $\tilde{F}(\boldsymbol{x}) \geq \frac{8}{3}\tilde{F}(\boldsymbol{x}_0)$, which together with the assumption $\tilde{F}(\boldsymbol{x}) \leq 2\tilde{F}(\boldsymbol{x}_0)$ leads to $\tilde{F}(\boldsymbol{x}) = \tilde{F}(\boldsymbol{x}_0) =$

0. Then by (A.159) we get $d(\boldsymbol{u}^*, \boldsymbol{x}) = 0$, which contradicts (A.157) since $\delta > 0$. Thus Lemma A.5.2 is proved.

Now we get back to the proof of (A.156). We prove (A.156) by induction on t. The basis of the induction holds due to (4.56) and the fact $\delta_0 = \delta/6$. Suppose $\boldsymbol{x}_t \in K(2\delta/3)$, we need to prove $\boldsymbol{x}_{t+1} \in K(2\delta/3)$. Assume the contrary that $\boldsymbol{x}_{t+1} \notin K(2\delta/3)$, i.e.

$$d(\boldsymbol{u}^*, \boldsymbol{x}_{t+1}) > \frac{2}{3}\delta. \tag{A.160}$$

Let $i = t + 1$ in (A.154), we get $\tilde{F}(\boldsymbol{x}_{t+1}) \leq 2\tilde{F}(\boldsymbol{x}_0)$. Then by Lemma A.5.2 we have

$$d(\boldsymbol{x}_{t+1}, \boldsymbol{u}^*) \notin [\frac{2}{3}\delta, \delta]; \tag{A.161}$$

Combining (A.161) and (A.160), we get

$$d(\boldsymbol{x}_{t+1}, \boldsymbol{u}^*) > \delta. \tag{A.162}$$

In the rest of the proof, we will derive a contradiction for the three cases (4.57a), (4.57b) and (4.57c) separately.

Case 1: (4.57a) holds. By the induction hypothesis, $d(\boldsymbol{x}_t, \boldsymbol{u}^*) \leq \frac{2}{3}\delta$. Since $d(\boldsymbol{x}, \boldsymbol{u}^*)$ is a continuous function over $\boldsymbol{x}$, the relation $d(\boldsymbol{x}_t, \boldsymbol{u}^*) \leq \frac{2}{3}\delta$ and (A.162) imply that there must exist some $\boldsymbol{x}' = (1 - \lambda)\boldsymbol{x}_{t+1} + \lambda\boldsymbol{x}_t, \lambda \in [0, 1]$ such that

$$d(\boldsymbol{x}', \boldsymbol{u}^*) = \delta. \tag{A.163}$$

According to (4.57a), we have $\tilde{F}(\boldsymbol{x}') \leq 2\tilde{F}(\boldsymbol{x}_0)$. By Lemma A.5.2, we have $d(\boldsymbol{u}^*, \boldsymbol{x}') \notin [\frac{2}{3}\delta, \delta]$, which contradicts (A.163).

Case 2: (4.57b) holds. Define

$$\boldsymbol{\lambda}' = \arg \min_{\lambda \in \mathbb{R}, d(\boldsymbol{x}_t + \lambda\Delta_t, \boldsymbol{u}^*) \leq \delta} \psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; \lambda). \tag{A.164}$$

By the induction hypothesis, $d(\boldsymbol{x}_t, \boldsymbol{u}^*) \leq \delta$, thus 0 lies in the feasible region of the optimization problem in (A.164), which implies

$$\psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; \lambda') \leq \psi(\boldsymbol{x}_t, \Delta_t; 0) \overset{(4.55b)}{=} \tilde{F}(\boldsymbol{x}_t). \tag{A.165}$$

Define $\boldsymbol{x}' = \boldsymbol{x}_t + \lambda'\Delta_t$, then the feasibility of λ' for the optimization problem in (A.164) implies $\delta \geq d(\boldsymbol{x}', \boldsymbol{u}^*)$. Since $d(\boldsymbol{x}, \boldsymbol{u}^*)$ is a continuous function over $\boldsymbol{x}$ and $d(\boldsymbol{x}', \boldsymbol{u}^*) \leq$

$\delta \overset{(A.162)}{<} d(\boldsymbol{x}_{t+1}, \boldsymbol{u}^*)$, there must exist some $\boldsymbol{x}'' = (1-\epsilon)\boldsymbol{x}_{t+1} + \epsilon \boldsymbol{x}' = \boldsymbol{x}_t + (1 - \epsilon + \epsilon\lambda')\boldsymbol{\Delta}_t, \epsilon \in [0, 1]$ such that

$$d(\boldsymbol{x}'', \boldsymbol{u}^*) = \delta. \tag{A.166}$$

Then we have

$$\tilde{F}(\boldsymbol{x}'') \overset{(4.55b)}{\leq} \psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; 1 - \epsilon + \epsilon\lambda') \overset{(4.55a)}{\leq} (1-\epsilon)\psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; 1) + \epsilon\psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; \lambda') \overset{(A.155),(A.165)}{\leq} \tilde{F}(\boldsymbol{x}_t) \overset{(A.154)}{\leq} 2\tilde{F}(\boldsymbol{x}_0).$$

Again we apply Lemma A.5.2 to obtain $d(\boldsymbol{u}^*, \boldsymbol{x}'') \notin [\frac{2}{3}\delta, \delta]$, which contradicts (A.166).

Case 3: (4.57c) holds. By (4.56) and the fact $\delta_0 = \delta/6$ we get $d(\boldsymbol{x}_0, \boldsymbol{u}^*) \leq \delta/6$. Then we have

$$d(\boldsymbol{x}_{t+1}, \boldsymbol{u}^*) \leq d(\boldsymbol{x}_{t+1}, \boldsymbol{x}_0) + d(\boldsymbol{x}_0, \boldsymbol{u}^*) \overset{(4.57c)}{\leq} \frac{5}{6}\delta + \frac{1}{6}\delta = \delta,$$

which contradicts (A.162).

In all three cases we have arrived at a contradiction, thus the assumption (A.160) does not hold, which finishes the induction step for $t+1$. Therefore, (A.156) holds for all t.

A.5.4 Proof of Claim 4.4.3

The sequence $\{\boldsymbol{x}_t\}$ generated by Algorithm 1 with either restricted Armijo rule or restricted line search satisfies (4.57c) because the sequence $\tilde{F}(\boldsymbol{x}_t)$ is decreasing and the requirement $d(\boldsymbol{x}_t, \boldsymbol{x}_0) \leq 5\delta/6$ is enforced throughout computation.

Algorithm 2 and Algorithm 3 satisfy (4.57b) since all of them perform exact minimization of a convex upper bound of the objective function along some directions. Note that $\boldsymbol{x}_t$ should be understood as the produced solution after t "iterations" (one block of variables is updated in one "iteration"). In contrast, (X_k, Y_k) defined in these algorithms is the produced solution after k "loops" (all variables are updated once in one "loop"). For (X_k, Y_k) generated by Algorithm 2, we define $\boldsymbol{x}_{2k} = (X_k, Y_k), \boldsymbol{x}_{2k+1} = (X_{k+1}, Y_k)$ and $\psi(\boldsymbol{x}_t, \Delta_t; \lambda) = \tilde{F}(\boldsymbol{x}_t + \lambda\Delta_t)$, then ψ satisfies (4.55) and $\{\boldsymbol{x}_t\}_{t=0}^{\infty} = \{(X_k, Y_k), (X_{k+1}, Y_k)\}_{k=0}^{\infty}$ satisfies (4.57b). Similarly, for (X_k, Y_k) generated by Algorithm 3, define

$$\boldsymbol{x}_{(m+n)k+i} = (X_{k+1}^{(1)}, \ldots, X_{k+1}^{(i-1)}, X^{(i)}, X_k^{(i+1)}, \ldots, X_k^{(m)}, Y_k),\ i = 1, \ldots, m$$
$$\boldsymbol{x}_{(m+n)k+m+j} = (X_{k+1}, Y_{k+1}^{(1)}, \ldots, Y_{k+1}^{(j-1)}, Y^{(j)}, Y_k^{(j+1)}, \ldots, Y_k^{(m)}),\ j = 1, \ldots, n,$$

and $\psi(\boldsymbol{x}_t, \boldsymbol{\Delta}_t; \lambda) = \tilde{F}(\boldsymbol{x}_t + \lambda\boldsymbol{\Delta}_t) + \lambda_0\|\lambda\boldsymbol{\Delta}_t\|^2/2$, then ψ satisfies (4.55) and $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ satisfies (4.57b).

We then show that Algorithm 1 with constant stepsize $\eta < \bar{\eta}_1$ satisfies (4.57a) for some $\bar{\eta}_1$ when Ω satisfies (4.19). We prove by induction on t. Define $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$, then (4.57a) holds for $t = 0$. Assume (4.57a) holds for $t-1$, i.e., $\tilde{F}(\boldsymbol{x}_{t-1} + \lambda\boldsymbol{\Delta}_{t-1}) \le 2\tilde{F}(\boldsymbol{x}_0), \forall\lambda \in [0,1]$, where $\boldsymbol{\Delta}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$. In particular, we have $\tilde{F}(\boldsymbol{x}_t) \le 2\tilde{F}(\boldsymbol{x}_0)$, which together with the assumption that Ω satisfies (4.19) leads to (by Proposition (A.5.1))

$$\boldsymbol{x}_t \in K_1 \cap K_2.$$

Thus $\max\{\|X_t\|_F, \|Y_t\|_F\} \le \beta_T$, $\|X_t^{(i)}\| \le \beta_1, \forall i$, and $\|Y_t^{(j)}\| \le \beta_2, \forall j$. Then we have

$$\begin{aligned}
&\|\nabla_X \tilde{F}(\boldsymbol{x}_t)\|_F = \|\nabla_X F(\boldsymbol{x}_t) + \nabla_X G(\boldsymbol{x}_t)\|_F \\
&\le \|\mathcal{P}_\Omega(X_tY_t^T - M)Y_t\|_F + \left\|\rho\sum_{i=1}^{m} G_0'(\frac{3\|X_t^{(i)}\|^2}{2\beta_1^2})\frac{3\bar{X}_t^{(i)}}{\beta_1^2}\right\|_F + \left\|\rho G_0'(\frac{3\|X_t\|_F^2}{2\beta_T^2})\frac{3X_t}{\beta_T^2}\right\|_F \\
&\le \|\mathcal{P}_\Omega(X_tY_t^T - M)\|_F\|Y_t\|_F + \frac{3\rho\|X_t\|_F}{\beta_1^2} + \frac{3\rho\|X_t\|_F}{\beta_T^2} \\
&\le \sqrt{\tilde{F}(\boldsymbol{x}_t)}\beta_T + \frac{6\rho\|X_t\|_F}{\beta_1^2} \\
&\le \sqrt{2\tilde{F}(\boldsymbol{x}_0)}\beta_T + \frac{6\rho\beta_T}{\beta_1^2},
\end{aligned}$$

where in the second inequality we use $G_0'(\frac{3\|X_t^{(i)}\|^2}{2\beta_1^2}) \le G_0'(\frac{3}{2}) = 1$ and $G_0'(\frac{3\|X\|_F^2}{2\beta_T^2}) \le G_0'(\frac{3}{2}) = 1$. Assume

$$\bar{\eta}_1 \le \frac{1}{4\beta_T^2}. \tag{A.167}$$

Recall that $\eta \le \bar{\eta}_1$, thus we have

$$\begin{aligned}
\|X_{t+1}\|_F \le \|X_t\|_F + \eta\|\nabla_X\tilde{F}(\boldsymbol{x}_t)\|_F &\le \beta_T + \frac{1}{4\beta_T^2}\left(\sqrt{2\tilde{F}(\boldsymbol{x}_0)}\beta_T + \frac{6\rho\beta_T}{\beta_1^2}\right) \\
&\overset{(A.153)}{\le} \beta_T + \frac{1}{4\beta_T}\left(\sqrt{2p}\delta_0 + \frac{6\rho}{\beta_1^2}\right) \triangleq c_1.
\end{aligned} \tag{A.168}$$

By a similar argument, we can prove $\|Y_{t+1}\|_F \le c_1$, thus $\boldsymbol{x}_{t+1} = (X_{t+1}, Y_{t+1}) \in \Gamma(c_1)$ (recall the definition of $\Gamma(\cdot)$ in (4.10) is $\Gamma(\beta) = \{(X,Y) \mid \|X\|_F \le \beta, \|Y\|_F \le \beta\}$). Since $(X_t, Y_t) \in \Gamma(\beta_T) \subseteq \Gamma(c_1)$ and $\Gamma(c_1)$ is a convex set, we have that the line segment

connecting $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$, denoted as $[\boldsymbol{x}_t, \boldsymbol{x}_{t+1}]$, lies in $\Gamma(c_1)$. Then by Claim 4.1.1 we have that $\nabla\tilde{F}$ is Lipschitz continuous in $[\boldsymbol{x}_t, \boldsymbol{x}_{t+1}]$ with Lipschitz constant

$$L_1 = L(c_1) = 4c_1^2 + 54\rho\frac{c_1^2}{\beta_1^4} \geq L(\beta_T) \geq 4\beta_T^2, \tag{A.169}$$

where the last inequality is due to the fact $c_1 \geq \beta_T$. Define (note c_1 is defined by (A.168))

$$\bar{\eta}_1 \triangleq \frac{1}{L_1} = \frac{1}{4c_1^2 + 54\rho\frac{c_1^2}{\beta_1^4}}, \tag{A.170}$$

then $\bar{\eta}_1 \leq \frac{1}{L(\beta_T)} \leq \frac{1}{4\beta_T^2} = \frac{1}{4\beta_T^2}$, which is consistent with (A.167).

It follows from a classical descent lemma (see, e.g., [86, Prop. A.24]) that

$$\begin{aligned}
\tilde{F}(\boldsymbol{x}_t - \lambda\eta\nabla\tilde{F}(\boldsymbol{x}_t)) &\leq \tilde{F}(\boldsymbol{x}_t) - \langle\lambda\eta\nabla\tilde{F}(\boldsymbol{x}_t), \nabla\tilde{F}(\boldsymbol{x}_t)\rangle + \frac{L_1}{2}\|\boldsymbol{\lambda}\eta\nabla\tilde{F}(\boldsymbol{x}_t)\|^2 \\
&= \tilde{F}(\boldsymbol{x}_t) + \|\nabla\tilde{F}(\boldsymbol{x}_t)\|^2(\frac{L_1}{2}\lambda^2\eta^2 - \lambda\eta) \\
&\leq \tilde{F}(\boldsymbol{x}_t) - \frac{\lambda\eta}{2}\|\nabla\tilde{F}(\boldsymbol{x}_t)\|^2 \\
&\leq \tilde{F}(\boldsymbol{x}_t) \\
&\leq 2\tilde{F}(\boldsymbol{x}_0),\ \forall\ \lambda \in [0,1],
\end{aligned} \tag{A.171}$$

where the second inequality follows from the fact that $\lambda\eta \leq \eta \leq \bar{\eta}_1 = 1/L_1$. This finishes the induction step (note that $\boldsymbol{\Delta}_t = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\eta\nabla\tilde{F}(\boldsymbol{x}_t)$), thus (4.57a) is proved.

Finally, we show that Algorithm 4 (SGD) satisfies (4.57a) with $\boldsymbol{x}_t = (X_k, Y_k)$ representing the produced solution after the t-th loop, provided that Ω satisfies (4.19). Denote $N = |\Omega| + m + n + 2$ and $\boldsymbol{x}_{k,i} = (X_{k,i}, Y_{k,i}), i = 1, \ldots, N$. We prove (4.57a) by induction on t. Define $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$, then (4.57a) holds for $t = 0$. Assume (4.57a) holds for $0, 1, \ldots t-1$, i.e., $\tilde{F}(\boldsymbol{x}_k + \lambda\boldsymbol{\Delta}_k) \leq 2\tilde{F}(\boldsymbol{x}_0), \forall\lambda \in [0,1]$, where $\boldsymbol{\Delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, $0 \leq k \leq t-1$. In particular, we have $\tilde{F}(\boldsymbol{x}_t) \leq 2\tilde{F}(\boldsymbol{x}_0)$, which together with the assumption that Ω satisfies (4.19) leads to (by Proposition (A.5.1))

$$\boldsymbol{x}_t \in K_1 \cap K_2. \tag{A.172}$$

Now we show that there exist constants $c_{1,i}, c_{2,i}, i = 0, 1, \ldots, N$ (independent of t) so

that

$$\max\{\|X_{t,i}\|_F,\|Y_{t,i}\|_F\} \le c_{1,i}, \tag{A.173a}$$

$$\max\{\|\nabla_X f_{i+1}(\boldsymbol{x}_{t,i})\|_F, \|\nabla_Y f_{i+1}(\boldsymbol{x}_{t,i-1})\|_F\} \le c_{2,i}. \tag{A.173b}$$

We prove (A.173) by induction on i. When $i = 0$, since by (A.172) we have

$$\max\{\|X_{t,0}\|_F, \|Y_{t,0}\|_F\} = \max\{\|X_t\|_F, \|Y_t\|_F\} \le \beta_T,$$

thus (A.173a) holds for $c_{1,0} = \beta_T$.

Suppose (A.173a) holds for i, we prove (A.173b) holds for i with suitably chosen $c_{2,i}$. Note that f_{i+1} can be one of the five different functions in (4.16). When f_{i+1} equals some F_{jl}, we have

$$\begin{aligned}\|\nabla_X f_{i+1}(\boldsymbol{x}_{t,i})\|_F = \|\nabla_X F_{j,l}(\boldsymbol{x}_{t,i})\|_F &= |(X_{t,i}^{(j)})^T Y_{t,i}^{(l)} - M_{jl}|\|Y_{t,i}^{(l)}\| \\ &\le (\|X_{t,i}\|_F\|Y_{t,i}\|_F + M_{\max})\|Y_{t,i}\|_F \le (c_{1,i}^2 + M_{\max})c_{1,i}.\end{aligned}$$

When $f_{i+1}(X,Y)$ equals some $G_{1j}(X)$, we have (see (4.14) for the expression of $\nabla_X G_{1j}$)

$$\begin{aligned}\|\nabla_X f_{i+1}(\boldsymbol{x}_{t,i})\|_F = \|\nabla_X G_{1j}(X_{t,i})\|_F &= \rho G_0'(\frac{3\|X_{t,i}^{(j)}\|^2}{2\beta_1^2})\frac{3\|X_{t,i}^{(j)}\|}{\beta_1^2} \\ &\le \rho G_0'(\frac{3c_{1,i}^2}{2\beta_1^2})\frac{3c_{1,i}}{\beta_1^2} \le \rho G_0'(\frac{3c_{1,i}^2}{2\beta_T^2})\frac{3c_{1,i}}{\beta_T^2}.\end{aligned}$$

When $f_{i+1}(X,Y)$ equals some $G_3(X)$, we have

$$\|\nabla_X f_{i+1}(\boldsymbol{x}_{t,i})\|_F = \|\nabla_X G_3(X_{t,i})\|_F = \rho G_0'(\frac{3\|X_{t,i}\|_F^2}{2\beta_T^2})\frac{3\|X_{t,i}\|_F}{\beta_T^2} \le \rho G_0'(\frac{3c_{1,i}^2}{2\beta_T^2})\frac{3c_{1,i}}{\beta_T^2}.$$

When $f_{i+1}(X,Y)$ equals some $G_{2j}(Y)$ or $G_4(Y)$ that only depend on Y, we have $\nabla_X f_{i+1}(\boldsymbol{x}_{t,i}) = 0$. Let

$$c_{2,i} \triangleq \max\left\{(c_{1,i}^2 + M_{\max})c_{1,i},\ \ \rho G_0'(\frac{3c_{1,i}^2}{2\beta_T^2})\frac{3c_{1,i}}{\beta_T^2}\right\},$$

then no matter what kind of function f_{i+1} is, we always have $\|\nabla_X f_{i+1}(\boldsymbol{x}_{t,i})\|_F \le c_{2,i}$. Similarly, $\|\nabla_Y f_{i+1}(\boldsymbol{x}_{t,i})\|_F \le c_{2,i}$. Thus (A.173b) holds for i.

Suppose (A.173b) holds for $i-1$, we prove that (A.173a) holds for i with suitably chosen $c_{1,i}$. In fact,

$$\|X_{t,i}\|_F = \|X_{t,i-1} - \eta_t \nabla_X f_i(\boldsymbol{x}_{t,i-1})\|_F \le \|X_{t,i-1}\|_F + \eta_t\|\nabla_X f_i(\boldsymbol{x}_{t,i-1})\|_F \le c_{1,i-1} + \bar{\eta}c_{2,i-1},$$

thus (A.173a) holds for $c_{1,i} = c_{1,i-1} + \bar{\eta} c_{2,i-1}$. This finishes the induction proof of (A.173).

In Claim 4.1.1, we have proved that $\nabla \tilde{F}$ is Lipschitz continuous with Lipschitz constant $L(\beta_0) = 4\beta_0 + 54\rho\frac{\beta_0^2}{\beta_1^4}$ in the set $\Gamma(\beta_0)$ (the definition of $\Gamma(\cdot)$ is given in (4.10)). By a similar argument (or set irrelevant rows of X, Y, U, V to zero in the proof of Claim (4.1.1)), we can prove that each ∇f_i is also Lipschitz continuous with Lipschitz constant $L(\beta_0) = 4\beta_0 + 54\rho\frac{\beta_0^2}{\beta_1^4}$ in the set $\Gamma(\beta_0)$. Then we have

$$\|\nabla f_i(\boldsymbol{x}_{t,i-1}) - \nabla f_i(\boldsymbol{x}_t)\|_F \le c'_{i-1}\|\boldsymbol{x}_{t,i-1} - \boldsymbol{x}_t\|_F, \quad i = 1, \ldots, N, \tag{A.174}$$

where $c'_{i-1} = L(c_{1,i-1})$.

Note that $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \sum_{i=1}^N (\boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,i-1}) = \boldsymbol{x}_t - \eta_t \sum_{i=1}^N \nabla f_i(\boldsymbol{x}_{t,i-1})$. We can express SGD as an approximate gradient descent method:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t(\nabla \tilde{F}(\boldsymbol{x}_t) + w_t), \tag{A.175}$$

where the error

$$w_t = \sum_{i=1}^N \nabla f_i(\boldsymbol{x}_{t,i-1}) - \nabla \tilde{F}(\boldsymbol{x}_t) = \sum_{i=1}^N (\nabla f_i(\boldsymbol{x}_{t,i-1}) - \nabla f_i(\boldsymbol{x}_t)).$$

Following the analysis in [90, Lemma 1], we can bound each term $\nabla f_i(\boldsymbol{x}_{t,i-1}) - \nabla f_i(\boldsymbol{x}_t)$ as

$$\begin{aligned}\|\nabla f_i(\boldsymbol{x}_{t,i-1}) - \nabla f_i(\boldsymbol{x}_t)\|_F &\overset{\text{(A.174)}}{\le} c'_{i-1}\|\boldsymbol{x}_{t,i-1} - \boldsymbol{x}_t\|_F = \eta_t c'_{i-1}\|\sum_{l=1}^{i-1} \nabla f_l(\boldsymbol{x}_{t,l-1})\|_F \\ &\overset{\text{(A.173b)}}{\le} \eta_t c'_{i-1} \sum_{l=1}^{i-1} \sqrt{2} c_{2,l}.\end{aligned} \tag{A.176}$$

Plugging this inequality for $i = 1, \ldots, N$ into the expression of w_t, we obtain an upper bound of the error w_t:

$$\|w_t\|_F \le \eta_t c_0, \tag{A.177}$$

where $c_0 \triangleq \sum_{i=1}^N (c'_{i-1} \sum_{l=1}^{i-1} \sqrt{2} c_{2,l})$ is a constant.

Applying (A.173a) for $i = N$, we get $\max\{\|X_{t+1}\|_F, \|Y_{t+1}\|_F\} \le c_{1,N}$, thus $\boldsymbol{x}_{t+1} \in \Gamma(c_{1,N})$. Since $\boldsymbol{x}_t \in \Gamma(\beta_T) \subseteq \Gamma(c_{1,N})$ and $\Gamma(c_{1,N})$ is a convex set, we have that the line

segment connecting $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ lies in $\Gamma(c_{1,N})$. Then by Claim 4.1.1 we have that $\nabla\tilde{F}$ is Lipschitz continuous over this line segment with Lipschitz constant $L' = L(c_{1,N})$. It follows from a classical descent lemma (see, e.g., [86, Prop. A.24]) that

$$\tilde{F}(\boldsymbol{x}_{t+1}) \leq \tilde{F}(\boldsymbol{x}_t) + \langle \boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \nabla\tilde{F}(\boldsymbol{x}_t)\rangle + \frac{L'}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_F^2.$$

Using the expression (A.175), the above relation becomes

$$\tilde{F}(\boldsymbol{x}_{t+1}) - \tilde{F}(\boldsymbol{x}_t) \leq -\eta_t\langle \boldsymbol{\nabla}\tilde{F}(\boldsymbol{x}_t) + w_t, \nabla\tilde{F}(\boldsymbol{x}_t)\rangle + \frac{L'}{2}\eta_t^2\|\nabla\tilde{F}(\boldsymbol{x}_t) + w_t\|_F^2. \tag{A.178}$$

Plugging

$$-\eta_t\langle w_t, \nabla\tilde{F}(\boldsymbol{x}_t)\rangle \leq \eta_t\|w_t\|_F\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F \overset{\text{(A.177)}}{\leq} \eta_t^2 c_0\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F \leq \frac{1}{2}\eta_t^2 c_0(1+\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2)$$

and

$$\frac{1}{2}\|\nabla\tilde{F}(\boldsymbol{x}_t) + w_t\|_F^2 \leq \|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2 + \|w_t\|_F^2 \overset{\text{(A.177)}}{\leq} \|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2 + \eta_t^2 c_0^2$$

into (A.178), we get

$$\begin{aligned}\tilde{F}(\boldsymbol{x}_{t+1}) - \tilde{F}(\boldsymbol{x}_t) &\leq -\eta_t\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2 + \frac{1}{2}\eta_t^2 c_0(1 + \|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2) + L'\eta_t^2(\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2 + \eta_t^2 c_0^2)\\ &= (\frac{1}{2}\eta_t^2 c_0 + \eta_t^2 L' - \eta_t)\|\nabla\tilde{F}(\boldsymbol{x}_t)\|_F^2 + \eta_t^2(\frac{1}{2}c_0 + L'\eta_t^2 c_0^2).\end{aligned} \tag{A.179}$$

Pick

$$\bar{\eta} \triangleq \frac{1}{c_0 + 2L'}.$$

Since $\eta_t \leq \bar{\eta}$, we have $\frac{1}{2}\eta_t^2 c_0 + \eta_t^2 L' - \eta_t \leq -\eta_t/2$ and $L'\eta_t^2 c_0^2 \leq L' c_0^2 \frac{1}{(c_0+2L')^2} \leq \frac{c_0}{8}$ (the last inequality follows from $(c_0 + 2L')^2 \geq 8c_0 L'$). Plugging these two inequalities into (A.179), we obtain

$$\tilde{F}(\boldsymbol{x}_{t+1}) - \tilde{F}(\boldsymbol{x}_t) \leq \eta_t^2 c_0.$$

By the same argument we can prove

$$\tilde{F}(\boldsymbol{x}_{k+1}) - \tilde{F}(\boldsymbol{x}_k) \leq \eta_k^2 c_0, \;\; k = 0, 1, \ldots, t.$$

Summing up these inequalities, we get

$$\tilde{F}(\boldsymbol{x}_{t+1}) \leq \tilde{F}(\boldsymbol{x}_0) + \sum_{k=0}^{t}\eta_k^2 c_0 \leq \tilde{F}(\boldsymbol{x}_0) + \eta_{\text{sum}} c_0.$$

where the last inequality follows from the assumption $\sum_{k=0}^{\infty} \eta_k^2 \leq \eta_{\text{sum}}$. Pick

$$\eta_{\text{sum}} \triangleq \frac{\tilde{F}(\boldsymbol{x}_0)}{c_0},$$

the above relation becomes

$$\tilde{F}(\boldsymbol{x}_{t+1}) \leq 2\tilde{F}(\boldsymbol{x}_0).$$

By a similar argument, we can prove

$$\tilde{F}(\boldsymbol{x}_t + \lambda(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)) \leq 2\tilde{F}(\boldsymbol{x}_0), \ \forall \ \lambda \in [0, 1],$$

which completes the induction. Thus we have proved that Algorithm 4 (SGD) satisfies (4.57a) with suitably chosen $\bar{\eta}$ and η_{sum}.

A.5.5 Proof of Claim 4.4.1

For Algorithm 1 with constant stepsize $\eta < \bar{\eta}_1$ (defined in (A.170)), since the objective value $\tilde{F}(\boldsymbol{x}_t)$ is decreasing, we have $\tilde{F}(\boldsymbol{x}_t) \leq \tilde{F}(\boldsymbol{x}_0)$. By Proposition A.5.1 this implies that the algorithm generates a sequence in $K_1 \cap K_2$. By Claim 4.1.1 and the fact $K_2 = \Gamma(\beta_T)$ (see the definitions of K_2 in (4.20) and the definition of $\Gamma(\cdot)$ in (4.10)), $\nabla\tilde{F}$ is Lipschitz continuous with Lipschitz constant $L(\beta_T)$ over the set K_2. According to [86, Proposition 1.2.3], each limit point of the sequence generated by Algorithm 1 with constant stepsize $\eta < \bar{\eta}_1 \overset{\text{(A.170)}}{\leq} 2/L(\beta_T)$ is a stationary point of problem (P1).

We then consider Algorithm 1 with stepsize chosen by the restricted Armijo rule. The proof of [86, Proposition 1.2.1] for the standard Armijo rule can not be directly applied, and some extra effort is needed. For the restricted Armijo rule, the procedure of picking the stepsize η_k can be viewed as a two-phase approach. In the first phase, we find the smallest nonnegative integer so that the distance requirement is fulfilled, i.e.

$$i_1 \triangleq \min\{i \in \mathbb{Z}^+ \mid d(\boldsymbol{x}_k(\xi^i s_0), \boldsymbol{x}_0) \leq \frac{5}{6}\delta\}, \tag{A.180}$$

where $\mathbb{Z}^+$ denotes the set of nonnegative integers, and let $\bar{s}_k = \xi^{i_1} s_0$. Since

$$d(\boldsymbol{x}_k(0), s_0) = d(\boldsymbol{x}_{k-1}, \boldsymbol{x}_0) \leq \frac{2}{3}\delta, \tag{A.181}$$

(according to Proposition 4.4.1 and Claim 4.4.3), such an integer i_1 must exist. In the second phase, find the smallest nonnegative integer so that the reduction requirement

is fulfilled, i.e.

$$i_2 \triangleq \min\{i \in \mathbb{Z}^+ \mid \tilde{F}(\boldsymbol{x}_k(\xi^i \bar{s}_k)) \leq \tilde{F}(\boldsymbol{x}_{k-1}) - \sigma \xi^i \bar{s}_k \|\nabla \tilde{F}(\boldsymbol{x}_{k-1})\|_F^2\}, \tag{A.182}$$

and let $\eta_k = \xi^{i_2} \bar{s}_k = \xi^{i_1+i_2} s_0$.

Note that the second phase follows the same procedure as the standard Armijo rule (see (1.11) of [86]). Hence the difference between the standard Armijo rule and the restricted Armijo rule can be viewed as the following: in each iteration the former starts from a fixed initial stepsize s while the latter starts from a varying initial stepsize $\bar{s}_k$. We notice that the proof of [86, Proposition 1.2.1] does not require the initial stepsizes to be constant, but rather the following property: if the final stepsize η_k goes to zero for a subsequence $k \in \mathcal{K}$, then for large enough $k \in \mathcal{K}$ the initial stepsize must be reduced at least once (see the remark after (1.17) in [86]). This property also holds when the initial stepsize is lower bounded (asymptotically). In the following, we will prove that for the restricted Armijo rule the initial stepsize $\bar{s}_k$ is lower bounded (asymptotically), and then show how to apply the proof of [86, Proposition 1.2.1] to the restricted Armijo rule.

We first prove that the sequence $\{\bar{s}_k\}$ is lower bounded (asymptotically), i.e.

$$\liminf_{k \to \infty} \bar{s}_k > 0. \tag{A.183}$$

Assume the contrary that $\liminf_{k\to\infty} \bar{s}_k = 0$, i.e. there exists a subsequence $\{\bar{s}_k\}_{k\in\mathcal{K}}$ that converges to zero. Since s_0 is a fixed scalar, we can assume $\bar{s}_k < s_0, \forall k \in \mathcal{K}$, thus the corresponding $i_1 > 0$ for all $k \in \mathcal{K}$. By the definition of i_1 in (A.180), we know that $i_1 - 1$ does not satisfy the distance requirement; in other words, we have

$$d(\boldsymbol{x}_k(\xi^{-1}\bar{s}_k), \boldsymbol{x}_0) > \frac{5}{6}\delta.$$

Denote $g_{k-1} \triangleq \nabla \tilde{F}(\boldsymbol{x}_{k-1})$, then the above relation becomes

$$\frac{5}{6}\delta < d(\boldsymbol{x}_{k-1} - \xi^{-1}\bar{s}_k g_{k-1}, \boldsymbol{x}_0) \leq d(\boldsymbol{x}_{k-1}, x_0) + \xi^{-1}\bar{s}_k \|g_{k-1}\|_F \overset{\text{(A.181)}}{\leq} \frac{2}{3}\delta + \xi^{-1}\bar{s}_k \|g_{k-1}\|_F,$$

implying

$$\frac{1}{6}\xi\delta \leq \bar{s}_k \|g_{k-1}\|_F.$$

Since $\frac{1}{6}\xi\delta$ is a constant and $\{\bar{s}_k\}_{k\in\mathcal{K}}$ converges to zero, the above relation implies that $\{\|g_{k-1}\|_F\}_{k\in\mathcal{K}}$ goes to infinity. However, it is easy to verify that $\|g_{k-1}\|_F =$

$\|\nabla\tilde{F}(\boldsymbol{x}_{k-1})\|_F$ is bounded above by a universal constant when $\|\boldsymbol{x}_{k-1}\|_F \leq \beta_T$ (note that $\|\boldsymbol{x}_{k-1}\|_F \leq \beta_T$ holds due to Proposition 4.4.1 and Claim 4.4.3)), which is a contradiction. Therefore, (A.183) is proved.

Now we prove that each limit point of the sequence $\{\boldsymbol{x}_k\}$ generated by Algorithm 1 with restricted Armijo rule is a stationary point. Assume the contrary that there exists a limit point $\bar{\boldsymbol{x}}$ with $\nabla\tilde{F}(\bar{\boldsymbol{x}}) \neq 0$, and suppose the subsequence $\{\boldsymbol{x}_k\}_{k\in\mathcal{K}}$ converges to $\bar{\boldsymbol{x}}$. By the same argument as that for [86, Proposition 1.2.1], we can prove that the subsequence of final stepsizes $\{\eta_k\}_{k\in\mathcal{K}} \to 0$ (see the inequality before (1.17) in [86]). Since $\{\bar{s}_k\}$ is lower bounded (asymptotically), we must have that $\bar{s}_k > \eta_k,\ \forall\ k \in \mathcal{K}, k \geq \bar{k}$ for large enough $\bar{k}$. Thus the corresponding $i_2 > 0$ for all $k \in \mathcal{K}, k \geq \bar{k}$. By the definition of i_2 in (A.182), we know that $i_2 - 1$ does not satisfy the reduction requirement; in other words, we have $\tilde{F}(\boldsymbol{x}_k(\eta_k\xi^{-1})) > \tilde{F}(\boldsymbol{x}_{k-1}) - \sigma\eta_k\xi^{-1}\|\nabla\tilde{F}(\boldsymbol{x}_{k-1})\|_F^2$, or equivalently,

$$\tilde{F}(\boldsymbol{x}_{k-1}) - \tilde{F}(\boldsymbol{x}_{k-1} - \eta_k\xi^{-1}\nabla\tilde{F}(\boldsymbol{x}_{k-1}))) < \sigma\eta_k\xi^{-1}\|\nabla\tilde{F}(\boldsymbol{x}_{k-1})\|_F^2,\ \forall\ k \in \mathcal{K}, k \geq \bar{k}.$$

This relation is the same as (1.17) in [86] (except that (1.17) in [86] considers a more general descent direction), and the rest of the proof is also the same as [86] and is omitted here.

For Algorithm 1 with stepsize chosen by the restricted line search rule, since it "gives larger reduction in cost at each iteration" than the restricted Armijo rule, it "inherits the convergence properties" of the restricted Armijo rule (as remarked in the last paragraph of the proof of [86, Proposition 1.2.1]). The rigorous proof is similar to that in the second last paragraph of the proof of [86, Proposition 1.2.1]) and is omitted here.

Algorithm 2 is a two-block BCD method to solve problem (P1). According to [98, Corollary 2], each limit point of the sequence generated by Algorithm 2 is a stationary point of problem (P1).

Algorithm 3 belongs to the class of BSUM methods [82]. According to Proposition A.5.1, the level set $\mathcal{X}^0 = \{\boldsymbol{x} \mid \tilde{F}(\boldsymbol{x}) \leq \tilde{F}(\boldsymbol{x}_0)\}$ is a subset of the bounded set $K_1 \cap K_2$, thus $\mathcal{X}^0$ is bounded. Moreover, $\mathcal{X}^0$ is a closed set, thus $\mathcal{X}^0$ is compact. It is easy to verify that the objective function of each subproblem in Algorithm 3 is a convex tight upper bound of $\tilde{F}(\boldsymbol{x})$ (more precisely, satisfies Assumption 2 in [82]). It is also obvious that the objective function of each subproblem is strongly convex, thus each subproblem

of Algorithm 3 has a unique solution. Based on these facts, it follows from [82, Theorem 2] that each limit point of the sequence generated by Algorithm 3 is a stationary point.

Algorithm 4 is a SGD method (or more precisely, incremental gradient method) with a specific stepsize rule. According to (A.175) and (A.177) in Appendix (A.5.4), Algorithm 4 can be viewed as an approximate gradient descent method with bounded error. By [99, Proposition 1], each limit point of the sequence generated by Algorithm 4 is a stationary point.

A.6 Proof of Lemma 4.2.3

Remark: A new section. We will prove a statement that is stronger than Lemma 4.2.1: with probability at least $1 - 1/n^4$, for any $(X, Y) \in K_1 \cap K_2 \cap K(\delta)$ and U, V defined in Table 4.7, we have

$$\langle \nabla_X \tilde{F}(X,Y), X - U\rangle + \langle \nabla_Y \tilde{F}(X,Y), Y - V\rangle \geq \frac{p}{4}d^2 + \frac{2\sqrt{\rho}}{\Sigma_{\min}} d\sqrt{G(X,Y)}, \quad \text{(A.184)}$$

where $d = \|M - XY^T\|_F$.

We have already proved (4.27a), i.e. with probability at least $1 - 1/n^4$,

$$\phi_F = \langle \nabla_X F, X - U\rangle + \langle \nabla_Y F, Y - V\rangle \geq \frac{p}{4}d^2.$$

It remains to prove a bound on ϕ_G, which is stronger than the bound $\phi_G \geq 0$. Note that ϕ_F depends on the observed set Ω, thus the bound on ϕ_F holds with high probability; in contrast, ϕ_G does not depend on Ω, thus the bound on ϕ_G always holds.

Claim A.6.1 *For any $(X, Y) \in K_1 \cap K_2 \cap K(\delta)$ and U, V defined in Table 4.7, we have*

$$\phi_G = \langle \nabla_X G, X - U\rangle + \langle \nabla_Y G, Y - V\rangle \geq \frac{2\sqrt{\rho}}{\Sigma_{\min}} d\sqrt{G(X,Y)}. \quad \text{(A.185)}$$

Proof of Claim A.6.1: By the definition of G in (4.3), $G(X,Y) = \rho(\sum_i G_{1i}(X) + G_2(X) + \sum_j G_{3j}(Y) + G_4(Y))$, where the component functions

$$\begin{aligned} G_{1i}(X) &= G_0\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right), \quad G_2(X) = G_0\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right), \\ G_{3j}(Y) &\triangleq G_0\left(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2}\right), \quad G_4(Y) \triangleq G_0\left(\frac{3\|Y\|_F^2}{2\beta_T^2}\right). \end{aligned} \quad \text{(A.186)}$$

By the expressions of $\nabla_X G, \nabla_Y G$ in (4.14), we have

$$\begin{aligned}
&\phi_G = \langle \nabla_X G, X-U\rangle + \langle \nabla_Y G, Y-V\rangle = \\
&\rho\sum_{i=1}^{m} G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})\frac{3}{\beta_1^2}\langle X^{(i)}, X^{(i)}-U^{(i)}\rangle + \rho G_0'(\frac{3\|X\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle X, X-U\rangle \\
&+\rho\sum_{j=1}^{n} G_0'(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})\frac{3}{\beta_2^2}\langle Y^{(j)}, Y^{(j)}-V^{(j)}\rangle + \rho G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle Y, Y-V\rangle,
\end{aligned} \tag{A.187}$$

where $G_0'(z) = I_{[1,\infty]}(z)2(z-1) = 2\sqrt{G_0(z)}$.

Firstly, we prove

$$h_{1i} \triangleq G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2})\frac{3}{\beta_1^2}\langle X^{(i)}, X^{(i)}-U^{(i)}\rangle \geq \frac{1}{2}\sqrt{G_{1i}(X)},\ \forall\ i, \tag{A.188a}$$

$$h_{3j} \triangleq G_0'(\frac{3\|Y^{(j)}\|^2}{2\beta_2^2})\frac{3}{\beta_2^2}\langle Y^{(j)}, Y^{(j)}-V^{(j)}\rangle \geq \frac{1}{2}\sqrt{G_{3j}(Y)},\ \forall\ j. \tag{A.188b}$$

We only need to prove (A.188a); the proof of (A.188b) is similar. We consider two cases.

Case 1: $\|X^{(i)}\|^2 \leq \frac{2\beta_1^2}{3}$. Note that $\frac{3\|X^{(i)}\|^2}{2\beta_1^2} \leq 1$ implies $G_0(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) = G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) = 0$, thus $h_{1i} = G_{1i} = 0$, in which case (A.188a) holds.

Case 2: $\|X^{(i)}\|^2 > \frac{2\beta_1^2}{3}$. By Corollary 4.3.1 and the fact that $\beta_1^2 = \beta_T^2\frac{3\mu r}{m}$, we have

$$\|U^{(i)}\|^2 \leq \frac{3r\mu}{2m}\beta_T^2 \overset{(4.5)}{=} \frac{3}{4}\frac{2\beta_1^2}{3} < \frac{3}{4}\|X^{(i)}\|^2. \tag{A.189}$$

As a result, $\frac{\sqrt{3}}{2}\langle X^{(i)}, X^{(i)}\rangle = \frac{\sqrt{3}}{2}\|X^{(i)}\|\|X^{(i)}\| > \|X^{(i)}\|\|U^{(i)}\| \geq \langle X^{(i)}, U^{(i)}\rangle$, which implies $\langle X^{(i)}, X^{(i)}-U^{(i)}\rangle \geq (1-\frac{\sqrt{3}}{2})\|X^{(i)}\|^2 > (1-\frac{\sqrt{3}}{2})\frac{2}{3}\beta_1^2 > \frac{1}{12}\beta_1^2$. Combining this inequality with the fact that $G_0'(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}) = 2\sqrt{G_0\left(\frac{3\|X^{(i)}\|^2}{2\beta_1^2}\right)} = 2\sqrt{G_{1i}(X)}$, we get (A.188a).

Secondly, we prove

$$\begin{aligned}
&h_2 + h_4 \geq \frac{2d}{\Sigma_{\min}}\left(\sqrt{G_2(X)} + \sqrt{G_4(Y)}\right), \\
\text{where}\quad &h_2 \triangleq G_0'(\frac{3\|X\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle X, X-U\rangle, \quad h_4 \triangleq G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle Y, Y-V\rangle.
\end{aligned} \tag{A.190}$$

Without loss of generality, we can assume $\|Y\|_F \geq \|X\|_F$, and we will apply Corollary 4.3.1 to prove (A.190). If $\|Y\|_F < \|X\|_F$, we can apply a symmetric result of Corollary 4.3.1 to prove (A.190). We consider three cases.

Case 1: $\|X\|_F \le \|Y\|_F \le \sqrt{\frac{2}{3}}\beta_T$. In this case $G_0(\frac{3\|X\|_F^2}{2\beta_T^2}) = G_0'(\frac{3\|X\|_F^2}{2\beta_T^2}) = G_0(\frac{3\|Y\|_F^2}{2\beta_T^2}) = G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2}) = 0$, which implies $h_2 = h_4 = G_2(X) = G_4(Y) = 0$, thus (A.190) holds.

Case 2: $\|X\|_F \le \sqrt{\frac{2}{3}}\beta_T < \|Y\|_F$. Then we have $\frac{3\|X\|_F^2}{2\beta_T^2} \le 1$, which implies $h_2 = 0 = G_2(X)$. By (4.41d) in Corollary 4.3.1 we have $\|V\|_F \le (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F$, which implies $(1-\frac{d}{\Sigma_{\min}})\langle Y, Y\rangle = (1-\frac{d}{\Sigma_{\min}})\|Y\|_F^2 \ge \|Y\|_F\|V\|_F \ge \langle Y, V\rangle$. This further implies $\langle Y, Y - V\rangle \ge \frac{d}{\Sigma_{\min}}\|Y\|_F^2 \ge \frac{d}{\Sigma_{\min}}\frac{2\beta_T^2}{3}$. Combined with the fact that $G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2}) = 2\sqrt{G_0(\frac{3\|Y\|_F^2}{2\beta_T^2})} = 2\sqrt{G_4(Y)}$, we get

$$h_4 = G_0'(\frac{3\|Y\|_F^2}{2\beta_T^2})\frac{3}{\beta_T^2}\langle Y, Y - V\rangle \ge 2\sqrt{G_4(Y)}\frac{3}{\beta_T^2}\frac{d}{\Sigma_{\min}}\frac{2\beta_T^2}{3} = \frac{4d}{\Sigma_{\min}}\sqrt{G_4(Y)}.$$

Thus

$$\begin{aligned} h_2 + h_4 = h_4 \ge \frac{4d}{\Sigma_{\min}}\sqrt{G_4(Y)} &= \frac{4d}{\Sigma_{\min}}\left(\sqrt{G_4(Y)} + \sqrt{G_2(X)}\right) \\ &\ge \frac{2d}{\Sigma_{\min}}\left(\sqrt{G_4(Y)} + \sqrt{G_2(X)}\right). \end{aligned}$$

Case 3: $\sqrt{\frac{2}{3}}\beta_T < \|X\|_F \le \|Y\|_F$. Since $\|Y\|_F \ge \|X\|_F$, we have $G_4(Y) = G_0\left(\frac{3\|Y\|_F^2}{2\beta_T^2}\right) \ge G_0\left(\frac{3\|X\|_F^2}{2\beta_T^2}\right) = G_2(X)$. By Corollary 4.3.1, we have $\|U\|_F \le \|X\|_F$ and $\|V\|_F \le (1 - \frac{d}{\Sigma_{\min}})\|Y\|_F$. Similar to the argument in Case 2 we can prove $h_2 \ge 0, h_4 \ge \frac{4d}{\Sigma_{\min}}\sqrt{G_4(Y)}$; thus $h_2 + h_4 \ge \frac{4d}{\Sigma_{\min}}\sqrt{G_4(Y)} \ge \frac{2d}{\Sigma_{\min}}\left(\sqrt{G_4(Y)} + \sqrt{G_2(X)}\right)$.

In all three cases, we have proved (A.190), thus (A.190) holds.

We conclude that for U, V defined in Table 4.7,

$$\begin{aligned}
\phi_G &\overset{\text{(A.187)}}{=} \rho\left(\sum_i h_{1i} + \sum_j h_{3j} + h_2 + h_4\right) \\
&\overset{\text{(A.188),(A.190)}}{\geq} \rho\left(\frac{1}{2}\sum_i \sqrt{G_{1i}(X)} + \frac{1}{2}\sum_j \sqrt{G_{2j}(Y)} + \frac{2d}{\Sigma_{\min}}\sqrt{G_2(X)} + \frac{2d}{\Sigma_{\min}}\sqrt{G_4(Y)}\right) \\
&\geq \rho\frac{2d}{\Sigma_{\min}}\left(\sum_i \sqrt{G_{1i}(X)} + \sum_j \sqrt{G_{2j}(Y)} + \sqrt{G_2(X)} + \sqrt{G_4(Y)}\right) \\
&\geq \rho\frac{2d}{\Sigma_{\min}}\sqrt{\sum_i G_{1i}(X) + \sum_j G_{2j}(Y) + G_2(X) + G_4(Y)} \\
&= \rho\frac{2d}{\Sigma_{\min}}\sqrt{\frac{1}{\rho}G(X,Y)} = \frac{2\sqrt{\rho}}{\Sigma_{\min}}d\sqrt{G(X,Y)}.
\end{aligned} \tag{A.191}$$

which finishes the proof of Claim A.6.1. □

Let us come back to the proof of Lemma 4.2.3. The rest of the proof is just algebraic computation. According to (A.184), we have

$$\begin{aligned}
\frac{p}{4}d^2 + \frac{2\sqrt{\rho}}{\Sigma_{\min}}d\sqrt{G(X,Y)} &\leq \langle \nabla_X \tilde{F}(X,Y), X-U\rangle + \langle \nabla_Y \tilde{F}(X,Y), Y-V\rangle \\
&\leq (\|\nabla_X \tilde{F}(X,Y)\|_F + \|\nabla_Y \tilde{F}(X,Y)\|_F)\max\{\|X-U\|_F, \|Y-V\|_F\} \\
&\overset{\text{(4.41b)}}{\leq} \sqrt{2}\sqrt{\|\nabla_X \tilde{F}(X,Y)\|_F^2 + \|\nabla_Y \tilde{F}(X,Y)\|_F^2}\frac{17}{2}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d \\
&= \|\nabla \tilde{F}(X,Y)\|_F\frac{17}{\sqrt{2}}\sqrt{r}\frac{\beta_T}{\Sigma_{\min}}d.
\end{aligned}$$

Eliminating a factor of d from both sides and taking square, we get

$$\|\nabla \tilde{F}(X,Y)\|_F^2\frac{289}{2}r\frac{\beta_T^2}{\Sigma_{\min}^2} \geq \left(\frac{p}{4}d + \frac{2\sqrt{\rho}}{\Sigma_{\min}}\sqrt{G(X,Y)}\right)^2 \geq \frac{pd^2}{16} + \frac{4\rho}{\Sigma_{\min}^2}G(X,Y). \tag{A.192}$$

By the definition of β_T in (4.5), we have

$$r\frac{\beta_T^2}{\Sigma_{\min}^2} = r\frac{C_T r\Sigma_{\max}}{\Sigma_{\min}^2} = C_T\frac{r^2\kappa}{\Sigma_{\min}}.$$

According to Claim 4.2.1, we have

$$pd^2 = p\|M - XY^T\|_F^2 \geq \frac{1}{2}\|\mathcal{P}_\Omega(M - XY^T)\|_F^2 = F(X,Y).$$

By the definition of ρ in (4.7) and the definition of δ_0 in (4.6), we have

$$\frac{4\rho}{\Sigma_{\min}^2} = \frac{4}{\Sigma_{\min}^2} 8p\delta_0^2 = \frac{32p}{\Sigma_{\min}^2} \frac{1}{36} \frac{\Sigma_{\min}^2}{C_d^2 r^3 \kappa^2} = \frac{8}{9} \frac{1}{C_d^2 r^3 \kappa^2} p.$$

Substituting the above three relations into (A.192), we get (when $C_d \geq 32/3$)

$$\begin{aligned} \|\nabla \tilde{F}(X,Y)\|_F^2 \frac{289}{2} C_T \frac{r^2\kappa}{\Sigma_{\min}} &\geq \frac{p}{32} F(X,Y) + \frac{8}{9} \frac{1}{C_d^2 r^3 \kappa^2} pG(X,Y) \\ &\geq \frac{8}{9} \frac{1}{C_d^2 r^3 \kappa^2} p(F(X,Y) + G(X,Y)) = \frac{8}{9} \frac{1}{C_d^2 r^3 \kappa^2} p\tilde{F}(X,Y). \end{aligned}$$

This can be further simplified to

$$\|\nabla \tilde{F}(X,Y)\|_F^2 \geq \frac{\Sigma_{\min}}{C_g r^5 \kappa^3} p\tilde{F}(X,Y),$$

where the numerical constant $C_g = \frac{2601}{16} C_T C_d^2$. This finishes the proof of Lemma 4.2.3.