The Dissertation Committee for Praneeth Kumar Netrapalli
certifies that this is the approved version of the following dissertation:

Provable Alternating Minimization for Non-Convex Learning Problems

Committee:

Sujay Sanghavi, Supervisor

Constantine Caramanis

Inderjit Dhillon

Georgios-Alex Dimakis

Joydeep Ghosh

Pradeep Ravikumar

Provable Alternating Minimization for Non-Convex Learning Problems

by

Praneeth Kumar Netrapalli, B.Tech.; M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2014

To my grandparents.

Acknowledgments

This thesis would not have been possible and my PhD would not have been nearly as enjoyable if it were not for Sujay. I could not have hoped for anything more from an advisor. From the very first day, Sujay encouraged me to find problems that make me passionate. On one hand, I had the freedom to work on any problem that interested me and on the other, when ever I was stuck on a problem, I could walk into Sujay's office to brainstorm. His advice has always been invaluable to me, both in research and career. It is said that students internalize a lot from their advisors – I would be very happy if I have internalized even 1% of his approach and outlook on research.

The starting point of this thesis was my internship with Prateek at MSR India. It is from him that I have learnt how intuitive linear algebra is. The entire work in this thesis has been done in collaboration with him. This thesis would not have been possible with out him.

I would like to thank Sham for having me intern at MSR New England in Summer 2013. The sheer amount of energy he displays is infectious and has an inductive effect on all around him. I would also like to thank Alekh, Anima and Rashish for collaborating on the last part of this thesis.

My stay in Austin has been one of the most enjoyable in my life. I have learnt new things, picked up new interests and greatly enjoyed the time I have spent with Abhik, Ameya, Anish, Bofi, Kumar, Rawat, Reddy, Richa, Sarabjot, Sharayu, Srinadh and Vada. I apologize if I have unwittingly missed some one.

Finally, no words can convey my love and gratitude to my parents, sister and grandparents for loving me as unconditionally as they do. With out their love and support, none of this would have been possible.

Provable Alternating Minimization for Non-Convex Learning Problems

Publication No. _____

Praneeth Kumar Netrapalli, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Sujay Sanghavi

Alternating minimization (AltMin) is a generic term for a widely popular approach in non-convex learning: often, it is possible to partition the variables into two (or more) sets, so that the problem is convex/tractable in one set if the other is held fixed (and vice versa). This allows for alternating between optimally updating one set of variables, and then the other. AltMin methods typically do not have associated global consistency guarantees; even though they are empirically observed to perform better than methods (e.g. based on convex optimization) that do have guarantees.

In this thesis, we obtain rigorous performance guarantees for AltMin in three statistical learning settings: low rank matrix completion, phase retrieval and learning sparsely-used dictionaries. The overarching theme behind our results consists of two parts: (i) devising new initialization procedures (as opposed to doing so randomly, as is typical), and (ii) establishing exponential local convergence from this initialization. Our work shows that the pursuit of statistical guarantees can yield algorithmic improvements (initialization in our case) that perform better in practice.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A general description of a learning problem is as follows: There is an underlying model with unknown parameters. We obtain independent samples distributed according to this model. The goal is to estimate the model from those samples. The tremendous amount of increase in both amount and variety of data available over the last decade has resulted in a huge amount of interest in learning problems in the high dimensional regime. The high dimensional regime refers to the scenario where the number of unknown parameters is much larger than the number of samples available. However, in such cases, a variety of assumptions on the unknown parameters such as sparsity, low-rank etc. arise naturally. When posed as optimization problems, most of these assumptions result in non-convex constraints. A major theme in machine learning over the last decade has been the use of convex relaxations to provably solve such non-convex learning problems. Lasso [19, 16] for compressed sensing, trace norm minimization for matrix completion [15, 20, 76] and robust PCA [22, 14] are some of the prototypical examples of this approach.

Any algorithm for such a high dimensional learning problem is evaluated on two counts: statistical complexity (or sample complexity) and computational complexity. Statistical complexity refers to the number of examples (or samples) required by an algorithm for consistent recovery of the model. Computational complexity, on the other hand, refers to the time taken by the algorithm. Though convex relaxation methods such as lasso, trace norm minimization etc. are known to have good (in many cases optimal) statistical complexity, their computational complexity is high and hence, they do not scale well to large scale problems. Moreover, there are no known ways of implementing most of these algorithms in a distributed fashion.

To overcome this problem, researchers have come up with efficient heuristics that scale well to large problem sizes and have good performance

Algorithm 1 AltMinGeneral

input Function $f(\cdot, \cdot)$

Choose U^0
for $t \leftarrow 1, \cdots, T$ **do**
  $V^t \leftarrow \arg\min_V f(U^{t-1}, V)$
  $U^t \leftarrow \arg\min_U f(U, V^t)$
end for

output (U^T, V^T)

(i.e., statistical complexity) in practice. In spite of the success of such heuristics on real world data, there have been very few theoretical guarantees for such heuristics. The only exception is compressed sensing, for which a variety of non-convex methods have been shown to work [83, 94, 34]. To summarize, for many problems such as matrix sensing, matrix completion, robust PCA, though there are heuristics that work quite well in practice, till date the only methods with theoretical guarantees are the ones based on convex relaxation.

1.1 Alternating Minimization

In many non-convex inference problems, it turns out that it is possible to partition the variables into two (or more) sets such that the problem is convex in one set if we fix the other set. Alternating minimization is a heuristic for such problems where we optimize one set of variables while holding the other fixed and vice versa. Algorithm 1 gives the pseudocode of the algorithm.

AltMin is widely used and forms the basis of many popular algorithms: k-means for learning mixtures of Gaussians [91], Netflix prize winning BellKor algorithm [53] and so on. In spite of its empirical success, till date there are very few results on its performance in any setting [4, 48].

1.2 Our Contributions

In this dissertation, we make progress in addressing the above issue: i.e., we prove guarantees on the performance of alternating minimization for

three machine learning problems. Our results are as follows:

- **Matrix completion using alternating minimization**: In Chapter 2, we obtain statistical guarantees for alternating minimization as applied to the matrix completion problem. The matrix completion problem is, given partial entries of a matrix, to fill in the remaining ones under the assumption that the matrix is low-rank. We show that if the underlying matrix is incoherent and the samples are drawn uniformly at random from among all the entries, then the statistical complexity of alternating minimization is $O\left(k^7 n \log n\right)$ where k is the rank of the underlying $n \times n$ matrix. Further more, we show linear convergence of the estimate matrix to the underlying matrix.

- **Phase retrieval using alternating minimization**: The phase retrieval problem is to recover a complex n-dimensional signal using linear magnitude measurements. In Chapter 3, we show that $O\left(n \log^2 n\right)$ Gaussian magnitude measurements are sufficient to recover the underlying signal using alternating minimization with high probability. Further more, we show linear convergence of the estimate vector to the underlying vector.

- **Learning Sparsely Used Dictionaries**: In Chapter 4, we consider the problem of learning sparsely used dictionaries, where, given examples (which are vectors in $\mathbb{R}^d$), we wish to find a set of dictionary elements such that each example has a sparse representation as a linear combination of very few dictionary elements. For the case of incoherent dictionaries and sufficiently sparse representations, we present an approximate recovery algorithm and show that alternating minimization followed by this approximate recovery step succeeds in recovering the underlying dictionary if the number of examples is larger than $O\left(r^2\right)$.

The rest of the document is organized as follows: In Chapter 2 we present our results on matrix completion using alternating minimization and in Chapter 3, we present our results on phase retreival using alternating minimization. In Chapter 4, we present our results on the problem of learning sparsely used dictionaries. We conclude in Chapter 5. Most of the technical results are deferred to the appendices.

Chapter 2

Matrix Completion using Alternating Minimization

2.1 Introduction

Finding [1] a low-rank matrix to fit / approximate observations is a fundamental task in data analysis. In a slew of applications, a popular empirical approach has been to represent the target rank k matrix $X \in \mathbb{R}^{m \times n}$ in a *bi-linear form* $X = UV^\dagger$, where $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$. Typically, this is done for two reasons:
(a) Size and computation: If the rank k of the target matrix (to be estimated) is much smaller than m, n, then U, V are significantly smaller than X and hence are more efficient to optimize for. This is crucial for several practical applications, e.g., recommender systems where one routinely encounters matrices with billions of entries.
(b) Modeling: In several applications, one would like to impose extra constraints on the target matrix, besides just low rank. Oftentimes, these constraints might be easier and more natural to impose on factors U, V. For example, in Sparse PCA [96], one looks for a low-rank X that is the product of *sparse* U and V.

Due to the above two reasons, in several applications, the target matrix X is parameterized by $X = UV^\dagger$. For example, clustering [52], sparse PCA [96] etc.

Using the bi-linear parametrization of the target matrix X, the task of estimating X now reduces to finding U and V that, for example, minimize

[1] An extended abstract of the results in this chapter appeared as [45]. The coauthors on the paper had equal contributions in obtaining these results.

an error metric. The resulting problem is typically non-convex due to bi-linearity. Correspondingly, a popular approach has been to use alternating minimization: iteratively keep one of U, V fixed and optimize over the other, then switch and repeat, see e.g. [54]. While the overall problem is non-convex, each sub-problem is typically convex and can be solved efficiently.

Despite wide usage of bi-linear representation and alternating minimization, there has been to date almost no theoretical understanding of when such a formulation works. Motivated by this disconnect between theory and practice in the estimation of low-rank matrices, in this chapter, we provide one of the first guarantees for performance of alternating minimization, for two low-rank matrix recovery problems: matrix completion, and matrix sensing.

Matrix completion involves completing a low-rank matrix, by observing only a few of its elements. Its recent popularity, and primary motivation, comes from recommendation systems [54], where the task is to complete a user-item ratings matrix using only a small number of ratings. As elaborated in Section 2.3, alternating minimization becomes particularly appealing for this problem as it provides a fast, distributed algorithm that can exploit both sparsity of ratings as well as the low-rank bi-linear parametrization of X.

Matrix sensing refers to the problem of recovering a low-rank matrix $M \in \mathbb{R}^{m \times n}$ from affine equations. That is, given d linear measurements $b_i = tr(A_i^\dagger M)$ and measurement matrices A_i's, the goal is to recover back M. This problem is particularly interesting in the case of $d \ll mn$ and was first studied in [77] and subsequently in [44, 56]. In fact, matrix completion is a special case of this problem, where each observed entry in the matrix completion problem represents one single-element measurement matrix A_i.

Without any extra conditions, both matrix sensing and matrix completion are ill-posed problems, with potentially multiple low-rank solutions, and are in general NP hard [66]. Current work on these problems thus impose some extra conditions, which makes the problems both well defined, and amenable to solution via the respective proposed algorithms [77, 15]. In this chapter, we show that under similar conditions to the ones used by the existing methods, alternating minimization also guarantees recovery of the true matrix; we also show that it requires only a small number of computationally cheap iterations

and hence, as observed empirically, is computationally much more efficient than the existing methods.

Notations: We represent a matrix by capital letter (e.g. M) and a vector by small letter (u). u_i represents i-th element of u and U_{ij} denotes (i,j)-th entry of U. U_i represents i-th column of U and $U^{(i)}$ represents i-th row of U. $A^\dagger$ denotes matrix transpose of A. $u = vec(U)$ represents vectorized U, i.e., $u = [U_1^\dagger \; U_2^\dagger \; \dots \; U_k^\dagger]^\dagger$. $\|u\|_p$ denotes L_p norm of u, i.e., $\|u\|_p = (\sum_i |u_i|^p)^{1/p}$. By default, $\|u\|$ denotes L_2 norm of u. $\|A\|_F$ denotes Frobenius norm of A, i.e., $\|vec(A)\|_2$. $\|A\|_2 = \max_{x,\|x\|_2=1} \|Ax\|_2$ denotes spectral norm of A. $tr(A)$ denotes the trace (sum of diagonal elements) of square matrix A. Typically, $\widehat{U}$, $\widehat{V}$ represent factor matrices (i.e., $\widehat{U} \in \mathbb{R}^{m\times k}$ and $\widehat{V} \in \mathbb{R}^{n\times k}$) and U, V represent their orthonormal basis.

2.2 Related Work

Alternating Minimization: Alternating minimization and its variants have been applied to several low-rank matrix estimation problems. For example, clustering [52], sparse PCA [96], non-negative matrix factorization [51], signed network prediction [41] etc. There are three main reasons for such wide applicability of this approach: a) low-memory footprint and fast iterations, b) flexible modeling, c) amenable to parallelization. However, despite such empirical success, this approach has largely been used as a heuristic and has had no theoretical analysis other than the guarantees of convergence to the *local minima* [93].

After this work was completed, we became aware of [49] which provides an analysis of alternating minimization for matrix completion. Along with [49], ours is the first analysis of this approach for the problem of matrix completion. Moreover, ours is the first analysis of this approach for the problem of matrix sensing. .

Matrix Completion: This is the problem of completing a low-rank matrix from a few sampled entries. Candes and Recht [15] provided the first results on this problem, showing that under the random sampling and incoherence conditions (detailed above), $O(kn^{1.2}\log n)$ samples allow for recovery via convex trace-norm minimization; this was improved to $O(kn\log n)$ in [20]. For

large matrices, this approach is not very attractive due to the need to store and update the entire matrix, and because iterative methods for trace norm minimization require $O(\frac{1}{\sqrt{\epsilon}})$ steps to achieve additive error of ϵ. Moreover, each such step needs to compute an SVD.

Another approach, in [50], involved taking a single SVD, followed by gradient descent on a Grassmanian manifold. However, *(a)* this is more expensive than alternating minimization as it needs to compute gradient over Grassmanian manifold which in general is a computationally intensive step, and *(b)* the analysis of the algorithm only guarantees asymptotic convergence, and in the worst case might take exponential time in the problem size.

The most closely related work to ours is [49], which provides guarantees for alternating minimization for the case of matrix completion. [49] shows that consistent recovery is possible if the sampling probability p scales as $\Omega\left(k\left(\frac{\sigma_1^*}{\sigma_k^*}\right)^8 \frac{\log n}{m}\right)$. Our result is worse than theirs in the dependence on k while being better in the dependence on the condition number.

Recently, several other matrix completion type of problems have been studied in the literature. For example, robust PCA [22, 14], spectral clustering [46] etc. Here again, under additional assumptions, convex relaxation based methods have rigorous analysis but alternating minimization based algorithms continue to be algorithms of choice in practice.

Matrix Sensing: The general problem of matrix sensing was first proposed by [77]. They established recovery via trace norm minimization, assuming the sensing operator satisfies "restricted isometry" conditions. Subsequently, several other methods [44, 56] were proposed for this problem that also recovers the underlying matrix with optimal number of measurements and can give an ϵ-additive approximation in time $O(\log(1/\epsilon)$. But, similar to matrix completion, most of these methods require computing SVD of a large matrix at each step and hence have poor scalability to large problems.

We show that AltMinSense and AltMin-Completion provide more scalable algorithms for their respective problems. We demonstrate that these algorithms have geometric convergence to the optima, while each iteration is relatively cheap. For this, we assume conditions similar to those required by existing algorithms; albeit, with one drawback: number of samples required by

our analysis depend on the condition number of the underlying matrix M. For the matrix sensing problem, we remove this requirement by using a stagewise algorithm; we leave similar analysis for matrix completion as an open problem.

2.3 Our Results

In this section, we will first define the matrix sensing problem, and present our results for it. Subsequently, we will do the same for matrix completion. The matrix sensing setting – i.e. recovery of any low-rank matrix from linear measurements that satisfy matrix RIP – represents an easier analytical setting than matrix completion, but still captures several key properties of the problem that helps us in developing an analysis for matrix completion. We note that for either problem, ours represent one of the first global optimality guarantees for alternating minimization based algorithms. Due to lack of space, we do not present the proofs of these results in this document. Please refer [45] for complete proofs of all the results in this chapter.

Matrix Sensing via Alternating Minimization

Given d linear measurements $b_i = \langle M, A_i \rangle = tr(A_i^\dagger M),\ 1 \leq i \leq d$ of an *unknown* rank-k matrix $M \in \mathbb{R}^{m \times n}$ and the sensing matrices $A_i, 1 \leq i \leq d$, the goal in matrix sensing is to recover back M. In the following we collate these coefficients, so that $b \in \mathbb{R}^d$ is the vector of b_i's, and $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \to d$ is the corresponding linear map, with $b = \mathcal{A}(M)$. With this notation, the Low-Rank Matrix Sensing problem is:

$$\text{Find } X \in \mathbb{R}^{m \times n},\ \ s.t\ \ \mathcal{A}(X) = b,\ \ \ rank(X) \leq k. \tag{LRMS}$$

As in the existing work [77] on this problem, we are interested in the **underdetermined case**, where $d < mn$. Note that this problem is a strict generalization of the popular compressed sensing problem [18]; compressed sensing represents the case when M is restricted to be a diagonal matrix.

For matrix sensing, alternating minimization approach involves representing X as a product of two matrices $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$, i.e., $X = UV^\dagger$. If k is (much) smaller than m, n, these matrices will be (much)

smaller than X. With this bi-linear representation, alternating minimization can be viewed as an approximate way to solve the following non-convex optimization problem:

$$\min_{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}} \quad \|\mathcal{A}(UV^\dagger) - b\|_2^2$$

As mentioned earlier, alternating minimization algorithm for matrix sensing now alternately solves for U and V while fixing the other factor. See Algorithm 2 for a pseudo-code of AltMinSense algorithm that we analyze.

We note two key properties of AltMinSense : a) Each minimization – over U with V fixed, and vice versa – is a simple least-squares problem, which can be solved in time $O(dn^2k^2 + n^3k^3)$[2], b) We initialize U^0 to be the top-k left singular vectors of $\sum_i A_i b_i$ (step 2 of Algorithm 2). This provides a good initialization point for the sensing problem which is crucial; if the first iterate $\widehat{U}^0$ is orthogonal, or almost orthogonal, to the true U^* subspace, AltMinSense may never converge to the true space (this is easy to see in the simplest case, when the map is identity, i.e. $\mathcal{A}(X) = X$ – in which case AltMinSense just becomes the power method).

Algorithm 2 AltMinSense : Alternating minimization for matrix sensing

1: Input $b, \mathcal{A}$
2: Initialize $\widehat{U}^0$ to be the top-k left singular vectors of $\sum_i A_i b_i$
3: **for** $t = 0, \cdots, T-1$ **do**
4: $\quad \widehat{V}^{t+1} \leftarrow \arg\min_{V \in \mathbb{R}^{n \times k}} \ \|\mathcal{A}(\widehat{U}^t V^\dagger) - b\|_2^2$
5: $\quad \widehat{U}^{t+1} \leftarrow \arg\min_{U \in \mathbb{R}^{m \times k}} \ \|\mathcal{A}(U (\widehat{V}^{t+1})^\dagger) - b\|_2^2$
6: **end for**
7: Return $X = \widehat{U}^T (\widehat{V}^T)^\dagger$

In general, since $d < mn$, problem (LRMS) is not well posed as there can be multiple rank-k solutions that satisfy $\mathcal{A}(X) = b$. However, inspired by a similar condition in compressed sensing [18], Recht et al. [77] showed that if the linear map $\mathcal{A}$ satisfies a *(matrix) restricted isometry property* (RIP), then a trace-norm based convex relaxation of (LRMS) leads to exact recovery. This property is defined below.

[2]Throughout this chapter, we assume $m \leq n$.

Definition 2.3.1. [77] A linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^d$ is said to satisfy k-RIP, with δ_k RIP constant, if for all $X \in \mathbb{R}^{m\times n}$ s.t. $\text{rank}(X) \leq k$, the following holds:

$$(1-\delta_k)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1+\delta_k)\|X\|_F^2 . \tag{1}$$

Several random matrix ensembles with sufficiently many measurements (d) satisfy matrix RIP [77]. For example, if $d = \Omega(\frac{1}{\delta_k^2} kn \log n)$ and each entry of A_i is sampled i.i.d. from a 0-mean sub-Gaussian distribution then k-RIP is satisfied with RIP constant δ_k.

We now present our main result for AltMinSense.

Theorem 2.3.1. *Let $M = U^*\Sigma^*V^{*\dagger}$ be a rank-k matrix with non zero singular values $\sigma_1^* \geq \sigma_2^* \cdots \geq \sigma_k^*$. Also, let the linear measurement operator $\mathcal{A}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^d$ satisfy $2k$-RIP with RIP constant $\delta_{2k} < \frac{(\sigma_k^*)^2}{(\sigma_1^*)^2}\frac{1}{100k}$. Then, in the AltMinSense algorithm (Algorithm 2), for all $T > 2\log(\|M\|_F/\epsilon)$, the iterates $\widehat{U}^T$ and $\widehat{V}^T$ satisfy:*

$$\|M - \widehat{U}^T(\widehat{V}^T)^\dagger\|_F \leq \epsilon.$$

The above theorem establishes geometric convergence (in $O(\log(1/\epsilon))$ steps) of AltMinSense to the optimal solution of (LRMS) under standard RIP assumptions. This is in contrast to existing iterative methods for trace-norm minimization all of which require at least $O(\frac{1}{\sqrt{\epsilon}})$ steps; interior point methods for trace-norm minimization converge to the optimum in $O(\log(1/\epsilon))$ steps but require storage of the full $m \times n$ matrix and require $O(n^5)$ time per step, which makes it infeasible for even moderate sized problems.

Recently, several projected gradient based methods have been developed for matrix sensing [44, 56] that also guarantee convergence to the optimum in $O(\log(1/\epsilon))$ steps. But each iteration in these algorithms requires computation of the top k singular components of an $m \times n$ matrix, which is typically significantly slower than solving a least squares problem (as required by each iteration of AltMinSense).

Stagewise AltMinSense Algorithm: A drawback of our analysis for AltMinSense is the dependence of δ_{2k} on the condition number ($\kappa = \frac{\sigma_1^*}{\sigma_k^*}$) of M, which implies that the number of measurements d required by AltMinSense

Algorithm 3 Stage-AltMin: Stagewise Alternating Minimization for Matrix Sensing

Input: $b, \mathcal{A}$
$\widehat{U}^T \leftarrow []$, $\widehat{V}^T \leftarrow []$
for $i = 1, \cdots, k$ **do**
  $[\widehat{U}^0_{1:i} \quad \widehat{V}^0_{1:i}]$ = top i-singular vectors of $\Big(\widehat{U}^T_{1:i-1}(\widehat{V}^T_{1:i-1})^\dagger - \frac{3}{4}\mathcal{A}^T(\mathcal{A}(\widehat{U}^T_{1:i-1}(\widehat{V}^T_{1:i-1})^\dagger) - b)\Big)$ i.e., one step of SVP [44]
  for $t = 0, \cdots, T-1$ **do**
    $\widehat{V}^{t+1}_{1:i} \leftarrow \arg\min_{V \in \mathbb{R}^{n\times i}} \ \|\mathcal{A}(\widehat{U}^t_{1:i}V^\dagger) - b\|_2^2$
    $\widehat{U}^{t+1}_{1:i} \leftarrow \arg\min_{U \in \mathbb{R}^{m\times i}} \ \|\mathcal{A}(U_{1:i}(\widehat{V}^{t+1}_{1:i})^\dagger) - b\|_2^2$
  end for
end for
Output: $X = \widehat{U}^T_{1:i}(\widehat{V}^T_{1:i})^\dagger$

grows quadratically with κ. We address this issue by using a stagewise version of AltMinSense (Algorithm 3) for which we are able to obtain near optimal measurement requirement.

The key idea behind our stagewise algorithm is that if one of the singular vectors of M is very dominant, then we can treat the underlying matrix as a rank-1 matrix plus noise and approximately recover the top singular vector. Once we remove this singular vector from the measurements, we will have a relatively well-conditioned problem. Hence, at each stage of Algorithm 3, we seek to remove the remaining most dominant singular vector of M. The main result regarding the performance of Stage-AltMin is stated in the following theorem.

Theorem 2.3.2. *Let $M = U^*\Sigma^*V^{*^\dagger}$ be a rank-k incoherent matrix with non zero singular values $\sigma_1^* \geq \sigma_2^* \cdots \geq \sigma_k^*$. Also, let $\mathcal{A}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^d$ be a linear measurement operator that satisfies $2k$-RIP with RIP constant $\delta_{2k} < \frac{1}{3200k^2}$. Suppose, Stage-AltMin (Algorithm 3) is supplied inputs $\mathcal{A}$, $b = \mathcal{A}(M)$. Then, the i-th stage iterates $\widehat{U}^T_{1:i}$, $V^T_{1:i}$ satisfy:*

$$\|M - \widehat{U}^T_{1:i}\left(V^T_{1:i}\right)^\dagger\|_F^2 \leq \max(\epsilon, 16k(\sigma^*_{i+1})^2),$$

where $T = \Omega\left(\log(\|M\|_F^2/\epsilon)\right)$. *That is, the* T*-th step iterates of the* k*-th stage, satisfy:* $\|M - \widehat{U}_{1:k}^T \left(V_{1:k}^T\right)^\dagger \|_F^2 \leq \epsilon$.

The above theorem guarantees exact recovery using $O(k^4 n \log n)$ measurements which is only $O(k^3)$ worse than the information theoretic lower bound. We also note that for simplicity of analysis, we did not optimize the constant factors in δ_{2k}.

Matrix Completion via Alternating Minimization

The matrix completion problem is the following: there is an unknown rank-k matrix $M \in \mathbb{R}^{m\times n}$, of which we know a set $\Omega \subset [m] \times [n]$ of elements; that is, we know the values of elements M_{ij}, for $(i,j) \in \Omega$. The task is to recover M. Formally, the Low-Rank Matrix Completion problem is:

$$\text{Find rank-}k \text{ matrix } X \text{ s.t. } P_\Omega(X) = P_\Omega(M), \tag{LRMC}$$

where for any matrix S and a set of elements $\Omega \subset [m] \times [n]$ the matrix $P_\Omega(S) \in \mathbb{R}^{m\times n}$ is as defined below:

$$P_\Omega(S)_{ij} = \begin{cases} S_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We are again interested in the under-determined case; in fact, for a fixed rank k, as few as $O(n \log n)$ elements may be observed. This problem is a special case of matrix sensing, with the measurement matrices $A_i = e_j e_\ell^\dagger$ being non-zero only in single elements; however, such matrices do not satisfy matrix RIP conditions like (1). For example, consider a low-rank $M = e_1 e_1^\dagger$ for which a uniformly random Ω of size $O(n \log n)$ will most likely miss the non-zero entry of M.

Nevertheless, like matrix sensing, matrix completion has been shown to be possible once additional conditions are applied to the low-rank matrix M and the observation set Ω. Starting with the first work [15], the typical assumption has been to have Ω generated uniformly at random, and M to satisfy a particular incoherence property that, loosely speaking, makes it very

far from a sparse matrix. In this chapter, we show that *once* such assumptions are made, alternating minimization *also* succeeds. We now restate, and subsequently use, this incoherence definition.

Definition 2.3.2. [15] A matrix $M \in \mathbb{R}^{m \times n}$ is incoherent with parameter μ if:

$$\left\|u^{(i)}\right\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{m}} \; \forall \; i \in [m], \; \left\|v^{(j)}\right\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}} \; \forall \; j \in [n], \tag{3}$$

where $M = U\Sigma V^T$ is the SVD of M and $u^{(i)}$, $v^{(j)}$ denote the i^{th} row of U and the j^{th} row of V respectively.

The alternating minimization algorithm can be viewed as an approximate way to solve the following non-convex problem:

$$\min_{U,V \in \mathbb{R}^{n \times k}} \quad \|P_\Omega(UV^\dagger) - P_\Omega(M)\|_F^2$$

Similar to AltMinSense, the altmin procedure proceeds by alternatively solving for U and V. As noted earlier, this approach has been popular in practice and has seen several variants and extensions being used in practice [95, 54, 53, 23]. However, for ease of analysis, our algorithm further modifies the standard alternating minimization method. In particular, we introduce partitioning of the observed set Ω, so that we use different partitions of Ω in each iteration. See Algorithm 4 for a pseudo-code of our variant of the alternating minimization approach.

Our use of some technical lemmas from [50] renders all the constants dependent on $\frac{n}{m}$. In what follows, a constant by default is assumed to depend on $\frac{n}{m}$. We believe that our results hold even with out this assumption but proving this seems to take a little more work. We now present our main result for (LRMC):

Theorem 2.3.3. *Let $M = U^*\Sigma^*V^{*\dagger} \in \mathbb{R}^{m \times n}$ $(n \geq m)$ be a rank-k incoherent matrix, i.e., both U^* and V^* are μ-incoherent (see Definition 2.3.2). Also, let each entry of M be observed uniformly and independently with probability,*

$$p > C\frac{\left(\frac{\sigma_1^*}{\sigma_k^*}\right)^4 \mu^4 k^7 \log n \log \frac{k\|M\|_F}{\epsilon}}{m\delta_{2k}^2},$$

Algorithm 4 AltMinComplete: Alternating minimization for matrix completion

1: Input: observed set Ω, values $P_\Omega(M)$
2: Partition Ω into $2T+1$ subsets $\Omega_0, \cdots, \Omega_{2T}$ with each element of Ω belonging to one of the Ω_t with equal probability (sampling with replacement)
3: $\widehat{U}^0 = SVD(\frac{1}{p}P_{\Omega_0}(M), k)$ i.e., top-k left singular vectors of $\frac{1}{p}P_{\Omega_0}(M)$
4: Clipping step : Set all elements of $\widehat{U}^0$ that have magnitude greater than $\frac{2\mu\sqrt{k}}{\sqrt{n}}$ to zero and orthonormalize the columns of $\widehat{U}^0$
5: **for** $t = 0, \cdots, T-1$ **do**
6: $\quad \widehat{V}^{t+1} \leftarrow \arg\min_{V \in \mathbb{R}^{n\times k}} \ \|P_{\Omega_{t+1}}(\widehat{U}^t V^\dagger - M)\|_F^2$
7: $\quad \widehat{U}^{t+1} \leftarrow \arg\min_{U \in \mathbb{R}^{m\times k}} \ \|P_{\Omega_{T+t+1}}(U\left(\widehat{V}^{t+1}\right)^\dagger - M)\|_F^2$
8: **end for**
9: Return $X = \widehat{U}^T(\widehat{V}^T)^\dagger$

where $\delta_{2k} \leq \frac{\sigma_k^*}{C\sigma_1^*}$ *and* $C > 0$ *is a global constant. Then w.h.p. for* $T = C' \log \frac{\|M\|_F}{\epsilon}$*, the outputs* $\widehat{U}^T$ *and* V^T *of Algorithm 4, with input* $(\Omega, P_\Omega(M))$ *(see Equation* (2)*) satisfy:* $\left\|M - \widehat{U}^T\left(V^T\right)^\dagger\right\|_F \leq \epsilon$.

The above theorem implies that by observing $|\Omega| = O\left((\frac{\sigma_1^*}{\sigma_k^*})^6 k^7 n \log n \log(k\|M\|_F/\epsilon)\right)$ random entries of an incoherent M, AltMinComplete can recover M in $O(\log(1/\epsilon))$ steps. In terms of sample complexity ($|\Omega|$), our results show alternating minimization may require a bigger Ω than convex optimization, as our result has $|\Omega|$ depend on the condition number, required accuracy (ϵ) and worse dependence on k than known bounds. In contrast, trace-norm minimization based methods require $O(kn \log n)$ samples only.

Empirically however, this is not seen to be the case – see Section 2.7.

In terms of time complexity, we show that AltMinComplete needs time $O(|\Omega|k^2 \log(1/\epsilon))$. This is in contrast to popular trace-norm minimization based methods that need $O(1/\sqrt{\epsilon})$ steps [10] and total time complexity of $O(|\Omega|n/\sqrt{\epsilon})$; note that the latter can be potentially quadratic in n. Furthermore, each step of such methods requires computation of the SVD of an $m \times n$

matrix. As mentioned earlier, interior point methods for trace-norm minimization also converge in $O(\log(1/\epsilon))$ steps but each iteration requires $O(n^5)$ steps and need storage of the entire $m \times n$ matrix X.

2.4 Matrix Sensing

In this section, we study the matrix sensing problem (LRMS) and prove that if the measurement operator, $\mathcal{A}$, satisfies RIP then AltMinSense (Algorithm 2) recovers the underlying low-rank matrix *exactly* (see Theorem 2.3.1).

At a high level, we prove Theorem 2.3.1 by showing that the "distance" between subspaces spanned by $\widehat{V}^t$ (iterate at time t) and V^* decreases exponentially with t. This done based on the observation that once the (standard) matrix RIP condition (Definition 2.3.1) holds, alternating minimization can be viewed, and analyzed, as a **perturbed version of the power method.** This is easiest to see for the rank-1 case below; we detail this proof, and then the more general rank-k case.

In this paper, we use the following definition of distance between subspaces:

Definition 2.4.1. [38] Given two matrices $\widehat{U}, \widehat{W} \in \mathbb{R}^{m \times k}$, the (principal angle) distance between the subspaces spanned by the columns of $\widehat{U}$ and $\widehat{W}$ is given by:

$$\text{dist}\left(\widehat{U}, \widehat{W}\right) \stackrel{\text{def}}{=} \left\| U_\perp^\dagger W \right\|_2 = \left\| W_\perp^\dagger U \right\|_2$$

where U and W are orthonormal bases of the spaces Span $\left(\widehat{U}\right)$ and Span $\left(\widehat{W}\right)$, respectively. Similarly, $U_\perp$ and $W_\perp$ are any orthonormal bases of the perpendicular spaces Span $(U)^\perp$ and Span $(W)^\perp$, respectively.

Note: *(a)* The distance depends only on the spaces spanned by the columns of $\widehat{U}, \widehat{W}$, *(b)* if the ranks of $\widehat{U}$ and $\widehat{W}$ (i.e. the dimensions of their spans) are not equal, then dist $\left(\widehat{U}, \widehat{W}\right) = 1$, and *(c)* dist $\left(\widehat{U}, \widehat{W}\right) = 0$ if and only if they span the same subspace of $\mathbb{R}^m$.

We now present a theorem that bounds the distance between the subspaces spanned by $\widehat{V}^t$ and V^* and show that it decreases exponentially with t.

Theorem 2.4.1. *Let $b = \mathcal{A}(M)$ where M and $\mathcal{A}$ satisfy assumptions given in Theorem 2.3.1. Then, the $(t+1)$-th iterates $\widehat{U}^{t+1}$, $\widehat{V}^{t+1}$ of AltMinSense satisfy:*

$$\begin{aligned}\text{dist}\left(\widehat{V}^{t+1}, V^*\right) &\leq \frac{1}{4} \cdot \text{dist}\left(\widehat{U}^t, U^*\right) \ , \\ \text{dist}\left(\widehat{U}^{t+1}, U^*\right) &\leq \frac{1}{4} \cdot \text{dist}\left(\widehat{V}^{t+1}, V^*\right)\end{aligned}$$

where $\text{dist}\,(U, W)$ *denotes the principal angle based distance (see Definition 2.4.1).*

Using Theorem 2.4.1, we are now ready to prove Theorem 2.3.1.

Proof Of Theorem 2.3.1. Assuming correctness of Theorem 2.4.1, Theorem 2.3.1 follows by using the following set of inequalities:

$$\begin{aligned}\|M - \widehat{U}^T(\widehat{V}^T)^\dagger\|_F^2 &\overset{\zeta_1}{\leq} \frac{1}{1-\delta_{2k}}\|\mathcal{A}(M - \widehat{U}^T(\widehat{V}^T)^\dagger)\|_2^2, \\ &\overset{\zeta_2}{\leq} \frac{1}{1-\delta_{2k}}\|\mathcal{A}(M(I - V^T(V^T)^\dagger))\|_2^2, \\ &\overset{\zeta_3}{\leq} \frac{1+\delta_{2k}}{1-\delta_{2k}}\|U^*\Sigma^*(V^*)^\dagger(I - V^T(V^T)^\dagger))\|_F^2, \\ &\overset{\zeta_4}{\leq} \frac{1+\delta_{2k}}{1-\delta_{2k}}\|M\|_F^2 \text{dist}^2\left(V^T, V^*\right) \overset{\zeta_5}{\leq} \epsilon,\end{aligned}$$

where V^T is an orthonormal basis of $\widehat{V}^T$, ζ_1 and ζ_3 follow by RIP, ζ_2 holds as $\widehat{U}^T$ is the least squares solution, ζ_4 follows from the definition of $\text{dist}(\cdot,\cdot)$ and finally ζ_5 follows from Theorem 2.4.1 and by setting T appropriately. □

To complete the proof of Theorem 2.3.1, we now need to prove Theorem 2.4.1. In the next section, we illustrate the main ideas of the proof of Theorem 2.4.1 by applying it to a rank-1 matrix i.e., when $k = 1$. We then provide a proof of Theorem 2.4.1 for arbitrary k in Section 2.4.2.

2.4.1 Rank-1 Case

In this section, we provide a proof of Theorem 2.4.1 for the special case of $k = 1$. That is, let $M = u^*\sigma^*(v^*)^\dagger$ s.t. $u^* \in \mathbb{R}^m$, $\|u^*\|_2 = 1$ and $v^* \in \mathbb{R}^n, \|v^*\|_2 = 1$. Also note that when $\widehat{u}$ and $\widehat{w}$ are vectors, $\text{dist}(\widehat{u}, \widehat{w}) = 1 - (u^\dagger w)^2$, where $u = \widehat{u}/\|\widehat{u}\|_2$ and $w = \widehat{w}/\|\widehat{w}\|_2$.

Consider the t-th update step in the AltMinSense procedure. As $\widehat{v}^{t+1} = \arg\min_{\widehat{v}} \sum_{i=1}^d \left(\widehat{u}^{t\dagger} A_i^\dagger \widehat{v} - \sigma^* u^{*\dagger} A_i^\dagger v^*\right)^2$, setting the gradient of the above objective function to 0, we obtain:

$$\left(\sum_{i=1}^d A_i u^t (u^t)^\dagger A_i^\dagger\right) \|\widehat{u}^t\|_2 \widehat{v}^{t+1} = \sigma^* \left(\sum_{i=1}^d A_i u^t u^{*\dagger} A_i^\dagger\right) v^*,$$

where $u^t = \widehat{u}^t/\|\widehat{u}^t\|_2$. Now, let $B = \sum_{i=1}^d A_i u^t (u^t)^\dagger A_i^\dagger$ and $C = \sum_{i=1}^d A_i u^t (u^*)^\dagger A_i^\dagger$. Then,

$$\begin{aligned} \|\widehat{u}^t\|_2 \widehat{v}^{t+1} &= \sigma^* B^{-1} C v^*, \\ &= \underbrace{\langle u^*, u^t\rangle \sigma^* v^*}_{\text{Power Method}} - \underbrace{B^{-1}\left(\langle u^*, u^t\rangle B - C\right)\sigma^* v^*}_{\text{Error Term}}. \end{aligned} \tag{4}$$

Note that the first term in the above expression is the power method iterate (i.e., $M^\dagger u^t$). The second term is an error term and the goal is to show that it becomes smaller as u^t gets closer to u^*. Note that when $u^t = u^*$, the error term is 0 *irrespective* of the measurement operator $\mathcal{A}$.

Below, we provide a precise bound on the error term:

Lemma 2.4.2. *Consider the error term defined in* (4) *and let* $\mathcal{A}$ *satisfy* 2*-RIP with constant* δ_2*. Then,*

$$\|B^{-1}\left(\langle u^*, u^t\rangle B - C\right) v^*\| \leq \frac{3\delta_2}{1 - 3\delta_2}\sqrt{1 - \langle u^t, u^*\rangle^2}$$

See Appendix A.2.1 for a detailed proof of the above lemma.

Using the above lemma, we now finish the proof of Theorem 2.4.1:

Proof of Rank-1 case of Theorem 2.4.1. Let $v^{t+1} = \widehat{v}^{t+1}/\|\widehat{v}^{t+1}\|_2$. Now, using (4) and Lemma 2.4.2:,

$$\begin{aligned}\langle v^{t+1}, v^* \rangle &= \frac{\langle \widehat{v}^{t+1}, v^* \rangle}{\|\widehat{v}^{t+1}\|} = \frac{\langle \widehat{v}^{t+1}/\sigma^*, v^* \rangle}{\|\widehat{v}^{t+1}/\sigma^*\|} \\ &\leq \frac{\langle u^*, u^t \rangle - \widehat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2}}{\sqrt{\left(\langle u^*, u^t \rangle - \widehat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2}\right)^2 + \widehat{\delta}_2^2 \left(1 - \langle u^*, u^t \rangle^2\right)}},\end{aligned}$$

where $\widehat{\delta}_2 = \frac{3\delta_2}{1-3\delta_2}$. That is,

$$\mathrm{dist}^2(v^{t+1}, v^*) \leq \frac{\widehat{\delta}_2^2 (1 - \langle u^*, u^t \rangle^2)}{(\langle u^*, u^t \rangle - \widehat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2})^2 + \widehat{\delta}_2^2 (1 - \langle u^*, u^t \rangle^2)},$$

Hence, *assuming* $\langle u^*, u^t \rangle \geq 5\widehat{\delta}_2$, $\mathrm{dist}(v^{t+1}, v^*) \leq \frac{1}{4}\mathrm{dist}(u^t, u^*)$. As $\mathrm{dist}(u^{t+1}, u^*)$ and $\mathrm{dist}(v^{t+1}, v^*)$ are decreasing with t (from the above bound), we only need to show that $\langle u^0, u^t \rangle \geq 5\widehat{\delta}_2$. Recall that $\widehat{u}^0$ is obtained by using one step of SVP algorithm [44]. Hence, using Lemma 2.1 of [44] (see Lemma A.1.1):

$$\|\sigma_1^*(I - u^0(u^0)^\dagger)u^*)\|_2^2 \leq \|M - \widehat{u}^0(\widehat{v}^0)^\dagger\|_F^2 \leq 2\delta_2 \|M\|_F^2.$$

Therefore, $\langle u^0, u^* \rangle \geq \sqrt{1 - 2\delta_2} \geq 5\widehat{\delta}_2$ assuming $\delta_2 \leq \frac{1}{100}$. □

2.4.2 Rank-k Case

In this section, we present the proof of Theorem 2.4.1 for arbitrary k, i.e., when M is a rank-k matrix (with SVD $U^*\Sigma^*(V^*)^\dagger$).

Similar to the analysis for the rank-1 case (Section 2.4.1), we show that even for arbitrary k, the updates of AltMinSense are essentially power-method type updates but with a bounded error term whose magnitude decreases with each iteration.

However, directly analyzing iterates of AltMinSense is a bit tedious due to non-orthonormality of intermediate iterates $\widehat{U}$. Instead, **for analysis only** we consider the iterates of a modified version of AltMinSense, where

we explicitly orthonormalize each iterate using the QR-decomposition[3]. In particular, suppose we replace steps 4 and 5 of AltMinSensewith the following

$$\begin{aligned}
\widehat{U}^t &= U^t R_U^t \quad \text{(QR decomposition)}, \\
\widehat{V}^{t+1} &\leftarrow \arg\min_V \ \|\mathcal{A}(U^t V^\dagger) - b\|_2^2, \\
\widehat{V}^{t+1} &= V^{t+1} R_V^{t+1} \quad \text{(QR decomposition)} \\
\widehat{U}^{t+1} &\leftarrow \arg\min_U \ \|\mathcal{A}(U(V^{t+1})^\dagger) - b\|_2^2
\end{aligned} \tag{5}$$

In our algorithm, in each iterate both $\widehat{U}^t, \widehat{V}^t$ remain full-rank because dist $(U^t, U^*) < 1$; with this, the following lemma implies that the spaces spanned by the iterates in our AltMinSense algorithm are *exactly the same* as the respective ones by the iterates of the above modified version (and hence the distances dist$(\widehat{U}^t, U^*)$ and dist$(\widehat{V}^t, V^*)$ are also the same for the two algorithms).

Lemma 2.4.3. *Let $\widehat{U}^t$ be the t^{th} iterate of our AltMinSense algorithm, and $\widetilde{U}^t$ of the modified version stated above. Suppose also that both $\widehat{U}^t, \widetilde{U}^t$ are full-rank, and span the same subspace. Then the same will be true for the subsequent iterates for the two algorithms, i.e. Span$(\widehat{V}^{t+1})$ = Span$(\widetilde{V}^{t+1})$, Span$(\widehat{U}^{t+1})$ = Span$(\widetilde{U}^{t+1})$, and all matrices at iterate $t+1$ will be full-rank.*

The proof of the above lemma can be found in Appendix A.2.2. In light of this, we will now prove Theorem 2.4.1 **with** the new QR-based iterates (5).

Lemma 2.4.4. *Let $\widehat{U}^t$ be the t-th step iterate of AltMinSense and let $U^t, \widehat{V}^{t+1}$ and V^{t+1} be obtained by Update* (5)*. Then,*

$$\widehat{V}^{t+1} = \underbrace{V^* \Sigma^* U^{*\dagger} U^t}_{\substack{\textit{Power-method} \\ \textit{Update}}} - \underbrace{F}_{\substack{\textit{Error} \\ \textit{Term}}}, \quad V^{t+1} = \widehat{V}^{t+1}(R^{(t+1)})^{-1}, \tag{6}$$

where F is an error matrix defined in (8) *and $R^{(t+1)}$ is a triangular matrix obtained using QR-decomposition of $\widehat{V}^{t+1}$.*

[3]The QR decomposition factorizes a matrix into an orthonormal matrix (a basis of its column space) and an upper triangular matrix; that is given $\widehat{S}$ it computes $\widehat{S} = SR$ where S has orthonormal columns and R is upper triangular. If $\widehat{S}$ is full-rank, so are S and R.

See Appendix A.2 for a detailed proof of the above lemma.

Before we give an expression for the error matrix F, we define the following notation. Let $v^* \in \mathbb{R}^{nk}$ be given by: $v^* = vec(V^*)$, i.e., $v^* = \left[v_1^{*\dagger} v_2^{*\dagger} \dots v_k^{*\dagger}\right]^\dagger$. Define B, C, D, S as follows:

$$B \stackrel{\text{def}}{=} \begin{bmatrix} B_{11} & \cdots & B_{1k} \\ \vdots & \ddots & \vdots \\ B_{k1} & \cdots & B_{kk} \end{bmatrix}, C \stackrel{\text{def}}{=} \begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \cdots & C_{kk} \end{bmatrix},$$
$$D \stackrel{\text{def}}{=} \begin{bmatrix} D_{11} & \cdots & D_{1k} \\ \vdots & \ddots & \vdots \\ D_{k1} & \cdots & D_{kk} \end{bmatrix}, S \stackrel{\text{def}}{=} \begin{bmatrix} \sigma_1^* I_n & \dots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \dots & \sigma_k^* I_n \end{bmatrix}. \tag{7}$$

where , for $1 \le p, q \le k$: $B_{pq} \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^t u_q^{t\dagger} A_i^\dagger$,
$C_{pq} \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^t u_q^{*\dagger} A_i^\dagger$, and, $D_{pq} \stackrel{\text{def}}{=} \langle u_p^t, u_q^* \rangle \mathbb{I}_{n\times n}$. Recall that, u_p^t is the p-th column of U^t and u_q^* is the q-th left singular vector of the underlying matrix $M = U^* \Sigma^* (V^*)^\dagger$. Finally F is obtained by "de-stacking" the vector $B^{-1}(BD - C) S v^*$ i.e., the i^{th} column of F is given by:

$$F_i \stackrel{\text{def}}{=} \begin{bmatrix} \left(B^{-1}(BD - C) S v^*\right)_{ni+1} \\ \left(B^{-1}(BD - C) S v^*\right)_{ni+2} \\ \vdots \\ \left(B^{-1}(BD - C) S v^*\right)_{ni+n} \end{bmatrix}, F \stackrel{\text{def}}{=} [F_1\ F_2\ \cdots\ F_k]. \tag{8}$$

Note that the notation above should have been B^t, C^t and so on. We suppress the dependence on t for notational simplicity. Now, from Update (6), we have

$$V^{t+1} = \widehat{V}^{t+1} R^{(t+1)^{-1}} = \left(V^* \Sigma^* U^{*\dagger} U^t - F\right) R^{(t+1)^{-1}}$$
$$\Rightarrow V_\perp^{*\dagger} V^{t+1} = -V_\perp^{*\dagger} F R^{(t+1)^{-1}}. \tag{9}$$

where $V_\perp^*$ is an orthonormal basis of Span $(v_1^*, v_2^*, \cdots, v_k^*)^\perp$. Therefore,

$$dist(V^*, V^{t+1}) = \|V_\perp^{*\dagger} V^{t+1}\|_2 = \|V_\perp^{*\dagger} F R^{(t+1)^{-1}}\|_2 \le \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* R^{(t+1)^{-1}}\|_2.$$

Now, we break down the proof of Theorem 2.4.1 into the following two steps:

- show that $\|F(\Sigma^*)^{-1}\|_2$ is small (Lemma 2.4.5) and

- show that $\|\Sigma^* R^{(t+1)^{-1}}\|_2$ is small(Lemma 2.4.6).

We will now state the two corresponding lemmas. Complete proofs can be found in Appendix A.2.2 The first lemma bounds the spectral norm of $F(\Sigma^*)^{-1}$.

Lemma 2.4.5. *Let linear measurement $\mathcal{A}$ satisfy RIP for all $2k$-rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m\times n}$ being a rank-k matrix. Then, spectral norm of error matrix $F(\Sigma^*)^{-1}$ (see Equation 6) after t-th iteration update satisfy:*

$$\left\|F(\Sigma^*)^{-1}\right\|_2 \leq \frac{\delta_{2k}k}{1-\delta_{2k}} dist(U^t, U^*). \tag{10}$$

The following lemma bounds the spectral norm of $\Sigma^* R^{(t+1)^{-1}}$.

Lemma 2.4.6. *Let linear measurement $\mathcal{A}$ satisfy RIP for all $2k$-rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m\times n}$ being a rank-k matrix. Then,*

$$\|\Sigma^*(R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*/\sigma_k^*}{\sqrt{1-\text{dist}^2\left(U^t, U^*\right)} - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k\text{dist}(U^t,U^*)}{1-\delta_{2k}}}. \tag{11}$$

With the above two lemmas, we now prove Theorem 2.4.1.

Proof Of Theorem 2.4.1. Using (9), (10) and (11), we obtain the following:

$$\begin{aligned}
\text{dist}\left(V^{t+1}, V^*\right) &= \left\|V_\perp^{*\dagger} V^{t+1}\right\|_2, \\
&\leq \left\|V_\perp^{*\dagger} F(\Sigma^*)^{-1}\Sigma^* R^{(t+1)^{-1}}\right\|_2, \\
&\leq \|V_\perp^*\|_2 \left\|F(\Sigma^*)^{-1}\right\|_2 \left\|\Sigma^* R^{(t+1)^{-1}}\right\|_2 \\
&\leq \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k \cdot \text{dist}\left(U^t, U^*\right)}{(1-\delta_{2k})L},
\end{aligned} \tag{12}$$

where $L = \sqrt{1-\text{dist}\left(U^t, U^*\right)^2} - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k\text{dist}\left(U^t,U^*\right)}{1-\delta_{2k}}$. Also, note that U^0 is

obtained using SVD of $\sum_i A_i b_i$. Hence, using Lemma A.1.1, we have:

$$\begin{aligned}
&\|\mathcal{A}(U^0\Sigma^0V^0 - U^*\Sigma^*(V^*)^\dagger\|_2^2 \le 4\delta_{2k}\|\mathcal{A}(U^*\Sigma^*(V^*)^\dagger)\|_2^2,\\
\Rightarrow&\|U^0\Sigma^0V^0 - U^*\Sigma^*(V^*)^\dagger\|_F^2 \le 4\delta_{2k}(1+3\delta_{2k})\|\Sigma^*\|_F^2,\\
\Rightarrow&\|U^0(U^0)^\dagger U^*\Sigma^*(V^*)^\dagger - U^*\Sigma^*(V^*)^\dagger\|_F^2 \le 6\delta_{2k}\|\Sigma^*\|_F^2,\\
\Rightarrow&(\sigma_k^*)^2\|(U^0(U^0)^\dagger - I)U^*\|_F^2 \le 6\delta_{2k}k(\sigma_1^*)^2,\\
\Rightarrow&\text{dist}(U^0,U^*) \le \sqrt{6\delta_{2k}k}\left(\frac{\sigma_1^*}{\sigma_k^*}\right) < \frac{1}{2}. \qquad (13)
\end{aligned}$$

Using (12) with $\text{dist}\,(U^0,U^*) < \frac{1}{2}$ and $\delta_{2k} < \frac{1}{24(\sigma_1^*/\sigma_k^*)^2k}$, we obtain: $\text{dist}\,(V^t,V^*) < \frac{1}{4}\text{dist}\,(U^t,U^*)$. Similarly we can show that $\text{dist}\,(U^{t+1},U^*) < \frac{1}{4}\text{dist}\,(V^t,V^*)$. □

2.5 Matrix Completion

In this section, we study the Matrix Completion problem (LRMC) and show that, assuming k and $\frac{\sigma_1^*}{\sigma_k^*}$ are constant, AltMinComplete (Algorithm 4) recovers the underlying matrix M using only $O(n\log n)$ measurements (i.e., we prove Theorem 2.3.3).

As mentioned, while observing elements in Ω constitutes a linear map, matrix completion is different from matrix sensing because the map does not satisfy RIP. The (now standard) approach is to assume incoherence of the true matrix M, as done in Definition 2.3.2. With this, and the random sampling of Ω, matrix completion exhibits similarities to matrix sensing. For our analysis, we can again use the fact that incoherence allows us to view alternating minimization as a perturbed power method, whose error we can control.

However, there are important differences between the two problems, which make the analysis of completion more complicated. Chief among them is the fact that we need to establish *the incoherence of each iterate*. For the first initialization $\widehat{U}^0$, this necessitates the "clipping" procedure (described in step 4 of the algorithm). For the subsequent steps, this requires the partitioning of the observed Ω into $2T+1$ sets (as described in step 2 of the algorithm).

As in the case of matrix sensing, we prove our main result for matrix completion (Theorem 2.3.3) by first establishing a geometric decay of the

distance between the subspaces spanned by $\widehat{U}^t, \widehat{V}^t$ and U^*, V^* respectively.

Theorem 2.5.1. *Under the assumptions of Theorem 2.3.3, the* $(t+1)^{th}$ *iterates* $\widehat{U}^{t+1}$ *and* $\widehat{V}^{t+1}$ *satisfy the following property w.h.p.:*

$$\text{dist}\left(\widehat{V}^{t+1}, V^*\right) \leq \frac{1}{4}\text{dist}\left(\widehat{U}^t, U^*\right) \text{ and}$$
$$\text{dist}\left(\widehat{U}^{t+1}, U^*\right) \leq \frac{1}{4}\text{dist}\left(\widehat{V}^{t+1}, V^*\right), \quad \forall\ 1 \leq t \leq T.$$

We use the above result along with incoherence of M to prove Theorem 2.3.3. See Appendix A.3 for a detailed proof.

Now, similar to the matrix sensing case, alternating minimization needs an initial iterate that is close enough to U^* and V^*, from where it will then converge. To this end, Steps $3-4$ of Algorithm 4 use SVD of $P_\Omega(M)$ followed by clipping to initialize $\widehat{U}^0$. While the SVD step guarantees that $\widehat{U}^0$ is close enough to U^*, it might not remain incoherent. To maintain incoherence, we introduce an extra clipping step which guarantees incoherence of $\widehat{U}^0$ while also ensuring that $\widehat{U}^0$ is close enough to U^* (see Lemma 2.5.2)

Lemma 2.5.2. *Let* M, Ω, p *be as defined in Theorem 2.3.3. Also, let* U^0 *be the initial iterate obtained by step* 4 *of Algorithm 4. Then, w.h.p. we have*

- $\text{dist}\left(U^0, U^*\right) \leq \frac{1}{2}$ *and*
- U^0 *is incoherent with parameter* $4\mu\sqrt{k}$.

The above lemma guarantees a "good" starting point for alternating minimization. Using this, we now present a proof of Theorem 2.5.1. Similar to the sensing section, we first explain key ideas of our proof using rank-1 example. Then in Section 2.5.2 we extend our proof to general rank-k matrices.

2.5.1 Rank-1 Case

Consider the rank-1 matrix completion problem where $M = \sigma^* u^* (v^*)^\dagger$. Now, the t-th step iterates $\widehat{v}^{t+1}$ of Algorithm 4 are given by:

$$\widehat{v}^{t+1} = \arg\min_{\widehat{v}} \sum_{(i,j)\in\Omega} (M_{ij} - \widehat{u}_i^t \widehat{v}_j)^2.$$

Let $u^t = \widehat{u}^t / \|\widehat{u}^t\|_2$. Then, $\forall j$:

$$\begin{aligned}
&\|\widehat{u}^t\|_2 \sum_{i:(i,j)\in\Omega} (u_i^t)^2 \widehat{v}_j^{t+1} = \sigma^* \sum_{i:(i,j)\in\Omega} u_i^t u_i^* v_j^* \\
&\Rightarrow \|\widehat{u}^t\|_2 \widehat{v}_j^{t+1} = \frac{\sigma^*}{\sum_{i:(i,j)\in\Omega}(u_i^t)^2} \sum_{i:(i,j)\in\Omega} u_i^t u_i^* v_j^* \\
&= \sigma^* \langle u^t, u^* \rangle v_j^* - \frac{\sigma^* (\langle u^t, u^* \rangle \sum_{i:(i,j)\in\Omega} (u_i^t)^2 v_j^* - \sum_{i:(i,j)\in\Omega} u_i^t u_i^* v_j^*)}{\sum_{i:(i,j)\in\Omega}(u_i^t)^2}. \qquad (14)
\end{aligned}$$

Hence,

$$\|\widehat{u}^t\|_2 \widehat{v}^{t+1} = \underbrace{\langle u^*, u^t \rangle \sigma^* v^*}_{\text{Power Method}} - \underbrace{\sigma^* B^{-1} \left(\langle u^t, u^* \rangle B - C \right) v^*}_{\text{Error Term}}, \qquad (15)$$

where $B, C \in \mathbb{R}^{n \times n}$ are diagonal matrices, such that,

$$B_{jj} = \frac{\sum_{i:(i,j)\in\Omega} (u_i^t)^2}{p}, \quad C_{jj} = \frac{\sum_{i:(i,j)\in\Omega} u_i^t u_i^*}{p}. \qquad (16)$$

Note the similarities between the update (15) and the rank-1 update (4) for the sensing case. Here again, it is essentially a power-method update (first term) along with a bounded error term (see Lemma 2.5.3). Using this insight, we now prove Theorem 2.5.1 for the special case of rank-1 matrices. Our proof can be divided in three major steps:

- *Base Case*: Show that $u^0 = \widehat{u}^0 / \|\widehat{u}^0\|_2$ is incoherent and have small distance to u^* (see Lemma 2.5.2).
- *Induction Step (distance)*: Assuming $u^t = \widehat{u}^t / \|\widehat{u}^t\|_2$ to be incoherent and that u^t has a small distance to u^*, v^{t+1} decreases distances to v^* by at least a constant factor.
- *Induction Step (incoherence)*: Show incoherence of v^{t+1}, while assuming incoherence of u^t (see Lemma 2.5.4)

We first prove the second step of our proof. To this end, we provide the following lemma that bounds the error term. See Appendix A.3.2 for a proof of the below given lemma.

Lemma 2.5.3. *Let M, p, Ω, u^t be as defined in Theorem 2.3.3. Also, let u^t be a unit vector with incoherence parameter $\mu_1 = \frac{6(1+\delta_2)\mu}{1-\delta_2}$. Then, w.p. at least $1 - \frac{1}{n^3}$:*

$$\|B^{-1}\left(\langle u^*, u^t\rangle B - C\right) v^*\|_2 \leq \frac{\delta_2}{1-\delta_2}\sqrt{1 - \langle u^t, u^*\rangle^2}.$$

Multiplying (15) with v^* and using Lemma 2.5.3, we get:

$$\|\widehat{u}^t\|_2 \langle \widehat{v}^{t+1}, v^*\rangle \geq \sigma^* \langle u^t, u^*\rangle - 2\sigma^*\delta_2\sqrt{1 - \langle u^t, u^*\rangle^2}, \tag{17}$$

where $\delta_2 < \frac{1}{12}$ is a constant defined in the Theorem statement and is similar to the RIP constant in Section 2.4.

Similarly, by multiplying (15) with $v_\perp$ (where $\langle v_\perp^*, v^*\rangle = 0$ and $\|v_\perp^*\|_2 = 1$) and using Lemma 2.5.3:

$$\|\widehat{u}^t\|_2 \langle \widehat{v}^{t+1}, v_\perp^*\rangle \leq 2\sigma^*\delta_2\sqrt{1 - \langle u^t, u^*\rangle^2}.$$

Using the above two equations:

$$1 - \langle v^{t+1}, v^*\rangle^2 \leq \frac{4\delta_2^2(1 - \langle u^t, u^*\rangle^2)}{(\langle u^t, u^*\rangle - 2\delta_2\sqrt{1 - \langle u^t, u^*\rangle^2})^2 + (2\delta_2\sqrt{1 - \langle u^t, u^*\rangle^2})^2}.$$

Assuming, $\langle v^{t+1}, v^*\rangle \geq 6\delta_2$,

$$\text{dist}(v^{t+1}, v^*) = \sqrt{1 - \langle v^{t+1}, v^*\rangle^2} \leq \frac{1}{4}\sqrt{1 - \langle u^t, u^*\rangle^2}.$$

Using same arguments, we can show that, $\text{dist}(u^{t+1}, u^*) \leq \text{dist}(v^{t+1}, v^*)/4$. Hence, after $O(\log(1/\epsilon))$ iterations, $\text{dist}(u^t, u^*) \leq \epsilon$ and $\text{dist}(v^{t+1}, v^*) \leq \epsilon$. This proves our second step.

We now provide the following lemma to prove the third step. We stress that v^{t+1} does not increase the incoherence parameter (μ_1) when compared to that of u^t.

Lemma 2.5.4. *Let M, p, Ω be as defined in Theorem 2.3.3. Also, let u^t be a unit vector with incoherence parameter $\mu_1 = \frac{6(1+\delta_2)\mu}{1-\delta_2}$. Then, w.p. at least $1 - \frac{1}{n^3}$, v^{t+1} is also μ_1 incoherent.*

See Appendix A.3.2 for a detailed proof of the lemma.

Finally, for the base case we need that u^0 is μ_1 incoherent and also $\langle u^0, u^* \rangle \geq 6\delta_2$. This follows directly by using Lemma 2.5.2 and the fact that $\delta_2 \leq 1/12$.

Note that, to obtain an error of ϵ, AltMinComplete needs to run for $O\left(\log \frac{\|M\|_F}{\epsilon}\right)$ iterations. Also, we need to sample a fresh Ω at each iteration of AltMinComplete. Hence, the total number of samples needed by AltMinComplete is $O\left(\log \frac{\|M\|_F}{\epsilon}\right)$ larger than the number of samples required per step.

2.5.2 Rank-k case

We now extend our proof of Theorem 2.5.1 to matrices with arbitrary rank. Here again, we show that the AltMinComplete algorithm reduces to power method with bounded perturbation at each step.

Similar to the matrix sensing case, we analyze the following QR decomposition based update instead of directly analyzing the updates of Algorithm 4:

$$\begin{aligned}
\widehat{U}^t &= U^t R_U^t \quad \text{(QR decomposition)}, \\
\widehat{V}^{t+1} &= \arg\min_{\widehat{V}} \ \|P_\Omega(U^t \widehat{V}^\dagger) - P_\Omega(M)\|_F^2, \\
\widehat{V}^{t+1} &= V^{t+1} R_V^{t+1}. \quad \text{(QR decomposition)} \ , \\
\widehat{U}^{t+1} &= \arg\min_{\widehat{U}} \ \|P_\Omega(\widehat{U}(V^{t+1})^\dagger) - P_\Omega(M)\|_F^2.
\end{aligned} \tag{18}$$

Here again, we would stress that the updates output exactly the same matrices at the end of each iteration and we prefer QR-based updates due to notational ease.

Now, as matrix completion is a special case of matrix sensing, Lemma 2.4.4 characterizes the updates of the AltMinComplete algorithm (see Algorithm 4). That is,

$$\begin{aligned}
\widehat{V}^{t+1} &= \underbrace{V^* \Sigma^* U^{*\dagger} U^t}_{\text{Power-method Update}} - \underbrace{F}_{\text{Error Term}} \ , \\
V^{t+1} &= \widehat{V}^{t+1} (R^{(t+1)})^{-1},
\end{aligned} \tag{19}$$

where F is the error matrix defined in (8) and $R^{(t+1)}$ is a upper-triangular matrix obtained using QR-decomposition of $\widehat{V}^{t+1}$. See (7) for the definition of B, C, D, and S.

Also, note that for the special case of matrix completion, $B_{pq}, C_{pq}, 1 \leq p, q \leq k$ are *diagonal matrices* with

$$(B_{pq})_{jj} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} U^t_{ip} U^t_{iq}, \quad (C_{pq})_{jj} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} U^t_{ip} U^*_{iq}.$$

We use this structure to further simplify the update equation. We first define matrices $B^j, C^j, D^j \in \mathbb{R}^{k\times k}, 1 \leq i \leq n$:

$$B^j = \frac{1}{p} \sum_{i:(i,j)\in\Omega} (U^t)^{(i)} (U^t)^{(i)\dagger}, \quad C^j = \frac{1}{p} \sum_{i:(i,j)\in\Omega} (U^t)^{(i)} (U^*)^{(i)\dagger},$$

and $D^j = (U^t)^\dagger U^*$. Using the above notation, (19) decouples into n equations of the form $(1 \leq j \leq n)$:

$$(V^{t+1})^{(j)} = (V^*)^{(j)} (D^j - (B^j)^{-1}(B^j D^j - C^j))(R^{(t+1)})^{-1}, \tag{20}$$

where $(V^{t+1})^{(j)}$ and $(V^*)^{(j)}$ denote the j^{th} rows of V^{t+1} and V^* respectively.

Using the above notation, we now provide a proof of Theorem 2.5.1 for the general rank-k case.

Proof of Theorem 2.5.1. Multiplying the update equation (19) on the left by $(V^*_\perp)^\dagger$, we get:
$(V^*_\perp)^\dagger \widehat{V}^{t+1} = -(V^*_\perp)^\dagger F (R^{(t+1)})^{-1}$. That is,

$$\begin{aligned} \text{dist}(V^*, V^{t+1}) &= \|V^{*\dagger}_\perp V^{(t+1)}\|_2 = \|V^{*\dagger}_\perp F {R^{(t+1)}}^{-1}\|_2 \\ &\leq \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* {R^{(t+1)}}^{-1}\|_2. \end{aligned}$$

Now, similar to the sensing case (see Section 2.4.2) we break down our proof into the following two steps:

- Bound $\|F(\Sigma^*)^{-1}\|_2$ (Lemma 2.5.6) and
- Bound $\|\Sigma^* {R^{(t+1)}}^{-1}\|_2$, i.e., the minimum singular value of $(\Sigma^*)^{-1} R^{(t+1)}$ (Lemma 2.5.7).

Using Lemma 2.5.6 and Lemma 2.5.7, w.p. at least $1 - 1/n^3$,

$$\begin{aligned}\operatorname{dist}(V^*, V^{t+1}) &\leq \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* {R^{(t+1)}}^{-1}\|_2 \\ &\leq \frac{(\sigma_1^*/\sigma_k^*)k(\delta_{2k}/(1-\delta_{2k})) \cdot \operatorname{dist}\left(U^{(t)}, U^*\right)}{\sqrt{1 - \operatorname{dist}\left(U^{(t)}, U^*\right)^2} - \frac{(\sigma_1^*/\sigma_k^*)k\delta_{2k}\operatorname{dist}\left(U^{(t)}, U^*\right)}{1-\delta_{2k}}}.\end{aligned}$$

Now, using Lemma 2.5.2 we get: $\operatorname{dist}(U^t, U^*) \leq \operatorname{dist}(U^0, U^*) \leq \frac{1}{2}$. By selecting $\delta_{2k} < \frac{\sigma_k^*}{12k\sigma_1^*}$, i.e., $p \geq \frac{C(\sigma_1^*)^2 k^4 \log n}{m(\sigma_k^*)^2}$ and using above two inequalities:

$$\operatorname{dist}(V^{t+1}, V^*) \leq \frac{1}{4}\operatorname{dist}(U^t, U^*).$$

Furthermore, using Lemma 2.5.5 we get that V^{t+1} is μ_1 incoherent. Hence, using similar arguments as above, we also get: $\operatorname{dist}(U^{t+1}, U^*) \leq \left(\frac{1}{4}\right)\operatorname{dist}(V^{t+1}, V^*)$. □

We now provide lemmas required by our above given proof. See Appendix A.3.3 for a detailed proof of each of the lemmas.

We first provide a lemma to bound incoherence of V^{t+1}, assuming incoherence of U^t.

Lemma 2.5.5. *Let M, Ω, p be as defined in Theorem 2.3.3. Also, let U^t be the t-th step iterate obtained by (18). Let U^t be $\mu_1 = \frac{16\sigma_1^*\mu\sqrt{k}}{\sigma_k^*}$ incoherent. Then, w.p. at least $1 - 1/n^3$, iterate $V^{(t+1)}$ is also μ_1 incoherent.*

We now bound the error term (F) in AltMin update (19).

Lemma 2.5.6. *Let F be the error matrix defined by (8) (also see (19)) and let U^t be a μ_1-incoherent orthonormal matrix obtained after $(t-1)^{th}$ update. Also, let M, Ω, and p satisfy assumptions of Theorem 2.3.3. Then, w.p. at least $1 - 1/n^3$:*

$$\left\|F(\Sigma^*)^{-1}\right\|_2 \leq \frac{\delta_{2k}k}{1-\delta_{2k}} dist(U^t, U^*).$$

Next, we present a lemma to bound $\|(R^{(t+1)})^{-1}\|_2$.

Lemma 2.5.7. *Let $R^{(t+1)}$ be the lower-triangular matrix obtained by QR decomposition of $\widehat{V}^{t+1}$ (see (19)) and let U^t be a μ_1-incoherent orthonormal matrix obtained after $(t-1)^{th}$ update. Also, let M and Ω satisfy assumptions of Theorem 2.3.3. Then,*

$$\|\Sigma^*(R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*/\sigma_k^*}{\sqrt{1-\text{dist}^2\left(U^{(t)},U^*\right)} - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k\text{dist}(U^{(t)},U^*)}{1-\delta_{2k}}} \tag{21}$$

Proof. Lemma follows by exactly the same proof as that of Lemma 2.4.6 for the matrix sensing case. □

2.6 Stagewise AltMin Algorithm

In Section 2.4, we showed that if $\delta_{2k} \leq \frac{(\sigma_k^*)^2}{(\sigma_1^*)^2 k}$ then AltMinSense (Algorithm 2) recovers the underlying matrix. This means that, $d = \frac{(\sigma_1^*)^4}{(\sigma_k^*)^4}k^2 n \log n$ random Gaussian measurements (assume $m \leq n$) are required to recover M. For matrices with large condition number (σ_1^*/σ_k^*), this would be significantly larger than the information theoretic bound of $O(kn\log n/k)$ measurements.

To alleviate this problem, we present a modified version of AltMinSense called Stage-AltMin. Stage-AltMin proceeds in k stages where in the i-th stage, a rank-i problem is solved. The goal of the i-th stage is to recover top i-singular vectors of M, up to $O(\sigma_{i+1}^*)$ error.

Specifically, we initialize the i-th stage of our algorithm using one step of the SVP algorithm [44] (see Step 3 of Algorithm 3). We then show that, if $\delta_{2k} \leq \frac{1}{10k}$, then Stage-AltMin (Steps 3, 3 of Algorithm 3) decreases the error $\|M - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\|_F$ to $O(\sigma_{i+1}^*)$. Hence, after k steps, the error decreases to $O(\sigma_{k+1}^*) = 0$. Note that, $\widehat{U}_{1:i}^t \in \mathbb{R}^{m\times i}$ represents the t-th step iterate (U) in the i-th stage; $\widehat{V}_{1:i}^t \in \mathbb{R}^{n\times i}$ is also defined similarly.

Recall that, the main problem with our analysis of AltMinSense is that if $\sigma_i \gg \sigma_{i+1}$ (for some i) then $\delta_{2k} \leq \frac{(\sigma_{i+1}^*)^2}{(\sigma_i^*)^2 k}$ would need to be small. However, in such a scenario, the i-th stage of Algorithm 3 can be thought of as solving a noisy sensing problem where the goal is to recover $M_i \stackrel{\text{def}}{=} U_{1:i}^*\Sigma_{1:i}^*(V_{1:i}^*)^\dagger$ using noisy measurements $b = \mathcal{A}(U_{1:i}^*\Sigma_{1:i}^*(V_{1:i}^*)^\dagger + N)$ where noise matrix $N \stackrel{\text{def}}{=}$

$U^*_{i+1:k}\Sigma^*_{i+1:k}(V^*_{i+1:k})^\dagger$. Here M_i and N represent the top i singular components and last $k-i$ singular components of M respectively. Hence, using noisy-case type analysis (see Section A.2.3) we show that the error $\|M - \widehat{U}^t(\widehat{V}^t)^\dagger\|_F$ decreases to $O(\sigma^*_{i+1})$.

We now formally present the proof of our main result (see Theorem 2.3.2).

Proof Of Theorem 2.3.2. We prove the theorem using mathematical induction.

Base Case: After the 0-th step, error is: $\|M\|_F^2 \leq \sum_{j=1}^k \sigma_j^2 \leq k\sigma_1^2$. Hence, base case holds.

Induction Step: Here, assuming that the error bound holds for $(i-1)$-th stage, we prove the error bound for the i-th stage.

Our proof proceeds in two steps. First, we show that the initial point $\widehat{U}^0_{1:i}$, $\widehat{V}^0_{1:i}$ of the i-th stage, obtained using Step 3, has $c(\sigma^*_i)^2 + O\left(k(\sigma^*_{i+1})^2\right)$ error, with $c < 1$. In the second step, we show that using the initial points $\widehat{U}^0_{1:i}$, $\widehat{V}^0_{1:i}$, the AltMin algorithm iterations in the i-th stage (Steps 3, 3) reduces the error to $\max(\epsilon, 16k\sigma^2_{i+1})$.

We formalize the above mentioned first step in Lemma 2.6.1 and then prove the second step in Lemma 2.6.2. □

We now present two lemmas used by the above given proof. See Appendix A.2.4 for a proof of each of the lemmas.

Lemma 2.6.1. *Let assumptions of Theorem 2.3.2 be satisfied. Also, let $\widehat{U}^0_{1:i}$, $\widehat{V}^0_{1:i}$ be the output of Step 3 of Algorithm 3. Then, assuming that $\|M - \widehat{U}^T_{1:i-1}\widehat{V}^T_{1:i-1}\|_F^2 \leq 16k(\sigma^*_i)^2$, we obtain:*

$$\left\|M - \widehat{U}^0_{1:i}(\widehat{V}^0_{1:i})^\dagger\right\|_F^2 \leq \sum_{j=i+1}^{k}(\sigma^*_j)^2 + \frac{1}{100}(\sigma^*_i)^2.$$

Lemma 2.6.2. *Let assumptions of Theorem 2.3.2 be satisfied. Also, let $\widehat{U}^T_{1:i}$, $\widehat{V}^T_{1:i}$ be the T-th step iterates of the i-th stage of Algorithm 3. Then, assuming that $\|M - \widehat{U}^0_{1:i}V^0_{1:i}\|_F^2 \leq \sum_{j=i+1}^k(\sigma^*_j)^2 + \frac{1}{100}(\sigma^*_i)^2$, we obtain:*

$$\left\|M - \widehat{U}^T_{1:i}(\widehat{V}^T_{1:i})^\dagger\right\|_F^2 \leq \max(\epsilon, 16k(\sigma^*_{i+1})^2),$$

where $T = \Omega(\log(\|M\|_F/\epsilon))$.

2.7 Numerical Experiments

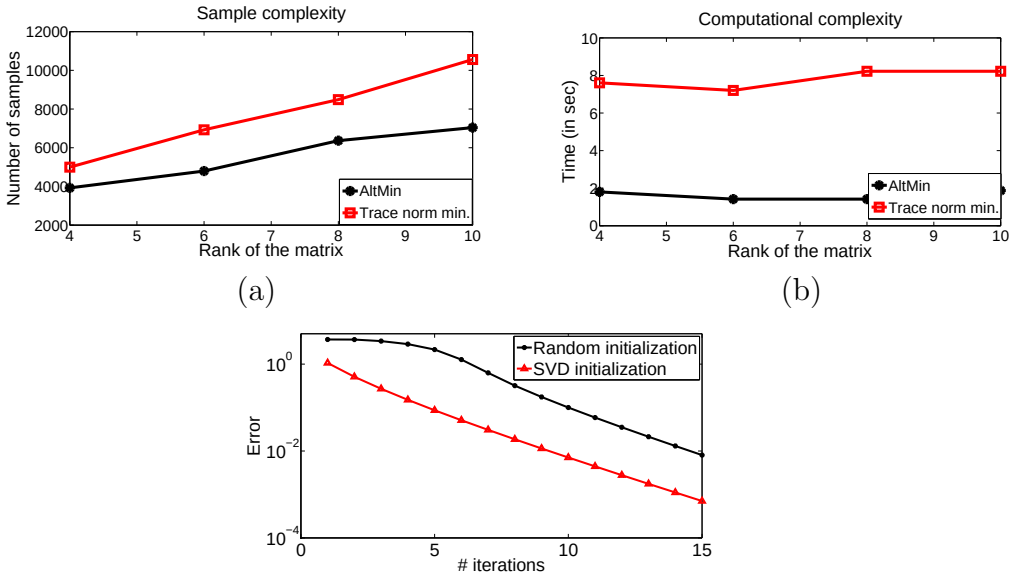


Figure 2.1: (a) Sample complexity and (b) computational complexity of AltMin as compared to trace norm minimization [15] on random low rank 225×225 matrices. Sample complexity denotes the number of observations needed for exact recovery of the matrix and computational complexity denotes the time taken by the algorithm. The plots were obtained after averaging over 10 trials. Clearly, AltMin has lower sample and computational complexity as compared to trace norm minimization. (c) denotes the error (on the observations) after each iteration of AltMin with random initialization and with our SVD based initialization. We see that with random initialization, the error decays very slowly initially but later on decays at a good rate, where as with SVD initialization, the error decays at a good rate from the beginning.

We now present some numerical experiments to verify if our theoretical results are tight. In particular, our results are weaker than those for convex optimization and one concrete question is if alternating minimization does

perform poorly in practice as compared to trace norm minimization algorithm. Fig. 2.1(a) suggests that this is not the case and we believe further tightening our bounds is possible.

We finally note that steps $3-4$ of AltMinComplete initialize the AltMin procedure in a principled manner. In contrast, empirically random initialization is quite popular. Even though random initialization works well in practice, proving rigorous guarantees is hard since the initial decay in error does not seem to have a good rate (see Fig. 2.1(c)).

2.8 Summary and Discussion

Alternating minimization provides an empirically appealing and popular approach to solving several different low-rank matrix recovery problems. The main motivation, and result, of this work was to provide the first theoretical guarantees on the global optimality of alternating minimization, for matrix completion and the related problem of matrix sensing. We would like to note the following aspects of our results and proofs:

- For both the problems, we show that alternating minimization recovers the true matrix under *similar problem conditions* (RIP, incoherence) to those used by existing algorithms (based on convex optimization or iterated SVDs); computationally, our results show faster convergence to the global optima, but with possibly higher statistical (i.e. sample) complexity.
- We develop a new framework for analyzing alternating minimization for low-rank problems. Key observation of our framework is that for some problems (under standard problem conditions) alternating minimization can be viewed as a perturbed version of the power method. In our case, we can control the perturbation error based on the extent of RIP / incoherence demonstrated by the problem. This idea is likely to have applications to other similar problems where trace-norm based convex relaxation techniques have rigorous theoretical results but alternating minimization has enjoyed more empirical success. For example, robust PCA [22, 14], spectral clustering [46] etc.

- Our analysis also sheds light on two key aspects of the alternating minimization approach:
  Initialization: Due to its connection to power method, it is now easy to see that for alternating minimization to succeed, the initial iterate should not be orthogonal to the target vector. Our results indeed show that alternating minimization succeeds if the initial iterate is not "almost orthogonal" to the target subspace. This suggests that, selecting initial iterate smartly is preferable to random initialization.
  Dependence on the condition number: Our results for the alternating minimization algorithm depend on the condition number. However, using a stagewise adaptation of alternating minimization, we can remove this dependence for the matrix sensing problem. This suggests that modifications of the basic alternating minimization algorithm may in fact perform better than the original one, while retaining the computational / implementational simplicity of the underlying method.

Chapter 3

Phase Retrieval using Alternating Minimization

3.1 Introduction

In this chapter [1], we are interested in recovering a complex[2] vector $x^* \in \mathbb{C}^n$ from *magnitudes of* its linear measurements. That is, for $a_i \in \mathbb{C}^n$, if

$$y_i \;=\; |\langle a_i, x^* \rangle|, \quad \text{for } i = 1, \ldots, m \tag{1}$$

then the task is to recover x^* using y and the measurement matrix $A = [a_1 \; a_2 \; \ldots \; a_m]$.

The above problem arises in many settings where it is harder / infeasible to record the phase of measurements, while recording the magnitudes is significantly easier. This problem, known as *phase retrieval*, is encountered in several applications in crystallography, optics, spectroscopy and tomography [69, 42]. Moreover, the problem is broadly studied in the following two settings:

(i) The measurements in (1) correspond to the Fourier transform (the number of measurements here is equal to n) and there is some apriori information about the signal.

(ii) The set of measurements y are overcomplete (i.e., $m > n$), while some apriori information about the signal may or may not be available.

[1] An extended abstract of the results in this chapter appeared as [71]. The coauthors on the paper had equal contributions in obtaining these results.

[2] Our results also cover the real case, i.e. where all quantities are real.

In the first case, various types of apriori information about the underlying signal such as positivity, magnitude information on the signal [31], sparsity [80] and so on have been studied. In the second case, algorithms for various measurement schemes such as Fourier oversampling [70], multiple random illuminations [12, 90] and wavelet transform [24] have been suggested.

By and large, the most well known methods for solving this problem are the error reduction algorithms due to Gerchberg and Saxton [37] and Fienup [31], and variants thereof. These algorithms are alternating projection algorithms that iterate between the unknown phases of the measurements and the unknown underlying vector. Though the empirical performance of these algorithms has been well studied [31, 61, 62]. and they are used in many applications [67, 68], there are not many theoretical guarantees regarding their performance.

More recently, a line of work [21, 17, 90] has approached this problem from a different angle, based on the realization that recovering x^* is equivalent to recovering the rank-one matrix x^*x^{*T}, i.e., its outer product. Inspired by the recent literature on trace norm relaxation of the rank constraint, they design SDPs to solve this problem. Refer Section 3.2 for more details.

In this work, we go back to the empirically more popular ideology of alternating minimization; we develop a new alternating minimization algorithm, for which we show that *(a)* empirically, it noticeably outperforms convex methods, and *(b)* analytically, a natural resampled version of this algorithm requires $O(n \log^3 n)$ i.i.d. random Gaussian measurements to geometrically converge to the true vector.

Our contribution:

- The iterative part of our algorithm is implicit in previous work [37, 31, 90, 12]; the novelty in our *algorithmic contribution* is the initialization step which makes it more likely for the iterative procedure to succeed.
- Our *analytical contribution* is the first theoretical guarantee regarding the global convergence, and subsequent exact recovery of the signal, via alternating minimization for phase retrieval.

Besides being an empirically better algorithm for this problem, our work is also interesting in a broader sense: there are many problems in machine learning, signal procesing and numerical linear algebra, where the natural formulation of a problem is non-convex; examples include rank constrained problems, applications of EM algorithms etc., and alternating minimization has good empirical performance. However, the methods with the best (or only) analytical guarantees involve convex relaxations (e.g., by relaxing the rank constraint and penalizing the trace norm). In most of these settings, correctness of alternating minimization is an open question. We believe that our results in this chapter are of interest, and may have implications, in this larger context.

Due to lack of space, we only present the algorithm and main results in this chapter. Refer [71] for complete proofs of all the results in this chapter.

3.2 Related Work

Phase Retrieval via Non-Convex Procedures: Inspite of the huge amount of work it has attracted, phase retrieval has been a long standing open problem. Early work in this area focused on using holography to capture the phase information along with magnitude measurements [33, 57]. However, computational methods for reconstruction of the signal using only magnitude measurements received a lot of attention due to their applicability in resolving spurious noise, fringes, optical system aberrations and so on and difficulties in the implementation of interferometer setups [26]. Though such methods have been developed to solve this problem in various practical settings [25, 32, 67, 68], our theoretical understanding of this problem is still far from complete. Many papers [9, 39, 78] have focused on determining conditions under which (1) has a unique solution. However, the uniqueness results of these papers do not resolve the algorithmic question of how to find the solution to (1).

Since the seminal work of Gerchberg and Saxton [37] and Fienup [31], many iterated projection algorithms have been developed targeted towards various applications [1, 29, 6]. [70] first suggested the use of multiple magnitude measurements to resolve the phase problem. This approach has been successfully used in many practical applications - see [26] and references there

in. Following the empirical success of these algorithms, researchers were able to explain its success in some of the instances [92, 86] using Bregman's theory of iterated projections onto convex sets [8]. However, many instances, such as the one we consider in this chapter, are out of reach of this theory since they involve magnitude constraints which are non-convex. To the best of our knowledge, there are no theoretical results on the convergence of these approaches in a non-convex setting.

Phase Retrieval via Convex Relaxation: An interesting recent approach for solving this problem formulates it as one of finding the rank-one solution to a system of linear matrix equations. The papers [21, 17] then take the approach of relaxing the rank constraint by a trace norm penalty, making the overall algorithm a convex program (called *PhaseLift*) over $n \times n$ matrices. Another recent line of work [90] takes a similar but different approach : it uses an SDP relaxation (called *PhaseCut*) that is inspired by the classical SDP relaxation for the max-cut problem. To date, these convex methods are the only ones with analytical guarantees on statistical performance [13, 90] (i.e. the number m of measurements required to recover x^*) – under an i.i.d. random Gaussian model on the measurement vectors a_i. However, by "lifting" a vector problem to a matrix one, these methods lead to a much larger representation of the state space, and higher computational cost as a result.

Sparse Phase Retrieval: A special case of the phase retrieval problem which has received a lot of attention recently is when the underlying signal x^* is known to be sparse. Though this problem is closely related to the compressed sensing problem, lack of phase information makes this harder. However, the ℓ_1 regularization approach of compressed sensing has been successfully used in this setting as well. In particular, if x^* is sparse, then the corresponding lifted matrix x^*x^{*T} is also sparse. [80, 72, 60] use this observation to design ℓ_1 regularized SDP algorithms for phase retrieval of sparse vectors. For random Gaussian measurements, [60] shows that ℓ_1 regularized PhaseLift recovers x^* correctly if the number of measurements is $\Omega(k^2 \log n)$. By the results of [74], this result is tight up to logarithmic factors for ℓ_1 and trace norm regularized SDP relaxations. [43, 79] develop algorithms for phase retrieval from Fourier magnitude measurements. However, achieving the optimal sample complexity of $O\left(k \log \frac{n}{k}\right)$ is still open [28].

Alternating Minimization (a.k.a. **ALS**): Alternating minimization has been successfully applied to many applications in the low-rank matrix setting. For example, clustering [52], sparse PCA [96], non-negative matrix factorization [51], signed network prediction [41] etc. However, despite empirical success, for most of the problems, there are no theoretical guarantees regarding its convergence except to a local minimum. The only exceptions are the results in [49, 45] which give provable guarantees for alternating minimization for the problems of matrix sensing and matrix completion.

3.3 Notation

For every complex vector $w \in \mathbb{C}^n$, $|w| \in \mathbb{R}^n$ denotes its element-wise magnitude vector. w^T and A^T denote the Hermitian transpose of the vector w and the matrix A respectively. e_1, e_2, etc. denote the canonical basis vectors in $\mathbb{C}^n$. $\overline{z}$ denotes the complex conjugate of the complex number z. In this chapter, we use the standard Gaussian (or normal) distribution over $\mathbb{C}^n$. a is said to be distributed according to this distribution if $a = a_1 + ia_2$, where a_1 and a_2 are independent and are distributed according to $\mathcal{N}(0, I)$. We also define $\mathrm{Ph}(z) \stackrel{\text{def}}{=} \frac{z}{|z|}$ for every $z \in \mathfrak{C}$, and $\mathrm{dist}(w_1, w_2) \stackrel{\text{def}}{=} \sqrt{1 - \left|\frac{\langle w_1, w_2 \rangle}{\|w_1\|_2 \|w_2\|_2}\right|^2}$ for every $w_1, w_2 \in \mathbb{C}^n$. inally, we use the shorthand wlog for without loss of generality and whp for with high probability.

3.4 Algorithm

In this section, we present our alternating minimization based algorithm for solving the phase retrieval problem. Let $A \in \mathbb{C}^{n \times m}$ be the measurement matrix, with a_i as its i^{th} column; similarly let y be the vector of recorded magnitudes. Then,

$$y \;=\; |\, A^T x^* \,|.$$

Recall that, given y and A, the goal is to recover x^*. If we had access to the true phase c^* of $A^T x^*$ (i.e., $c_i^* = \mathrm{Ph}(\langle a_i, x^* \rangle)$) and $m \geq n$, then our problem reduces to one of solving a system of linear equations:

$$C^* y \;=\; A^T x^*,$$

Algorithm 5 AltMinPhase

input A, y, t_0

1: Initialize $x^0 \leftarrow$ top singular vector of $\sum_i y_i^2 a_i {a_i}^T$
2: **for** $t = 0, \cdots, t_0 - 1$ **do**
3: $\quad C^{t+1} \leftarrow \mathrm{Diag}\left(\mathrm{Ph}\left(A^T x^t\right)\right)$
4: $\quad x^{t+1} \leftarrow \arg\min_{x \in \mathbb{R}^n} \left\| A^T x - C^{t+1} y \right\|_2$
5: **end for**

output x^{t_0}

where $C^* \stackrel{\text{def}}{=} \mathrm{Diag}(c^*)$ is the diagonal matrix of phases. Of course we do not know C^*, hence one approach to recovering x^* is to solve:

$$\arg\min_{C,x} \ \|A^T x - Cy\|_2, \tag{2}$$

where $x \in \mathbb{C}^n$ and $C \in \mathfrak{C}^{m \times m}$ is a diagonal matrix with each diagonal entry of magnitude 1. Note that the above problem is *not convex* since C is restricted to be a diagonal phase matrix and hence, one cannot use standard convex optimization methods to solve it.

Instead, our algorithm uses the well-known alternating minimization: alternatingly update x and C so as to minimize (2). Note that given C, the vector x can be obtained by solving the following least squares problem: $\min_x \|A^T x - Cy\|_2$. Since the number of measurements m is larger than the dimensionality n and since each entry of A is sampled from independent Gaussians, A is invertible with probability 1. Hence, the above least squares problem has a unique solution. On the other hand, given x, the optimal C is given by $C = \mathrm{Diag}(A^T x)$.

While the above algorithm is simple and intuitive, it is known that with bad initial points, the solution might not converge to x^*. In fact, this algorithm with a uniformly random initial point has been empirically evaluated for example in [90], where it performs worse than SDP based methods. Moreover, since the underlying problem is non-convex, standard analysis techniques fail to guarantee convergence to the global optimum, x^*. Hence, the key challenges here are: a) a good initialization step for this method, b) establishing this method's convergence to x^*.

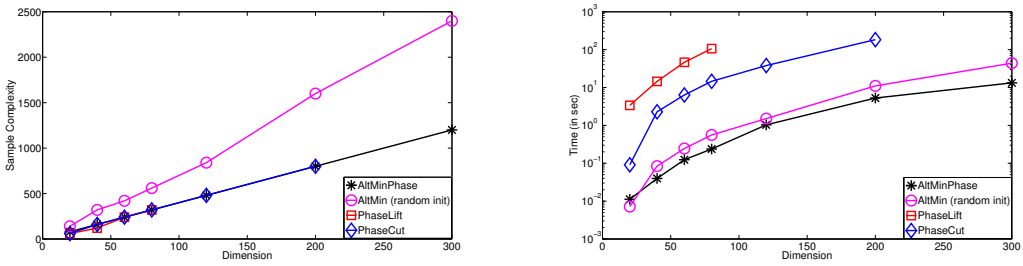
We address the first key challenge in our AltMinPhase algorithm (Algorithm 5) by initializing x as the largest singular vector of the matrix $S = \frac{1}{m}\sum_i y_i a_i a_i^T$. Theorem 3.5.1 shows that when A is sampled from standard complex normal distribution, this initialization is accurate. In particular, if $m \geq C_1 n \log^3 n$ for large enough $C_1 > 0$, then whp we have $\|x^0 - x^*\|_2 \leq 1/100$ (or any other constant).

Theorem 3.5.2 addresses the second key challenge and shows that a variant of AltMinPhase (see Algorithm 6) actually converges to the global optimum x^* at linear rate. See section 3.5 for a detailed analysis of our algorithm.

We would like to stress that not only does a natural variant of our proposed algorithm have rigorous theoretical guarantees, it also is effective practically as each of its iterations is fast, has a closed form solution and does not require SVD computation. AltMinPhase has similar statistical complexity to that of PhaseLift and PhaseCut while being much more efficient computationally. In particular, for accuracy ϵ, we only need to solve each least squares problem only up to accuracy $O(\epsilon)$. Now, since the measurement matrix A is sampled from Gaussian with $m > Cn$, it is well conditioned. Hence, using conjugate gradient method, each such step takes $O\left(mn \log \frac{1}{\epsilon}\right)$ time. When $m = O(n)$ and we have geometric convergence, the total time taken by the algorithm is $O\left(n^2 \log^2 \frac{1}{\epsilon}\right)$. SDP based methods on the other hand require $\Omega(n^3/\sqrt{\epsilon})$ time. Moreover, our initialization step increases the likelihood of successful recovery as opposed to a random initialization (which has been considered so far in prior work). Refer Figure 3.1 for an empirical validation of these claims.

3.5 Our Results

In this section we describe the main contribution of this work: provable statistical guarantees for the success of alternating minimization in solving the phase recovery problem. To this end, we consider the setting where each measurement vector a_i is iid and is sampled from the standard complex normal distribution. We would like to stress that all the existing guarantees for phase recovery also use exactly the same setting [17, 13, 90]. Table 3.1 presents



(a) **(b)**

Figure 3.1: Sample and Time complexity of various methods for Gaussian measurement matrices A. Figure 3.1(a) compares the number of measurements required for successful recovery by various methods. We note that our initialization improves sample complexity over that of random initialization (AltMin (random init)) by a factor of 2. AltMinPhase requires similar number of measurements as PhaseLift and PhaseCut. Figure 3.1(b) compares the running time of various algorithms on log-scale. Note that AltMinPhase is almost two orders of magnitude faster than PhaseLift and PhaseCut.

a comparison of the theoretical guarantees of Algorithm 6 as compared to PhaseLift and PhaseCut.

	Sample complexity	Comp. complexity
PhaseLift [13]	$O(n)$	$O(n^3/\epsilon^2)$
PhaseCut [90]	$O(n)$	$O(n^3/\sqrt{\epsilon})$

Table 3.1: Comparison of Algorithm 6 with PhaseLift and PhaseCut: Though the theoretical sample complexity of Algorithm 6 is off by log factors from that of PhaseLift and PhaseCut, it is $O(n)$ better than them in computational complexity. Note that, we can solve the least squares problem in each iteration approximately by using conjugate gradient method which requires only $O(mn)$ time.

Our proof for convergence of alternating minimization can be broken into two key results. We first show that if $m \geq Cn\log^3 n$, then whp the initialization step used by AltMinPhase returns x^0 which is at most a constant distance away from x^*. Furthermore, that constant can be controlled by using more samples (see Theorem 3.5.1).

We then show that if x^t is a *fixed* vector such that $\text{dist}(x^t, x^*) < c$ (small enough) and A is sampled independently of x^t with $m > Cn$ (C large enough) then whp x^{t+1} satisfies: $\text{dist}(x^{t+1}, x^*) < \frac{3}{4}\text{dist}(x^t, x^*)$ (see Theorem 3.5.2). Note that our analysis critically requires x^t to be "fixed" and be independent of the sample matrix A. Hence, we cannot re-use the same A in each iteration; instead, we need to resample A in every iteration. Using these results, we prove the correctness of Algorithm 6, which is a natural resampled version of AltMinPhase.

Algorithm 6 AltMinPhase with Resampling

input A, y, ϵ

1: $t_0 \leftarrow c \log \frac{1}{\epsilon}$
2: Partition y and (the corresponding columns of) A into $t_0 + 1$ equal disjoint sets: $(y^0, A^0), (y^1, A^1), \cdots, (y^{t_0}, A^{t_0})$
3: $x^0 \leftarrow$ top singular vector of $\sum_l (y_l^0)^2 a_\ell^0 (a_\ell^0)^T$
4: **for** $t = 0, \cdots, t_0 - 1$ **do**
5: $\quad C^{t+1} \leftarrow \text{Diag}\left(\text{Ph}\left((A^{t+1})^T x^t\right)\right)$
6: $\quad x^{t+1} \leftarrow \arg\min_{x \in \mathbb{R}^n} \left\|(A^{t+1})^T x - C^{t+1} y^{t+1}\right\|_2$
7: **end for**

output x^{t_0}

We now present the two results mentioned above. In the following theorems, wlog, we assume that $\|x^*\|_2 = 1$. Our first result guarantees a good initial vector.

Theorem 3.5.1. *There exists a constant C_1 such that if $m > \frac{C_1}{c^2} n \log^3 n$, then in Algorithm 6, with probability greater than $1 - 4/n^2$ we have:*

$$\left\|x^0 - x^*\right\|_2 < c.$$

The second result proves geometric decay of error assuming a good initialization.

Theorem 3.5.2. *There exist constants c, $\widehat{c}$ and $\widetilde{c}$ such that in iteration t of Algorithm 6, if $\text{dist}(x^t, x^*) < c$ and the number of columns of A^t is greater*

than $\widehat{c}\left(\log\frac{1}{\eta}\right)n$ *then, with probability more than* $1-\eta$*, we have:*

$$\operatorname{dist}\left(x^{t+1},x^*\right)<\frac{3}{4}\operatorname{dist}\left(x^t,x^*\right),\ \textit{and}\ \left\|x^{t+1}-x^*\right\|_2<\widetilde{c}\operatorname{dist}\left(x^t,x^*\right).$$

Proof. For simplicity of notation in the proof of the theorem, we will use A for A^{t+1}, C for C^{t+1}, x for x^t, x^+ for x^{t+1}, and y for y^{t+1}. Now consider the update in the $(t+1)^{\text{th}}$ iteration:

$$x^+=\arg\min_{\widetilde{x}\in\mathbb{R}^n}\left\|A^T\widetilde{x}-Cy\right\|_2=\left(AA^T\right)^{-1}ACy=\left(AA^T\right)^{-1}ADA^Tx^*, \tag{3}$$

where D is a diagonal matrix with $D_{ll}\overset{\text{def}}{=}\operatorname{Ph}\left(a_\ell{}^Tx\cdot\overline{a_\ell{}^Tx^*}\right)$. Now (3) can be rewritten as:

$$x^+=\left(AA^T\right)^{-1}ADA^Tx^*=x^*+\left(AA^T\right)^{-1}A\left(D-I\right)A^Tx^*, \tag{4}$$

that is, x^+ can be viewed as a perturbation of x^* and the goal is to bound the error term (the second term above). We break the proof into two main steps:

1. $\exists$ a constant c_1 such that $|\langle x^*,x^+\rangle|\geq 1-c_1\operatorname{dist}\left(x,x^*\right)$, and
2. $|\langle z,x^+\rangle|\leq\frac{5}{9}\operatorname{dist}\left(x,x^*\right)$, for all z s.t. $z^Tx^*=0$.

Assuming the above two bounds and choosing $c<\frac{1}{100c_1}$, we can prove the theorem:

$$\operatorname{dist}\left(x^+,x^*\right)^2<\frac{(25/81)\cdot\operatorname{dist}\left(x,x^*\right)}{(1-c_1\operatorname{dist}\left(x,x^*\right))^2}\leq\frac{9}{16}\operatorname{dist}\left(x,x^*\right)^2,$$

proving the first part of the theorem. The second part follows from (4) and by controlling $\left\|\left(AA^T\right)^{-1}A\left(D-I\right)A^Tx^*\right\|_2$. □

Combining Theorems 3.5.1 and 3.5.2, we have the following theorem establishing the correctness of Algorithm 6.

Theorem 3.5.3. *Suppose the measurement vectors in* (1) *are independent standard complex normal vectors. For every* $\eta>0$*, there exists a constant* c *such that if* $m>cn\left(\log^3 n+\log\frac{1}{\epsilon}\log\log\frac{1}{\epsilon}\right)$ *then, with probability greater than* $1-\eta$*, Algorithm 6 outputs* x^{t_0} *such that* $\left\|x^{t_0}-x^*\right\|_2<\epsilon$*.*

Algorithm 7 SparseAltMinPhase

input A, y, k

1: $S \leftarrow$ top-k $\arg\max_{j\in[n]} \sum_{i=1}^{m} |a_{ij} y_i|$ {Pick indices of k largest absolute value inner product}

2: Apply Algorithm 6 on A_S, y_S and output the resulting vector with elements in S^c set to zero.

	Sample complexity	Comp. complexity
Algorithm 7	$\widetilde{O}\left(k\left(k\log n + \log\frac{1}{\epsilon}\right)\right)$	$\widetilde{O}\left(k^2\left(kn\log n + \log^2\frac{1}{\epsilon}\right)\right)$
ℓ_1-PhaseLift [60]	$O\left(k^2\log n\right)$	$O\left(n^3/\epsilon^2\right)$

Table 3.2: Comparison of Algorithm 7 with ℓ_1-PhaseLift when $x^*_{\min} = \Omega\left(1/\sqrt{k}\right)$. Note that the complexity of Algorithm 7 is dominated by the support finding step. If $k = O(1)$, Algorithm 7 runs in quasi-linear time.

3.6 Sparse Phase Retrieval

In this section, we consider the case where x^* is known to be sparse, with sparsity k. A natural and practical question to ask here is: can the sample and computational complexity of the recovery algorithm be improved when $k \ll n$.

Recently, [60] studied this problem for Gaussian A and showed that for ℓ_1 regularized PhaseLift, $m = O(k^2 \log n)$ samples suffice for exact recovery of x^*. However, the computational complexity of this algorithm is still $O(n^3/\epsilon^2)$.

In this section, we provide a simple extension of our AltMinPhase algorithm that we call SparseAltMinPhase, for the case of sparse x^*. The main idea behind our algorithm is to first recover the support of x^*. Then, the problem reduces to phase retrieval of a k-dimensional signal. We then solve the reduced problem using Algorithm 6. The pseudocode for SparseAltMinPhase is presented in Algorithm 7. Table 3.2 provides a comparison of Algorithm 7 with ℓ_1-regularized PhaseLift in terms of sample complexity as well as computational complexity. The following lemma shows that if the number of measurements is large enough, step 1 of SparseAltMinPhase recovers the support of x^* correctly.

Lemma 3.6.1. *Suppose x^* is k-sparse with support S and $\|x^*\|_2 = 1$. If a_i*

are standard complex Gaussian random vectors and $m > \frac{c}{\left(x^*_{min}\right)^4} \log \frac{n}{\delta}$*, then Algorithm 7 recovers* S *with probability greater than* $1 - \delta$*, where* x^*_{min} *is the minimum non-zero entry of* x^**.*

The key step of our proof is to show that if $j \in supp(x^*)$, then random variable $Z_{ij} = \sum_i |a_{ij} y_i|$ has significantly higher mean than for the case when $j \notin supp(x^*)$. Now, by applying appropriate concentration bounds, we can ensure that $\min_{j \in supp(x^*)} |Z_{ij}| > \max_{j \notin supp(x^*)} |Z_{ij}|$ and hence our algorithm never picks up an element outside the true support set $supp(x^*)$. See Appendix B.2 for a detailed proof of the above lemma.

The correctness of Algorithm 7 now is a direct consequence of Lemma 3.6.1 and Theorem 3.5.3. For the special case where each non-zero value in x^* is from $\{-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}}\}$, we have the following corollary:

Corollary 3.6.2. *Suppose* x^* *is* k*-sparse with non-zero elements* $\pm\frac{1}{\sqrt{k}}$*. If the number of measurements* $m > c\left(k^2 \log \frac{n}{\delta} + k \log^2 k + k \log \frac{1}{\epsilon}\right)$*, then Algorithm 7 will recover* x^* *up to accuracy* ϵ *with probability greater than* $1 - \delta$*.*

3.7 Experiments

In this section, we present experimental evaluation of AltMinPhase (Algorithm 5) and compare its performance with the SDP based methods PhaseLift [17] and PhaseCut [90]. We also empirically demonstrate the advantage of our initialization procedure over random initialization (denoted by **AltMin (random init)**), which has thus far been considered in the literature [37, 31, 90, 12]. **AltMin (random init)** is the same as AltMinPhase except that step 1 of Algorithm 5 is replaced with:$x^0 \leftarrow$ Uniformly random vector from the unit sphere.

In the noiseless setting, a trial is said to **succeed** if the output x satisfies $\|x - x^*\|_2 < 10^{-2}$. For a given dimension, we do a linear search for smallest m (number of samples) such that empirical success ratio over 20 runs is at least 0.8. We implemented our methods in Matlab, while we obtained the code for PhaseLift and PhaseCut from the authors of [72] and [90] respectively.

We now present results from our experiments in three different settings.

Independent Random Gaussian Measurements: Each measurement vector a_i is generated from the standard complex Gaussian distribution. This measurement scheme was first suggested by [17] and till date, this is the only scheme with theoretical guarantees.

Multiple Random Illumination Filters: We now present our results for the setting where the measurements are obtained using multiple illumination filters; this setting was suggested by [12]. In particular, choose J vectors $z^{(1)}, \cdots, z^{(J)}$ and compute the following discrete Fourier transforms:

$$\widehat{x}^{(u)} = \mathrm{DFT}\left(x^* \cdot * z^{(u)}\right),$$

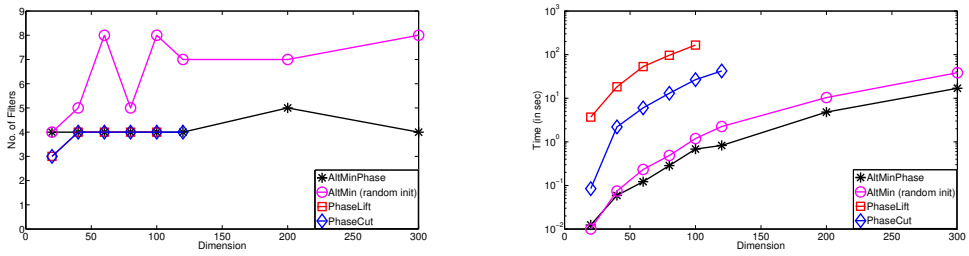
where $\cdot *$ denotes component-wise multiplication. Our measurements will then be the magnitudes of components of the vectors $\widehat{x}^{(1)}, \cdots, \widehat{x}^{(J)}$. The above measurement scheme can be implemented by modulating the light beam or by the use of masks; see [12] for more details.

For this setting, we conduct a similar set of experiments as the previous setting. That is, we vary dimensionality of the true signal $z^{(u)}$ (generated from the Gaussian distribution)and then empirically determine measurement and computational cost of each algorithm. Figures 3.2 (a) and (b) present our experimental results for this measurement scheme. Here again, we make similar observations as the last setting. That is, the measurement complexity of AltMinPhase is similar to PhaseCut and PhaseLift, but AltMinPhase is orders of magnitude faster than PhaseLift and PhaseCut. Note that Figure 3.2 is on a log-scale.

Noisy Phase Retrieval: Finally, we study our method in the following noisy measurement scheme:

$$y_i \;=\; |\langle a_i, x^* + w_i\rangle| \qquad \text{for } i = 1, \ldots, m, \tag{5}$$

where w_i is the noise in the i-th measurement and is sampled from $\mathcal{N}(0, \sigma^2)$. We fix $n = 64$ and $m = 6n$. We then vary the amount of noise added σ and measure the ℓ_2 error in recovery, i.e., $\|x - x^*\|_2$, where x is the recovered vector. Figure 3.3(a) compares the performance of various methods with varying amount of noise. We observe that our method outperforms PhaseLift and has similar recovery error as PhaseCut.



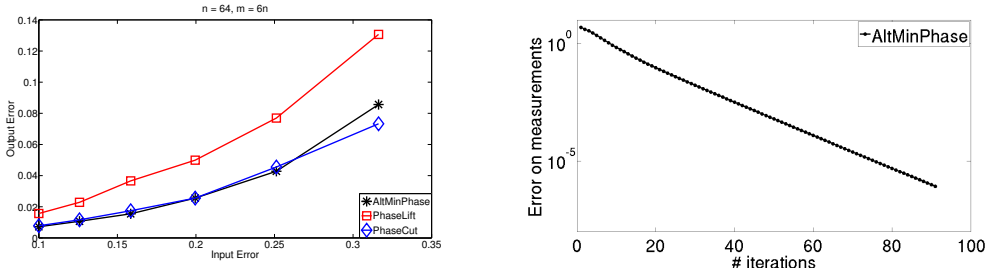
(a) **(b)**

Figure 3.2: Sample and time complexity for successful recovery using random Gaussian illumination filters. Similar to Figure 3.1, we observe that AltMinPhase has similar number of filters (J) as PhaseLift and PhaseCut, but is computationally much more efficient. We also see that AltMinPhase performs better than AltMin (randominit).

Geometric Decay: Finally, we provide empirical results verifying that AltMinPhase reduces the error at a geometric rate as guaranteed by Theorem 3.5.2 but no faster. The measurement vectors were chosen to be standard complex Gaussian with $n = 64$ and $m = 6n$. Figure 3.3(b) shows the plot of empirical error vs the number of iterations.

3.8 Summary

In this chapter, we presented an improved version of the alternating minimization procedure for the phase retrieval problem. We observe that empirically, it has simiar sample complexity as SDP based methods but is much more efficient than them. Analytically, we show that a natural resampled version of this algorithm has close to optimal sample complexity. We also extend our algorithm and results for the sparse phase retrieval problem.



(a) **(b)**

Figure 3.3: **(a)**: Recovery error $\|x - x^*\|_2$ incurred by various methods with increasing amount of noise (σ). AltMinPhase and PhaseCut perform comparably while PhaseLift incurs significantly larger error. **(b)**: Plot of empirical error $\left\| y - |A^T x| \right\|_2$ vs number of iterations for AltMinPhase. Each entry of A is chosen to be standard complex Gaussian with $n = 64$ and $m = 6n$. We can see that the error decreases geometrically suggesting that Theorem 3.5.2 is tight in some sense.

Chapter 4

Learning Sparsely Used Dictionaries using Alternating Minimization

4.1 Introduction

The [1] problem of dictionary learning can be stated as follows: given observations $Y \in \mathbb{R}^{d \times n}$, the task is to decompose it as

$$Y = A^* X^*, \quad A^* \in \mathbb{R}^{d \times r}, X^* \in \mathbb{R}^{r \times n}. \tag{1}$$

A^* is referred to as the *dictionary* matrix and X^* is the *coefficient* matrix. r denotes the number of basis elements in this dictionary, and we consider the overcomplete setting where $r \geq d$. Without further constraints, the solution to (1) is not unique. A popular framework is to assume that the coefficient matrix X^* is sparse, and that each observation $Y_i \in \mathbb{R}^d$ is a sparse combination of the dictionary elements (i.e. columns of the dictionary matrix). This problem is known as *sparse coding* and it has been argued that sparse coding can provide a succinct representation of the observed data, given only unlabeled samples [73, 55]. Through this lens of unsupervised learning, dictionary learning has recently received increased attention from the learning community [65, 5, 64].

Although several methods exist for sparse coding, most of them lack guarantees. [81] recently provided a method for guaranteed recovery when the dictionary matrix $A^* \in \mathbb{R}^{d \times r}$ is a basis. This implies that the number of dictionary elements $r \leq d$, where d is the observed dimension. However, in most settings, the dictionary is *overcomplete* ($r \gg d$) as overcomplete representations can provide greater flexibility in modeling as well as better robustness

[1] An extended abstract of the results in this chapter appeared as [2]. The coauthors on the paper had equal contributions in obtaining these results.

to noise [58, 7, 27]. In this paper, we establish *exact* recovery of sparsely used overcomplete dictionaries.

Summary of Results: We present a method for dictionary learning that consists of two phases. The initialization phase is a clustering-based procedure for recovering the dictionary to a rough accuracy. In particular, we establish that the recovery error of the initialization procedure, in ℓ_2 distance between true and recovered dictionary elements, is bounded by a small *constant* (dependent only on s) as long as the sparsity satisfies $s = O\left(d^{1/4}, r^{1/4}\right)$. The number of samples needed for this initialization procedure scales as $n = O\left(r(\log r + \log d)\right)$.

Our second result concerns the convergence to the global optimum of an alternating minimization scheme which outputs successively improved estimates of the coefficients and the dictionary through lasso and least-squares respectively. Our result requires the procedure to be initialized with a dictionary with an error of at most $O\left(1/s^2\right)$. Further when $s = O\left(d^{1/6}\right)$ and number of samples satisfies $n = O\left(r^2\right)$, we establish linear rate of convergence for the alternating minimization procedure to the true dictionary.

Combining the above two results, where we initialize the alternating method using our proposed dictionary estimation procedure with the required accuracy of $O\left(1/s^2\right)$, which entails $s = O\left(d^{1/9}, r^{1/8}\right)$, and sufficient number of samples $n = O\left(r^2\right)$, we guarantee exact recovery of the true dictionary. We believe that this is the first exact recovery result for dictionary learning in the overcomplete setting. Note that our alternating minimization guarantees are independent of the initialization procedure and it is entirely possible to use other initialization procedures for the alternating minimization algorithm. Indeed, the very recent and concurrent work of [3] can be seen as presenting alternative initialization procedures for our alternating minimization step.

Finally, we present some numerical simulations confirming the linear convergence of the alternating minimization procedure, and demonstrating the extent of gains beyond the initialization step. We also empirically test the recovery performance of the procedure, and find that it succeeds with $n = O\left(r\right)$ samples, hence suggesting room for tightening our analysis in future work.

Related Work: There have been many works on dictionary learning both from a theoretical and empirical viewpoint. Hillar and Sommer [40] consider conditions for identifiability of sparse coding. However, the number of samples required to establish identifiability is exponential in r for the general case. Most closely related to our work, [81] provide exact recovery results for an ℓ_1 based method, but they focus on the *undercomplete* setting, where $r \leq d$. We consider the overcomplete setting where $r > d$.

There exist many heuristics for dictionary learning, which work well in practice in many contexts, but lack theoretical guarantees. For instance, Lee et. al. propose an iterative ℓ_1 and ℓ_2 optimization procedure [55] similar to the the method of optimal directions [30]. Another popular method is the so-called K-SVD, which iterates between estimation of X and given an estimate of X, updates the dictionary estimate using a spectral procedure on the residual. Other works establish local optimality of the true solution (A^*, X^*) for certain non-convex programs [47, 36], but do not prescribe algorithms which can reach the true solution (A^*, X^*). Recent works [87, 65, 63, 82] provide generalization bounds for predictive sparse coding, without computational considerations.

Finally, our results are closely related to the very recent work of [3], carried out independently and concurrently with our work. Their approximate recovery work can be seen as providing a different initialization strategy for alternating minimization procedure. However, the key distinction between our alternating minimization procedure as compared to theirs is that we use the *same* samples in each iteration while they require fresh samples for each iteration of alternating minimization. This enables us to obtain exact recovery of the dictionary once $n = \Omega(r^2)$, whereas the error in their method can not be guaranteed to be below $\exp\left(-O\left(n/r^2\right)\right)$. Our algorithm is also robust in the sense that we do not expect to recover the complete support in the first iteration – we gradually recover more and more elements of the support as our dictionary estimate gets better.

The remainder of the paper is organized as follows. We present our algorithms next, followed by our assumptions and the recovery results. We provide proof sketches in Section 3 with details deferred to the supplement. Simulation results are described in Section 4.

4.2 Algorithm

Notation: Let $[n] := \{1, 2, \ldots, n\}$. For a vector v or a matrix W, we will use the shorthand $\text{Supp}(v)$ and $\text{Supp}(W)$ to denote the set of non-zero entries of v and W respectively. Let $\|w\|$ denote the ℓ_2 norm of vector w, and similarly for a matrix W, $\|W\|$ denotes its spectral norm. For a matrix X, X^i, X_i and X_j^i denote the i^{th} row, i^{th} column and $(i,j)^{\text{th}}$ element of X respectively. For a graph $G = (V, E)$, let $\mathcal{N}_G(i)$ denote set of neighbors for node i in G.

4.2.1 Initial Estimate of Dictionary Matrix

The first step is to obtain an initial estimate $\widehat{A}$ of the dictionary elements, and is given in Algorithm 8. The estimate $\widehat{A}$ is then employed in alternating steps to estimate the coefficient matrix and re-estimate the dictionary matrix respectively.

Given samples Y, we first construct the correlation graph $G_{\text{corr}(\rho)}$, where the nodes are samples $\{Y_1, Y_2, \ldots Y_n\}$ and an edge $(Y_i, Y_j) \in G_{\text{corr}(\rho)}$ implies that $|\langle Y_i, Y_j \rangle| > \rho$, for some threshold $\rho > 0$ (Figure 4.1 shows an example of a typical correlation graph under our assumptions). We then determine a good subset of samples via a *clustering* procedure on the graph as follows: we first randomly sample an edge $(Y_{i^*}, Y_{j^*}) \in G_{\text{corr}(\rho)}$ and then consider the intersection of the neighborhoods of Y_{i^*} and Y_{j^*}, denoted by $\widehat{S}$. We then employ UniqueIntersection routine in Procedure 1 to determine if $\widehat{S}$ is a "good set" for estimating a dictionary element. This is done by ensuring that the set $\widehat{S}$ has a sufficient number of edges[2] in the correlation graph. For instance, the procedure will return true when evaluated on the green edges labeled *Good*, but false on the red edges labeled *Bad*. Once $\widehat{S}$ is determined to be a good set, we then proceed by estimating the matrix $\widehat{Q}$ using samples in $\widehat{S}$ and output its top singular vector as the estimate of a dictionary element. The method is repeated over all edges in the correlation graph to ensure that all the dictionary elements get estimated with high probability.

[2]For convenience to avoid dependency issues, in Procedure 1, we partition $\widehat{S}$ into sets consisting of disjoint node pairs and determine if there are sufficient number of node pairs which are neighbors.

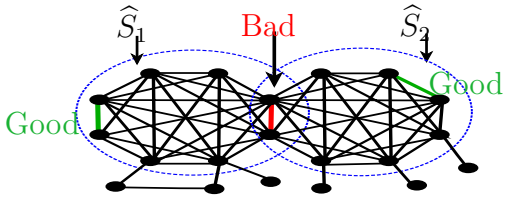


Figure 4.1: Sample correlation graph G_{corr} with nodes $\{Y_k\}$ and edge (Y_i, Y_j) s.t. $|\langle Y_i, Y_j \rangle| > \rho$. $\widehat{S}_1, \widehat{S}_2$ are the sets returned as true from UniqueIntersection procedure. The edges labeled "good" above refers to good anchor pairs which satisfy unique intersection in Algorithm 8,while the bad anchor pair does not satisfy the unique intersection. Good anchor pairs lead to formation of sets $\widehat{S}_1$ and $\widehat{S}_2$.

At a high level, the above procedure aims to find large cliques in the correlation graph. For instance, in Figure 4.1, the sets $\widehat{S}_1, \widehat{S}_2$ are the sets which are returned as true by the UniqueIntersection Procedure and the algorithm 8 computes SVD over the samples in such sets. Intuitively, when the correlation graph has cliques with small amount of overlap, our method succeeds in finding them, and then computes SVD over the samples in such sets. At a high level, the above procedure aims to find large cliques in the correlation graph. For instance, in Figure 4.1, the sets $\widehat{S}_1, \widehat{S}_2$ are the sets which are returned as true by the UniqueIntersection Procedure, when the node pairs labeled as "good" in the figure are used as anchor samples Y_{i^*} and Y_{j^*}. On the other hand, note that a bad anchor pair which sits at the overlap of multiple cliques is not returned as true by the UniqueIntersection Procedure. Thus, this procedure yields subsets of samples which correspond to large cliques in the correlation graph. Once, such a subset is found, the algorithm 8 computes SVD over the samples in such sets. As our proofs will demonstrate, any such clique $\widehat{S}_i$ involves samples that all contain a *unique* dictionary element in common, which can then be recovered approximately by the subsequent SVD step.

4.2.2 Alternating Minimization

Once an initial estimate of the dictionary is obtained, we alternate between two procedures, viz., a sparse recovery step for estimating the coefficients given a dictionary, and a least squares step for a dictionary given the

Algorithm 8 InitDictionaryLearn$(Y, \epsilon_{\text{dict}}, \rho)$: Initial step for estimating dictionary elements.

Input: Samples $Y = [Y_1|\dots|Y_n]$. Correlation threshold ρ. Desired separation parameter ϵ_{dict} between recovered dictionary elements.

Output: Initial Dictionary Estimate $\bar{A}$.

```
Construct correlation graph G_{corr(ρ)} s.t. (Y_i, Y_j) ∈ G_{corr(ρ)} when |⟨Y_i, Y_j⟩| > ρ.
Set Ā ← ∅.
for each edge (Y_{i*}, Y_{j*}) ∈ G_{corr(ρ)} do
  Ŝ ← N_{G_{corr(ρ)}}(Y_{i*}) ∩ N_{G_{corr(ρ)}}(Y_{j*}).
  if UniqueIntersection(Ŝ, G_{corr(ρ)}) then
    Q̂ ← Σ_{Y_i ∈ Ŝ} Y_i Y_i^⊤ and ā ← u_1, where u_1 is top singular vector of Q̂.
    if min_{b∈Ā} ‖ā − b‖ > 2ε_dict then
      Ā ← Ā ∪ ā
    end if
  end if
end for
Return Ā
```

estimates of the coefficients (details are presented in Algorithm 2).

The sparse recovery step of Algorithm 2 is based on ℓ_1-regularization, followed by thresholding. The thresholding is required for us to guarantee that the support set of our coefficient estimate $X(t)$ is a *subset* of the true support with high probability. Once we have an estimate of the coefficients, the dictionary is re-estimated through least squares. The overall algorithmic scheme is popular for dictionary learning, and there are a number of variants of the basic method. For instance, the ℓ_1-regularized problem in step 3 can also be replaced by other robust sparse recovery procedures such as OMP [84] or GraDeS [35]. More generally the exact lasso and least-squares steps may be replaced with other optimization methods for computational efficiency, e.g. [47].

Procedure 1 UniqueIntersection(S, G): Determine if samples in S have a unique intersection.

Input: Set S with $2m$ vectors $Y_1, \dots Y_{2m}$ and graph G with $Y_1, \dots, Y_{2m}$ as nodes.

Output: Indicator variable UNIQUE_INT

Partition S into sets $S_1, \dots, S_m$ such that each $|S_i| = 2$.
if Number of S_i which are edges in G is greater than $\frac{61m}{64}$ **then**
  UNIQUE_INT $\leftarrow 1$
else
  UNIQUE_INT $\leftarrow 0$
end if
Return UNIQUE_INT

4.3 Guarantees

In this section, we provide our exact recovery result and also clearly specify all the required assumptions on A^* and X^*. We then provide guarantees for each of the individual steps (initialization step and alternating minimization steps) in Section 4.3.2 and Section 4.3.3, respectively. We provide a brief sketch of our proof for each of the steps in Section 4.3.4.

4.3.1 Assumptions and exact recovery result

We start by formally describing the assumptions needed for the main recovery result of this paper.

Assumptions on the dictionary:

($A1$) **Mutual Incoherence:** Wlog, assume that all the elements are normalized: $\|A_i^*\| = 1$, for $i \in [r]$. We assume pairwise incoherence condition on the dictionary elements, for some constant $\mu_0 > 0$, $|\langle A_i^*, A_j^* \rangle| < \frac{\mu_0}{\sqrt{d}}$.

($A2$) **Bound on the Spectral Norm:** The dictionary matrix has bounded spectral norm, i.e., for some $\mu_1 > 0$, we have $\|A^*\| < \mu_1 \sqrt{\frac{r}{d}}$.

Algorithm 2 AltMinDict$(Y, A(0), \epsilon_0)$: Alternating minimization for dictionary learning

Input: Samples Y, initial dictionary estimate $A(0)$, accuracy sequence ϵ_t and sparsity level s. Thresholding function $\mathcal{T}_\rho(a) = a$ if $|a| > \rho$ and 0 o.w.

1: **for** iterations $t = 0, 1, 2, \ldots, T-1$ **do**
2: **for** samples $i = 1, 2, \ldots, n$ **do**
3: $X(t+1)_i = \arg\min_{x \in \mathbb{R}^r} \|x\|_1$ such that, $\|Y_i - A(t)x\|_2 \leq \epsilon_t$.
4: **end for**
5: Threshold: $X(t+1) = \mathcal{T}_{9s\epsilon_t}(X(t+1))$.
6: Estimate $A(t+1) = YX(t+1)^+$
7: Normalize: $A(t+1)_i = \frac{A(t+1)_i}{\|A(t+1)_i\|_2}$
8: **end for**

Output: $A(T)$

Assumptions on the coefficients:

($B1$) **Non-zero Entries in Coefficient Matrix:** We assume that the non-zero entries of X^* are drawn i.i.d. from a zero-mean unit-variance distribution, and satisfy the following a.s.: $m \leq |X^{*i}_j| \leq M, \forall i, j$.

($B2$) **Sparse Coefficient Matrix:** The columns of coefficient matrix have s non-zero entries which are selected uniformly at random from the set of all s-sized subsets of $[r]$, i.e. $|\operatorname{Supp}(X^*_i)| = s$, $\forall\, i \in [n]$.We require s to satisfy

$$s < \min\left(\frac{r^{1/8}}{c_1}\left(\frac{m}{M}\right)^{1/4}, \frac{d^{1/9}}{c_2 \mu_1^{2/9}}\left(\frac{m}{M}\right)^{4/9}\right),$$

for universal constants $c_1, c_2 > 0$. Constants m, M are as specified above.

Assumption ($A1$) on normalization of dictionary elements is without loss of generality since we can always rescale the dictionary elements and the corresponding coefficients and obtain the same observations. However, the incoherence assumption is crucial in establishing our guarantees. In particular, incoherence also leads to a bound on the restricted isometry property (RIP) constant [75]; see Lemma C.3.2 in Appendix C.2. The assumption ($A2$)

provides a bound on the spectral norm of A^*. Note that the incoherence and spectral assumptions are satisfied with high probability (w.h.p.) when the dictionary elements are randomly drawn from a mean-zero sub-gaussian distribution.

Assumption $(B1)$ imposes some natural constraints on lower and upper bounds on the non-zero entries of X^*. We use lower bound assumption on $X^*(i,j)$ for simplicity of exposition, as explained in Section 4.3.4, we can remove this assumption as the thresholding coefficient in Algorithm 2 decreases with each iteration. Assumption$(B2)$ on sparsity in the coefficient matrix is crucial for identifiability of the dictionary learning problem.

We now give the main result of this paper.

Theorem 4.3.1 (Exact recovery)**.** *Suppose assumptions* $(A1) - (A2)$ *and* $(B1) - (B2)$ *are satisfied. Then there exists a universal constant* c_3 *such that, if*

1. **Sample Complexity:** $n \geq c_3 \max\left(r^2 \log \frac{1}{\delta}, rM^2 s \log \frac{2r}{\delta}\right),$
2. **Choice of Parameters for Initial Dictionary Estimation:** *inputs* ρ *and* ϵ_{dict} *to Algorithm 8 are chosen such that*
$$\rho = \frac{1}{2} - \frac{s^2\mu_0}{\sqrt{d}} > 0, \text{ and } \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right) < \epsilon_{\text{dict}}^2 < \frac{1}{4}.$$
3. **Choice of Parameters for Alternating Minimization:** *Algorithm 2 uses a sequence of accuracy parameters* $\epsilon_0 = 1/2592s^2$ *and*
$$\epsilon_{t+1} = \frac{25050\mu_1 s^3}{\sqrt{d}}\epsilon_t. \tag{2}$$

then, the alternating minimization procedure (Algorithm 2) when seeded with Algorithm 8, outputs $A(t)$ *at the* t*-th step* ($t \geq 1$) *that satisfies the following with probability at least* $1 - 2\delta - 2n^2\delta$*:*

$$\min_{z\in\{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \sqrt{2}\epsilon_t, \ \forall 1 \leq i \leq r,$$

where ϵ_t *is as given in Assumption* $(A7)$*. In particular, after* $T = O(\log(\frac{\epsilon_0}{\epsilon}))$ *steps of Algorithm 2, we obtain:*

$$\min_{z\in\{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \epsilon, \ \forall 1 \leq i \leq r, \forall \epsilon > 0.$$

Remarks: Note that we have a sign ambiguity in recovery of the dictionary elements, since we can exchange the signs of the dictionary elements and the coefficients to obtain the same observations.

Note that Theorem 4.3.1 guarantees that we can recover the dictionary A^* to an arbitrary precision ϵ (based on the number of iterations T of Algorithm 2), given $n = O(r^2)$ samples. We contrast this with the results of [3], who also provide recovery guarantees to an arbitrary accuracy ϵ, but only if the number of samples is allowed to increase as $O\left(r^2 \log \frac{1}{\epsilon}\right)$.

Establishing the above result requires two main ingredients, viz., guaranteeing an error bound for the initial dictionary estimation step, and proving a local convergence result for the alternating minimization step, and obtaining a bound on the *basin of attraction* for the solution consisting of the true dictionary and coefficient matrices. Below, we provide these individual results explicitly.

4.3.2 Guarantees for the Initialization Step

We now give the result for approximate recovery of the dictionary in the initialization step.

Theorem 4.3.2 (Approximate recovery of dictionary)**.** *Suppose the output of Algorithm 8 is $A(0)$. Under assumptions $(A1)-(A2)$ and $(B1)-(B2)$, and if*

1. **Sample Complexity:** $n \geq c_3 \max\left(r^2 \log \frac{1}{\delta}, rM^2 s \log \frac{2r}{\delta}\right)$,
2. **Choice of Parameters for Initial Dictionary Estimation:** *inputs ρ and ϵ_{dict} to Algorithm 8 are chosen such that*

$$\rho = \frac{1}{2} - \frac{s^2\mu_0}{\sqrt{d}} > 0, \text{ and } \frac{32sM^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right) < \epsilon_{\text{dict}}^2 < \frac{1}{4}.$$

then, with probability greater than $1 - 2n^2\delta$, there exists a permutation matrix P such that:

$$\epsilon_A^2 := \max_{i\in[r]} \min_{z\in\{-1,+1\}} \|zA_i^* - (PA(0))_i\|_2^2 < 32s\frac{M^2}{m^2}\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \frac{1}{s^3}\right).$$

Remarks: We note that the error in Theorem 4.3.2 does not go down with the number of samples n, since it depends on geometric properties of the dictionary, that are determined by the dimension dependent factors such as s, r and d. However, the error probability does go down with the number of samples, since the sample correlation graph becomes an increasingly accurate representative of the population version.

For the approximate recovery of dictionary elements, it turns out that a less stringent requirement on the sparsity level and the sample complexity suffices. Specifically, we can replace assumption $(B2)$ with the weaker condition $s < \min\left(\frac{m}{M}\sqrt{\frac{\sqrt{d}}{2\mu_0}}, \sqrt[3]{\frac{r}{1536}}\right)$, which suffices for the error in Theorem 4.3.2 to be $o(1)$. The more stringent requirement on sparsity arises in Theorem 4.3.1 since we need the error from Theorem 4.3.2 to be at most $O\left(1/s^2\right)$ for the subsequent alternating minimization steps to succeed. The initialization step also has a milder requirement on the number of samples, and does not need the condition $n = O\left(r^2 \log(1/\delta)\right)$. Thus, we obtain a near linear sample complexity for our initialization method.

4.3.3 Guarantees for Alternating Minimization

We now prove a local convergence result for alternating minimization. We assume that we have access to a good initial estimate of the dictionary:

$(C1)$ **Initial dictionary with guaranteed error bound:** We assume that we have access to an initial dictionary estimate $A(0)$ such that

$$\epsilon_0 := \max_{i \in [r]} \min_{z \in \{-1,+1\}} \|z A_i^* - A(0)_i\|_2 < \frac{1}{2592 s^2}.$$

Theorem 4.3.3 (Local linear convergence)**.** *Under assumptions* $(A1)$-$(A2)$*,* $(B1)$-$(B2)$ *and* $(C1)$*, if*

1. **Sample Complexity:** $n \geq c_3 \max\left(r^2 \log \frac{1}{\delta}, r M^2 s \log \frac{2r}{\delta}\right)$,
2. **Choice of Parameters for Alternating Minimization:** *Algorithm 2 uses a sequence of accuracy parameters* $\epsilon_0 = 1/2592 s^2$ *and*

$$\epsilon_{t+1} = \frac{25050 \mu_1 s^3}{\sqrt{d}} \epsilon_t. \tag{3}$$

then, with probability at least $1 - 2\delta$ *the iterate* $A(t)$ *of Algorithm 2 satisfies the following for all* $t \geq 1$*:*

$$\min_{z \in \{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \epsilon_t, 1 \leq i \leq r.$$

Remarks: The consequences of Theorem 4.3.3 are powerful combined with our Assumption $(B2)$ and the recurrence 3 (since $(B2)$ ensures that ϵ_t forms a decreasing sequence). In particular, it is implied that with high probability we obtain,

$$\min_{z \in \{-1,1\}} \|zA(t)_i - A^*{}_i\|_2 \leq \|A(0) - A^*\|_2 2^{-t}.$$

Given the above bound, we need at most $O\left(\log_2 \frac{\epsilon_0}{\epsilon}\right)$ in order to ensure $\|zA(T)_i - A^*{}_i\|_2 \leq \epsilon$ for all the dictionary elements $i = 1, 2, \ldots, r$. In the convex optimization parlance, the result demonstrates a local linear convergence of Algorithm 2 to the globally optimal solution under an initialization condition. Another way of interpreting our result is that the global optimum has a *basin of attraction* of size $O\left(1/s^2\right)$ for our alternating minimization procedure under these assumptions (since we require $\epsilon_0 \leq O\left(1/s^2\right)$).

We note that Theorem 4.3.3 does not crucially rely on initialization specifically by the output of Algorithm 8, and admits any other initialization satisfying Assumption $(C1)$. In particular, some of the assumptions in $(B1) - (B2)$ are not essential for Theorem 4.3.3, but are only made for the overall result of Theorem 4.3.1. Indeed, it suffices to have a sparsity level satisfying $s < \frac{d^{1/6}}{c_2 \mu_1^{1/3}}$ for a universal constant $c_2 > 0$ (without any dependence on r). The theorem also does not rely on lower bounded entries, and only needs $\|X^*\|_\infty \leq M$. We also recall that the lasso step in Algorithm 2 can be replaced with a different robust sparse recovery procedure, with qualitatively similar results.

As remarked earlier, the recent work of [3] provides an alternative initialization strategy for our alternating minimization procedure. Indeed, under our sample complexity assumption, their OverlappingAverage method provides a solution with $\epsilon_0 = O\left(s/\sqrt{r}\right)$ assuming $s = O\left(\max(r^{2/5}, \sqrt{d})\right)$.

4.3.4 Overview of Proof

In this section we outline the key steps in proving Theorems 4.3.2 and 4.3.3. Given these theorems, Theorem 4.3.1 follows as an immediate consequence.

Analysis of initial dictionary estimation: The core intuitions for this step can be described in terms of the relationships between the two graphs, viz., the coefficient bipartite graph B_{coeff} and the sample correlation graph G_{corr}, shown in Figures 4.2 and 4.1 respectively. B_{coeff} consists of dictionary elements $\{A_i^*\}$ on one side and the samples $\{Y_i\}$ on the other. There is an edge between Y_i and A_j^* iff ${X^*}_j^i \neq 0$, and $\mathcal{N}_B(Y_i)$ denotes the neighborhood of Y_i in the graph B_{coeff}.

Now given this bipartite graph B_{coeff}, for each dictionary element A_i^*, consider a set of samples[3] which (pairwise) have only one dictionary element A_i^* in common, and denote such a set by $\mathcal{C}_i$ i.e. $\mathcal{C}_i := \{Y_k, k \in S : \mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l) = A_i^*,\ \forall\, k, l \in S\}$. Intuitively, the sets $\widehat{S}$ constructed in Algorithm 8 are our proxies for $\mathcal{C}_i$. Indeed, the first part of the proof is to demonstrate that for a random coefficient matrix X^*, adequately large cliques $\mathcal{C}_i$ exist in the graph B_{coeff}.

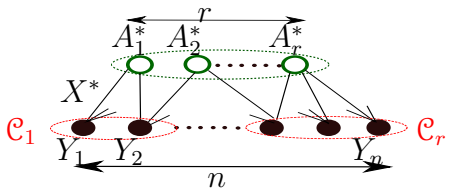


Figure 4.2: Bipartite graph B mapping dictionary elements $A_1^*, \ldots A_r^*$ to samples $Y_1, \ldots Y_n$. See text for definition of $\mathcal{C}_i$.

Our subsequent analysis is broadly divided into two parts, viz., establishing that (large) sets $\{\mathcal{C}_i\}$ can be found efficiently, and that the dictionary elements can be estimated accurately once such sets $\{\mathcal{C}_i\}$ are found. We start

[3]Note that such a set need not be unique.

with a proposition that demonstrates the correctness of Procedure 1 at identifying these cliques. We use the notation Uniq-intersect(Y_i, Y_j) to denote that Y_i and Y_j have exactly one dictionary element in common.

Proposition 4.3.4 (Correctness of Procedure 1)**.** *Suppose* $(Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}$. *Suppose that* $s^3 \leq r/1536$ *and* $\gamma \leq 1/64$. *Then Algorithm 8 returns the value of* Uniq-intersect(Y_{i^*}, Y_{j^*}) *correctly with probability at least* $1 - 2\exp(-\gamma^2 m)$.

Given a large sample of elements with a unique dictionary element (say A_1^*) in common ($\widehat{S}$ in Algorithm 8), we next show that the subsequent SVD step recovers this dictionary element approximately. Intuitively this happens since each sample $Y_i \in \widehat{S}$ contains A_1^* with a coefficient at least m (in absolute value). Hence the covariance matrix $\widehat{Q}$ has a larger component along A_1^* than other dictionary elements, which leads to approximate recovery via the top singular vector.

Proposition 4.3.5 (Accuracy of SVD)**.** *Consider anchor samples* Y_{i^*} *and* Y_{j^*} *such that* Uniq-intersect(Y_{i^*}, Y_{j^*}) *in Algorithm 8 is satisfied, and wlog, let* $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A_1^*\}$. *Recall the definition of* $\widehat{S}$ *in Algorithm 8, and further define* $\widehat{Q} := \sum_{i \in \widehat{S}} Y_i Y_i^\top$ *and* $|\widehat{S}| = k$. *If* $\widehat{a}$ *is the top singular vector of* $\widehat{Q}$, *then there exists a universal constant* c *such that for any* $0 < \alpha < 1/20$ *we have:*

$$\min_{z \in \{-1,1\}} \|z\widehat{a} - A_1^*\|_2^2 < 32s\left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right),$$

with probability at least $1 - d\exp\left(-c\alpha^2 k\right)$.

Note the ambiguity in signs above, since SVD cannot recover the sign of the top singular vector. With the above auxiliary results in place, the proof of Theorem 4.3.2 follows with simple arguments.

Analysis of alternating minimization: Given an approximate estimate of the dictionary, we then establish a local convergence result for alternating minimization.

For ease of notation, let us consider just one iteration of Algorithm 2 and denote $X(t+1)$ as X, $A(t+1)$ as A and $A(t)$ as $\widetilde{A}$. Then we have the least-squares update

$$A - A^* = YX^+ - A^* = A^*X^*X^+ - A^*XX^+ = A^* \triangle X X^+,$$

where $\triangle X = X^* - X$. This means that we can understand the error in dictionary recovery by the error in the least squares operator $\triangle X X^+$. In particular, we can further expand the error in a column p as: $A_p - {A^*}_p = {A^*}_p (\triangle X X^+)^p_p + {A^*}_{\setminus p} (\triangle X X^+)^{\setminus p}_p$, where the notation $\setminus p$ represents the collection of all indices apart from p. Hence we see two sources of error in our dictionary estimate. The element $(\triangle X X^+)^p_p$ causes the rescaling of A_p relative to ${A^*}_p$. However, this is a minor issue since the renormalization would correct it.

More serious is the contribution from the off-diagonal terms $(\triangle X X^+)^p_{\setminus p}$, which corrupt our estimate A_p with other dictionary elements beyond ${A^*}_p$. Indeed, a crucial argument in our proof is controlling the contribution of these terms at an appropriately small level. In order to do that, we start by controlling the magnitude of $\triangle X$.

Lemma 4.3.6 (Error in sparse recovery)**.** *Let* $\triangle X \stackrel{def}{=} X(t) - X^*$. *Assume that* $2\mu_0 s/\sqrt{d} \leq 0.1$ *and* $\sqrt{s\epsilon_t} \leq 0.1$ *Then, we have* $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X^*)$ *and the error bound* $\|\triangle X\|_\infty \leq 9 s \epsilon_t$.

This lemma is very uesful in our error analysis, since we establish that any matrix W satisfying $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$ has a good bound on its spectral norm (even if the entries depend on A^*, X^*).

Lemma 4.3.7. *With probability at least* $1 - r\exp\left(-\frac{Cn}{rs}\right)$, *for every* $r \times n$ *matrix* W *s.t.* $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$, *we have* $\|W\|_2 \leq 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}}$.

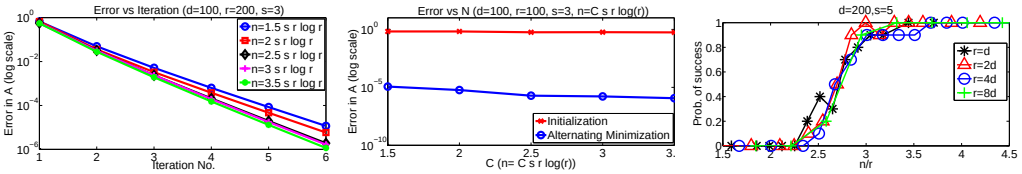
A particular consequence of this lemma is that it guarantees the invertibility of the matrix $XX^\top$, so that the pseudo-inverse X^+ is well-defined for subsequent least squares updates. Next, we present the most crucial step which is controlling the off-diagonal terms $(\triangle X X^+)^p_{\setminus p}$.

Lemma 4.3.8 (Off-diagonal error bound)**.** *With probability at least* $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2 s}\right) - \exp\left(-n/(3r^2)\right)$, *we have uniformly for every* $p \in [r]$ *and every* $\triangle X$ *such that* $\|\triangle X\|_\infty < \frac{1}{288 s}$.

$$\left\| \left(\triangle X X^+\right)^{\setminus p}_p \right\|_2 = \left\| \left(X^* X^+\right)^{\setminus p}_p \right\|_2 \leq \frac{1968 s^2 \|\triangle X\|_\infty}{\sqrt{r}}.$$

The lemma uses the earlier two lemmas along with some other auxilliary results. Given these lemmas, the proof of the main theorem follows with some algebra. Specifically, for any unit vector w such that $w \perp A^*{}_p$, we can bound the normalized inner product $\langle w, A_p \rangle / \|A_p\|_2$ which suffices to obtain the result of the theorem.

4.4 Experiments



(a) (b) (c)

Figure 4.3: (a): Average error after each step alternating minimization step of Algorithm 2 on log-scale. (b): Average error after the initialization procedure (Algorithm 8) and after 5 alternating minimization steps of Algorithm 2. (c): Sample complexity requirement of the alternating minimization algorithm. For ease of experiments, we initialize the dictionary using a random perturbation of the true dictionary rather than using Algorithm 8 which should in fact give better initial point with smaller error.

Alternating minimization/descent approaches have been widely used for dictionary learning and several existing works show effectiveness of these methods on real-world/synthetic datasets [5, 82]. Hence, instead of replicating those results, in this section we focus on illustrating the following three key properties of our algorithms via experiments in a controlled setting: a) Advantage of alternating minimization over one-shot initialization, b) linear convergence of alternating minimization, c) sample complexity of alternating minimization.

Data generation model: Each entry of the dictionary matrix A is chosen i.i.d. from $\mathcal{N}(0, 1)$. Note that, random Gaussian matrices are known to satisfy incoherence and the spectral norm bound [88]. The support of each column of X was chosen independently and uniformly from the set of all s-subsets of $[r]$. Similarly, each non-zero element of X was chosen independently from the uniform distribution on $[-2, -1] \cup [1, 2]$. We use the GraDeS algorithm of

[35] to solve the sparse recovery step, as it is faster than lasso. We measure error in the recovery of dictionary by $error(A) = \max_i \sqrt{1 - \frac{\langle A_i, A_i^* \rangle^2}{\|A_i\|_2^2 \|A_i^*\|_2^2}}$. The first two plots are for a typical run and the third plot averages over 10 runs. The implementation is in Matlab.

Linear convergence: In the first set of experiments, we fixed $d = 100$, $r = 200$ and measured error after each step of our algorithm for increasing values of n. Figure 4.3 (a) plots error observed after each iteration of alternating minimization; the first data point refers to the error incurred by the initialization method. As expected due to Theorem 4.3.3, we observe a geometric decay in the error.

One-shot vs iterative algorithm: It is conceivable that the initialization procedure of Algorithm 8 itself is sufficient to obtain an estimate of the dictionary upto reasonable accuracy. Figure 4.3(b) shows that this is not the case. The figure plots the error in recovery vs the number of samples used for both Algorithm 8 and Algorithm 2. It is clear that the recovery error of the alternating minimization procedure is significantly smaller than that of the initialization procedure. For example, for $n = 2.5sr \log r$ with $s = 3, r = 200, d = 100$, initialization incurs error of .56 while alternating minimization incurs error of 10^{-6}. Note however that the recovery accuracy of the initialization procedure is non-trivial and also crucial to the success of alternating minimization- a random vector in $\mathbb{R}^d$ would give an error of $1 - \frac{1}{d} = 0.99$, where as the error after initialization procedure is ≈ 0.55.

Sample complexity: Finally, we study sample complexity requirement of the alternating minimization algorithm which is $n = O(r^2 \log r)$ according to Theorem 4.3.3, assuming good enough initialization. Figure 4.3(c) suggests that in fact only $O(r)$ samples are sufficient for success of alternating minimization. The figure plots the probability of success with respect to $\frac{n}{r}$ for various values of r. A trial is said to succeed if at the end of 25 iterations, the error is smaller than 10^{-6}. Since we focus only on the sample complexity of alternating minimization, we use a faster initialization procedure: we initialize the dictionary by randomly perturbing the true dictionary as $A(0) = A^* + Z$, where each element of Z is an $\mathcal{N}(0, 0.5)$ random variable. Figure 4.3 (c) shows that the success probability transitions at nearly the same value for various values of r, suggesting that the sample complexity of the alternating minimization procedure in this regime of $r = O(d)$ is just $O(r)$.

4.5 Discussion

In this paper we present an exact recovery result for learning incoherent and overcomplete dictionaries with sparse coefficients. The first part of our result uses a novel initialization procedure, which uses a clustering-style algorithm to approximately recover the dictionary elements. The second step of our approach is an alternating minimization procedure which is quite widely used by practitioners for this problem already. We believe that our results are an important and timely advance in the understanding of this problem. There is an increasing interest on supervised and unsupervised feature learning methods in machine learning. However, we have an extremely rudimentary theoretical understanding of these problems as compared to standard classification of regression problems. A systematic understanding of dictionary learning and related models (both supervised and unsupervised) can help bridge this gap. Moreover, the applications of dictionary learning in other areas such as signal processing and coding make these results of broader interest beyond machine learning.

We believe that our work suggests several avenues for future research. We focus on the unsupervised setting in this paper, but extensions to supervised setting would be interesting for future work. Our theory also suggests room for strengthening the lasso step with further constraints on the global structure of the iterates $X(t)$, which might lead to better recovery properties with milder assumptions. Our simulations hint at the possibility of a better sample complexity, at least in certain regimes of parameters. Understanding these issues, as well as others such as noise robustness remain important questions for further research in this area.

Chapter 5

Conclusion

Alternating minimization algorithms are widely used for solving many non-convex learning problems. Despite their good empirical performance, there have been very few theoretical guarantees on their performance. In this thesis, we present rigorous performance guarantees for alternating minimization for three machine learning problems: matrix completion, phase retrieval and learning sparsely used dictionaries.

Understanding why alternating minimization and other such heuristics work so well in practice seems crucial to improving upon these methods as well as in designing new methods with better performance. A crucial component of our results for all the three problems is the designing of new initialization algorithms from where alternating minimization is guaranteed to converge at a good rate. For the phase retrieval problem, we indeed observe that principled initialization improves sample complexity over random initialization (see Figures 3.1 and 3.2). It will be interesting to see if our initialization algorithms improve the performance of alternating minimization in practice.

Alternating minimization is also closely related to Expectation Maximization (EM), which is the predominant method used in practice for many statistical problems. Similar to alternating minimization, despite its huge empirical success, there are very few results regarding its performance in any setting. It will be interesting to see if our methods help shed light on the performance of EM in any setting.

We believe that the ultimate goal of this line of research is to leverage our understanding of the performance of these methods to design faster algorithms with good performance. For instance, the success of alternating minimization methods naturally motivates the designing of gradient and proximal gradient methods to solve these problems. We believe that successful designing of such algorithms will have a big impact on many applications.

Appendices

Appendix A

Proofs for Matrix Completion using Alternating Minimization

A.1 Preliminaries

Lemma A.1.1 (Lemma 2.1 of [44]). *Let $b = \mathcal{A}(M) + e$, where e is a bounded error vector, M is a rank-k matrix and $\mathcal{A}$ is a linear measurement operator that satisfies $2k$-RIP with constant δ_{2k} (assume $\delta_{2k} < 1/3$). Let X^{t+1} be the $t+1$-th step iterate of SVP, then the following holds:*

$$\|\mathcal{A}(X^{t+1}) - b\|_2^2 \leq \|\mathcal{A}(M) - b\|_2^2 + 2\delta_{2k}\|\mathcal{A}(X^t) - b\|_2^2.$$

In our analysis, we heavily use the following two results. The first result is the well-known Bernstein's inequality.

Lemma A.1.2. *[Bernstein's inequality] Let $X_1, X_2, \cdots, X_n$ be independent random variables. Also, let $|X_i| \leq L \in \mathbb{R}\ \forall\ i$ w.p. 1. Then, we have the following inequality:*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right| > t\right] \leq 2\exp\left(\frac{-t^2/2}{\sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) + Lt/3}\right). \tag{1}$$

The second result is a restatement of Theorem 3.1 from [50].

Theorem A.1.3. *(Restatement of Theorem 3.1 from [50]) Suppose M is an incoherent rank-k matrix and let p, Ω be as in Theorem 2.3.3. Further, let M_k be the best rank-k approximation of $\frac{1}{p}P_\Omega\left(M\right)$. Then, w.h.p. we have:*

$$\left\|M - M_k\right\|_2 \leq C\sqrt{\frac{k}{p\sqrt{mn}}}\left\|M\right\|_F. \tag{2}$$

Remark: Note that Theorem 3.1 from [50] holds only for $T_r\left(P_\Omega(M)\right)$ where $T_r\left(P_\Omega(M)\right)$ is a trimmed version of $P_\Omega(M)$ obtained by setting all rows and columns of $P_\Omega(M)$ with too many observed entries to zero. However, using standard Chernoff bound we can argue that for our choice of p, none of the rows and columns of $P_\Omega(M)$ have too many observed entries and hence $T_r\left(P_\Omega(M)\right) = P_\Omega(M)$, whp.

A.2 Matrix Sensing

The following is an alternate characterization of RIP that we use heavily in our proofs. At a conceptual level, it says that if $\mathcal{A}$ satisfies RIP, then it also preserves inner-product between any two rank-k matrices (upto some additive error).

Lemma A.2.1. *Suppose $\mathcal{A}(\cdot)$ satisfies $2k$-RIP with constant δ_{2k}. Then, for any $U_1, U_2 \in \mathbb{R}^{m\times k}$ and $V_1, V_2 \in \mathbb{R}^{n\times k}$, we have the following:*

$$\left|\left\langle \mathcal{A}\left(U_1V_1^\dagger\right), \mathcal{A}\left(U_2V_2^\dagger\right)\right\rangle - \mathrm{Tr}\left(U_2^\dagger U_1 V_1^\dagger V_2\right)\right| \leq 3\delta_{2k}\left\|U_1V_1^\dagger\right\|_F\left\|U_2V_2^\dagger\right\|_F \quad (3)$$

Proof. Consider the matrices $X_1 \stackrel{\text{def}}{=} U_1V_1^T$, $X_2 \stackrel{\text{def}}{=} U_2V_2^T$ and $X = X_1 + X_2$. Since the rank of X is at most $2k$, we obtain the following using the RIP of $\mathcal{A}$:

$$(1-\delta)\left\|U_1V_1^T + U_2V_2^T\right\|_F^2 \leq \left\|\mathcal{A}(X)\right\|_2^2 \leq (1+\delta)\left\|U_1V_1^T + U_2V_2^T\right\|_F^2.$$

Concentrating on the second inequality, we obtain

$$\begin{aligned}
&\sum_i \left(\mathrm{Tr}\left(A_i U_1 V_1^T\right) + \mathrm{Tr}\left(A_i U_2 V_2^T\right)\right)^2 \\
&\qquad\qquad \leq (1+\delta)\left(\left\|U_1 V_1^T\right\|_F^2 + \left\|U_2 V_2^T\right\|_F^2 + \mathrm{Tr}\left(U_1 V_1^T V_2 U_2^T\right)\right) \\
\overset{(\zeta_1)}{\Rightarrow} &\sum_i \mathrm{Tr}\left(A_i U_1 V_1^T\right)\mathrm{Tr}\left(A_i U_2 V_2^T\right) - \mathrm{Tr}\left(U_1 V_1^T V_2 U_2^T\right) \\
&\qquad\qquad \leq \delta\left(\left\|U_1 V_1^T\right\|_F^2 + \left\|U_2 V_2^T\right\|_F^2 + \mathrm{Tr}\left(U_1 V_1^T V_2 U_2^T\right)\right) \\
\overset{(\zeta_2)}{\Rightarrow} &\sum_i \mathrm{Tr}\left(A_i U_1 V_1^T\right)\mathrm{Tr}\left(A_i U_2 V_2^T\right) - \mathrm{Tr}\left(U_1 V_1^T V_2 U_2^T\right) \\
&\qquad\qquad \leq \delta\left(\left\|U_1 V_1^T\right\|_F^2 + \left\|U_2 V_2^T\right\|_F^2 + \left\|U_1 V_1^T\right\|_F \left\|U_2 V_2^T\right\|_F\right)
\end{aligned} \tag{4}$$

where (ζ_1) follows from the fact that X_1 and X_2 are rank-k matrices and hence $\mathcal{A}(\cdot)$ satisfies RIP w.r.t. those matrices and (ζ_2) follows from the fact that $\mathrm{Tr}\left(U_1 V_1^T V_2 U_2^T\right) \leq \left\|U_1 V_1^T\right\|_F \left\|U_2 V_2^T\right\|_F$. Note that if we replace $U_1 V_1^T$ by $\lambda U_1 V_1^T$ and $U_2 V_2^T$ by $\frac{1}{\lambda} U_2 V_2^T$ in (4) for some non-zero $\lambda \in \mathbb{R}$, the LHS of (4) does not change where as the RHS of (4) changes. Optimizing the RHS w.r.t. λ, we obtain

$$\sum_i \mathrm{Tr}\left(A_i U_1 V_1^T\right)\mathrm{Tr}\left(A_i U_2 V_2^T\right) - \mathrm{Tr}\left(U_2^T U_1 V_1^T V_2\right) \leq 3\delta \left\|U_1 V_1^T\right\|_F \left\|U_2 V_2^T\right\|_F.$$

A similar argument proves the other side of the inequality. This proves the lemma. □

Proof of Lemma 2.4.4. We first show that the update (6) reduces to:

$$\sum_{q=1}^{k}\left(\sum_{i=1}^{s} A_i u_p^{(t)} u_q^{(t)\dagger} A_i^\dagger\right)\widehat{v}_q^{(t+1)} = \sum_{q=1}^{k}\left(\sum_{i=1}^{s} A_i u_p^{(t)} u_q^{*\dagger} A_i^\dagger\right) v_q^* \qquad \forall\, p \in [k]. \tag{5}$$

Let $Err(V) \overset{\text{def}}{=} \sum_i \left(\mathrm{Tr}\left(A_i M\right) - \mathrm{Tr}\left(A_i U^{(t)} V^\dagger\right)\right)^2$. Since $\widehat{V}^{(t+1)}$ minimizes $E(V)$,

we have $\nabla_V E(\widehat{V}^{(t+1)}) = 0$.

$$
\begin{aligned}
&\nabla_{v_p} Err(\widehat{V}^{(t+1)}) = 0 \\
\Rightarrow &\sum_{i=1}^{s} \left(\sum_{l=1}^{k} {v_q^{(t)}}^\dagger A_i u_q^{(t)} - \sum_{l=1}^{k} \sigma_q^* {v_q^*}^\dagger A_i u_q^* \right) A_i u_p = 0 \\
\Rightarrow &\sum_{l=1}^{k} \sum_{i=1}^{s} A_i u_p \left({v_q^{(t+1)}}^\dagger A_i u_q^{(t)} \right) = \sum_{l=1}^{k} \sum_{i=1}^{s} A_i u_p \left(\sigma_q^* {v_q^*}^\dagger A_i u_q^* \right) \\
\Rightarrow &\sum_{l=1}^{k} \sum_{i=1}^{s} A_i u_p \left({u_q^{(t)}}^\dagger A_i^\dagger v_q^{(t+1)} \right) = \sum_{l=1}^{k} \sum_{i=1}^{s} A_i u_p \left({u_q^*}^\dagger A_i^\dagger \sigma_q^* v_q^* \right) \\
\Rightarrow &\sum_{l=1}^{k} \left(\sum_{i=1}^{s} A_i u_p {u_q^{(t)}}^\dagger A_i^\dagger \right) v_q^{(t+1)} = \sum_{l=1}^{k} \left(\sum_{i=1}^{s} A_i u_p {u_q^*}^\dagger A_i^\dagger \right) \sigma_q^* v_q^*
\end{aligned}
$$

Define

$$
S = \begin{bmatrix} \sigma_1^* I_n & \dots & 0_n \\ \vdots & \vdots & \vdots \\ 0_n & \dots & \sigma_k^* I_n \end{bmatrix}, \quad v^* = \begin{bmatrix} v_1^* \\ \vdots \\ v_k^* \end{bmatrix}, \text{ and } \quad \widehat{v}_1^{(t+1)} = \begin{bmatrix} \widehat{v}_1^{(t+1)} \\ \vdots \\ \widehat{v}_k^{(t+1)} \end{bmatrix}.
$$

Then,

$$
\begin{aligned}
\widehat{v}_1^{(t+1)} &= B^{-1} C S v^* \\
&= DSv^* - B^{-1} \left(BD - C \right) S v^*
\end{aligned}
$$

where inverting B is valid since the minimum singular value of B is strictly positive (please refer Lemma A.2.2). Considering the p^{th} block of $\widehat{v}^{(t)}$, we obtain

$$
\begin{aligned}
\widehat{v}_p^{(t+1)} &= \left(\sum_q \langle u_p^{(t)}, u_q^* \rangle \sigma_q^* v_q^* \right) - \left(B^{-1} \left(BD - C \right) S v^* \right)_p \\
&= \left(\sum_q \sigma_q^* v_q^* {u_q^*}^\dagger \right) u_p^{(t)} - \left(B^{-1} \left(BD - C \right) S v^* \right)_p .
\end{aligned}
$$

This gives us the following equation for $\widehat{V}^{(t)}$:

$$
\widehat{V}^{(t+1)} = V^* \Sigma^* {U^*}^\dagger U^{(t)} - F, \text{ where}
$$

$$F := \begin{bmatrix} \left(B^{-1}\left(BD-C\right)Sv^*\right)_1 & \left(B^{-1}\left(BD-C\right)Sv^*\right)_2 & \cdots & \left(B^{-1}\left(BD-C\right)Sv^*\right)_k \end{bmatrix}.$$

Hence Proved. □

A.2.1 Rank-1 Matrix Sensing

Proof of Lemma 2.4.2. Using definition of the spectral norm:

$$\|B^{-1}\left(\langle u^*, u^t\rangle B - C\right)v^*\| \leq \|B^{-1}\|_2 \cdot \|\langle u^*, u^t\rangle B - C\|_2 \cdot \|v^*\|_2. \tag{6}$$

Consider $B = \sum_i A_i u^t (u^t)^\dagger A_i^\dagger$. Now, smallest eigenvalue of B, i.e., $\lambda_{min}(B)$ is given by:

$$\begin{aligned}
\lambda_{min}(B) = \min_{\|z\|=1} z^\dagger B z &= \min_{\|z\|=1} \sum_i z^\dagger A_i u^t (u^t)^\dagger A_i^\dagger z \\
&= \min_{\|z\|=1} \sum_i \mathrm{Tr}(A_i u^t z^\dagger)\,\mathrm{Tr}(A_i u^t z^\dagger), \\
&= \min_{\|z\|=1} \langle \mathcal{A}(u^t z^\dagger), \mathcal{A}(u^t z^\dagger)\rangle \\
&\geq 1 - 3\delta_2,
\end{aligned} \tag{7}$$

where the last inequality follows using Lemma A.2.1. Using (7),

$$\|B^{-1}\|_2 \leq \frac{1}{1-3\delta_2}. \tag{8}$$

Now, consider $G = \langle u^*, u^t\rangle B - C = \sum_i A_i \left(\langle u^*, u^t\rangle u^t (u^t)^\dagger - u^t (u^*)^\dagger\right) A_i^\dagger = \sum_i A_i u^t \left(\langle u^*, u^t\rangle u^t - u^*\right)^\dagger A_i^\dagger$. Using definition of the spectral norm:

$$\begin{aligned}
\|G\|_2 &= \max_{\|z\|=1,\|y\|=1} z^\dagger G y, \\
&= \max_{\|z\|=1,\|y\|=1} \sum_i z^\dagger A_i u^t \left(\langle u^*, u^t\rangle u^t - u^*\right)^\dagger A_i^\dagger y, \\
&= \max_{\|z\|=1,\|y\|=1} \langle \mathcal{A}(u^t z^\dagger), \mathcal{A}\left(\left(\langle u^*, u^t\rangle u^t - u^*\right) y^\dagger\right)\rangle, \\
&\leq 3\delta_2 \sqrt{1 - \langle u^t, u^*\rangle^2},
\end{aligned} \tag{9}$$

where the last inequality follows by using Lemma A.2.1 and the fact that $\langle u^t, (\langle u^*, u^t\rangle u^t - u^*)\rangle = 0$.

Lemma now follows using (6), (8), (9). □

A.2.2 Rank-k Matrix Sensing

Proof of Lemma 2.4.3. Since $\widehat{U}^t$ and $\widetilde{U}^t$ have full rank and span the same subspace, there exists a $k \times k$, full rank matrix R such that $\widehat{U}^t = \widetilde{U}^t R = U^t R_U^t R$. We have:

$$\left\|\mathcal{A}\left(\widehat{U}^t V^\dagger\right) - b\right\|_2 = \left\|\mathcal{A}\left(U^t\left(V\left(R_U^t R\right)^\dagger\right)^\dagger\right) - b\right\|_2 \geq \left\|\mathcal{A}\left(U^t\left(\widetilde{V}^{t+1}\right)^\dagger\right) - b\right\|_2$$

with equality holding in the last step for $V = \widetilde{V}^{t+1}\left(\left(R_U^t R\right)^\dagger\right)^{-1}$. The proof of Theorem 2.4.1 shows that $\widetilde{V}^{t+1}$ is unique and has full rank (since dist $\left(\widetilde{V}^{t+1}, V^*\right)$ < 1). This means that $\widehat{V}^{t+1}$ is also unique and is equal to $\widetilde{V}^{t+1}\left(\left(R_U^t R\right)^\dagger\right)^{-1}$. This shows that $Span\left(\widehat{V}^{t+1}\right) = Span\left(\widetilde{V}^{t+1}\right)$ and that both $\widehat{V}^{t+1}$ and $\widetilde{V}^{t+1}$ have full rank. □

Lemma A.2.2. *Let linear measurement $\mathcal{A}$ satisfy RIP for all $2k$-rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m\times n}$ being a rank-k matrix. Let δ_{2k} be the RIP constant for rank $2k$-matrices. Then, we have the following bound on the minimum singular value of B:*

$$\sigma_{min}(B) \geq 1 - \delta_{2k}. \tag{10}$$

Proof. Select any $w \in \mathbb{R}^{nk}$ such that $\|w\|_2 = 1$. Let

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}$$

where each $w_p \in \mathbb{R}^n$. Also denote $W \overset{\text{def}}{=} [w_1 w_2 \cdots w_k] \in \mathbb{R}^{n\times k}$, i.e., $w = vec(W)$.

We have,

$$\begin{aligned} w^\dagger B w &= \sum_{p,q=1}^{k} w_p^\dagger B_{pq} w_q = \sum_{p,q=1}^{k} w_p^\dagger \left(\sum_{i=1}^{d} A_i u_p^t (u_q^t)^\dagger A_i^\dagger \right) w_q \\ &= \sum_{i=1}^{d} \sum_{p,q=1}^{k} w_p^\dagger A_i u_p^t (u_q^t)^\dagger A_i^\dagger w_q \\ &= \sum_{i=1}^{d} \left(\sum_{p=1}^{k} w_p^\dagger A_i u_p^t \right) \left(\sum_{q=1}^{k} w_q^\dagger A_i u_q^t \right) = \sum_{i=1}^{d} \mathrm{Tr}\left(A_i U^t W^\dagger \right)^2 . \end{aligned}$$

Now, using RIP (see Definition 2.3.1) along with the above equation, we get:

$$\begin{aligned} w^\dagger B w = \sum_{i=1}^{d} \mathrm{Tr}\left(A_i U^t W^\dagger \right)^2 \geq (1-\delta) \left\| U^t W^\dagger \right\|_{\mathrm{F}}^2 &= (1-\delta_{2k}) \left\| W \right\|_F^2 \\ &= (1-\delta_{2k}) \|w\|^2 = (1-\delta_{2k}). \end{aligned}$$

Since w was arbitrary, this proves the lemma. □

Proof of Lemma 2.4.5. Note that,

$$\begin{aligned} \left\| F(\Sigma^*)^{-1} \right\|_2 \leq \left\| F(\Sigma^*)^{-1} \right\|_{\mathrm{F}} &= \left\| B^{-1} \left(BD - C \right) v^* \right\|_2 \\ &\leq \left\| B^{-1} \right\|_2 \left\| (BD - C) \right\|_2 \left\| v^* \right\|_2 \\ &\leq \frac{\sqrt{k}}{1-\delta_{2k}} \left\| (BD - C) \right\|_2 \end{aligned} \tag{11}$$

where the last step follows from Lemma A.2.2. Now we need to bound $\left\| (BD - C) \right\|_2$. Choose any $w, z \in \mathbb{R}^{nk}$ such that $\left\| w \right\|_2 = \left\| z \right\|_2 = 1$. As in Lemma A.2.2, define the following components of w and z:

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \text{ and } z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

where each $w_p, z_p \in \mathbb{R}^n$ and $W \stackrel{\text{def}}{=} [w_1 w_2 \cdots w_k]$ and $Z \stackrel{\text{def}}{=} [z_1 z_2 \cdots z_k] \in \mathbb{R}^{n \times k}$. We have,

$$w^\dagger \left(BD - C\right) z = \sum_{p,q=1}^{k} w_p^\dagger \left(BD - C\right)_{pq} z_q$$

We calculate $\left(BD - C\right)_{pq}$ as follows:

$$\begin{aligned}
\left(BD - C\right)_{pq} &= \sum_{l=1}^{k} B_{pl} D_{lq} - C_{pq} \\
&= \left(\sum_{l=1}^{k} B_{pl} \langle u_l^t, u_q^* \rangle \mathbb{I}_{n \times n} \right) - C_{pq} \\
&= \left(\sum_{l=1}^{k} u_q^{*\dagger} u_l^t \sum_{i=1}^{d} A_i u_p^t (u_l^t)^\dagger A_i^\dagger \right) - C_{pq} \\
&= \sum_{i=1}^{d} A_i u_p^t u_q^{*\dagger} \sum_{l=1}^{k} u_l^t (u_l^t)^\dagger A_i^\dagger - \sum_{i=1}^{d} A_i u_p^t (u_q^*)^\dagger A_i^\dagger \\
&= \sum_{i=1}^{d} A_i u_p^t u_q^{*\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger .
\end{aligned}$$

So we have,

$$
\begin{aligned}
& w^{\dagger}\left(BD-C\right)z \\
&= \sum_{p,q=1}^{k} w_p^{\dagger} \sum_{i=1}^{d} A_i u_p^t u_q^{*\dagger} \left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) A_i^{\dagger} z_q \\
&= \sum_{i=1}^{d} \sum_{p,q=1}^{k} w_p^{\dagger} A_i u_p^t u_q^{*\dagger} \left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) A_i^{\dagger} z_q \\
&= \sum_{i=1}^{d} \mathrm{Tr}\left(A_i U^t W^{\dagger}\right) \mathrm{Tr}\left(A_i \left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^* Z^{\dagger}\right) \\
&\overset{(\zeta_1)}{\leq} \mathrm{Tr}\left(U^{*\dagger}\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^t W^{\dagger} Z\right) + \delta_{2k} \left\|U^t W^{\dagger}\right\|_{\mathrm{F}} \left\|\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^* Z^{\dagger}\right\|_{\mathrm{F}} \\
&\overset{(\zeta_2)}{\leq} \delta_{2k} \|W\|_F \sqrt{\left\|(U^*)^{\dagger}\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right)^2 U^*\right\|_{\mathrm{F}} \|Z^{\dagger}Z\|_{\mathrm{F}}} \\
&\overset{(\zeta_3)}{\leq} \delta_{2k}\sqrt{k} \cdot dist(U^t, U^*),
\end{aligned}
$$

where (ζ_1) follows from the fact that $\mathcal{A}$ satisfies $2k$-RIP and Lemma A.2.1, (ζ_2) follows from the fact that $\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^t = 0$, (ζ_3) follows from the following: $\|W\|_F = \|w\|_2 = 1$, $\|Z^{\dagger}Z\|_F \leq \|Z\|_F^2 = 1$ and and finally : $\left\|\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^*\right\|_{\mathrm{F}} \leq \sqrt{k} \left\|\left(U^t(U^t)^{\dagger} - \mathbb{I}_{n\times n}\right) U^*\right\|_2$.

Since w and z were arbitrary unit vectors, we can conclude that $\|BD - C\|_2$ $\leq \delta_{2k}\sqrt{k} \cdot dist(U^t, U^*)$. Plugging this bound in (11) proves the lemma. □

Proof of Lemma 2.4.6. Note that $\|\Sigma^*(R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*}{\sigma_{\min}(R^{(t+1)})}$. Now,

$$\begin{aligned}
\sigma_{\min}(R^{(t+1)}) &= \min_{z,\|z\|_2=1} \|R^{(t+1)}z\|_2 = \min_{z,\|z\|_2=1} \|V^{(t+1)}R^{(t+1)}z\|_2, \\
&= \min_{z,\|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z - Fz\|_2, \\
&\geq \min_{z,\|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z\|_2 - \|Fz\|_2, \\
&\geq \min_{z,\|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z\|_2 - \|F\|_2, \\
&\geq \sigma_k^*\sigma_{\min}(U^{*\dagger}U^{(t)}) - \|F\|_2, \\
&\geq \sigma_k^*\sqrt{1 - \left\|U_\perp^{*\dagger}U^{(t)}\right\|_2^2} - \sigma_1^*\|F(\Sigma^*)^{-1}\|_2, \\
&= \sigma_k^*\sqrt{1 - \mathrm{dist}(U^*, U^{(t)})^2} - \sigma_1^*\|F(\Sigma^*)^{-1}\|_2. \qquad (12)
\end{aligned}$$

Lemma now follows using above inequality with Lemma 2.4.5.

□

A.2.3 Noisy Matrix Sensing

We now consider an extension of the matrix sensing problem where measurements can be corrupted arbitrarily using a bounded noise. That is, we observe $b = \mathcal{A}(M + N)$, where N is the noise matrix. For this noisy case as well, we show that a.s.recovers M upto an additive approximation depending on the Frobenius norm of N.

Theorem A.2.3. *Let M and $\mathcal{A}(\cdot)$ be as defined in Theorem 2.3.1. Suppose, a.s.algorithm (Algorithm 2) is supplied inputs $\mathcal{A}$, $b = \mathcal{A}(M + N)$, where N is the noise matrix s.t. $\|N\|_F < \frac{1}{100}\sigma_k^*$. Then, after $T = 4\log(2/\epsilon)$ steps, iterates $\widehat{U}^T$, $\widehat{V}^T$ of a.s.satisfy:*

$$\mathrm{dist}\left(\widehat{V}^T, V^*\right) \leq \frac{10\|N\|_F}{\sigma_k^*} + \epsilon, \ \mathrm{dist}\left(\widehat{U}^T, U^*\right) \leq \frac{10\|N\|_F}{\sigma_k^*} + \epsilon.$$

See Definition 2.4.1 for definition of $\mathrm{dist}\,(U, W)$.

Proof. At a high level, our proof for noisy case follows closely, the exact case proof given in Section 2.4. That is, we show that the update of a.s.algorithm

is similar to power-method type update but with two errors terms: one due to incomplete measurements and another due to the noise matrix.

Similar to our proof for sensing problem (Section 2.4), we analyze QR-decomposition based updates. That is,

$$\begin{aligned}
\widehat{U}^t &= U^t R_U^t \quad \text{(QR decomposition)},\\
\widehat{V}^{t+1} &= \arg\min_{V} \ \|\mathcal{A}(U^t V^\dagger) - b\|_2^2,\\
\widehat{V}^{t+1} &= V^{t+1} R_V^{t+1}. \quad \text{(QR decomposition)}
\end{aligned}$$

Similar to Lemma 2.4.4, we can re-write the above given update equation as:

$$\begin{aligned}
\widehat{V}^{t+1} &= V^* \Sigma^* (U^*)^\dagger U^t - F + V^N \Sigma^N (U^N)^\dagger U^{(t)} - G,\\
V^{t+1} &= \widehat{V}^{t+1} (R^{(t+1)})^{-1},
\end{aligned} \tag{13}$$

where, F is the error matrix and is as defined in (8) and G is the error matrix due to noise and is given by:

$$G \stackrel{\text{def}}{=} \left[\left(B^{-1}\left(BD^N - C^N\right) S^N v^N\right)_1 \ \cdots \ \left(B^{-1}\left(BD^N - C^N\right) S^N v^N\right)_k \right], \tag{14}$$

where B, C and D defined in the previous section (See (7)) and C^N and D^N are defined below:

$$C^N \stackrel{\text{def}}{=} \begin{bmatrix} C_{11}^N & \cdots & C_{1m}^N \\ \vdots & \ddots & \vdots \\ C_{k1}^N & \cdots & C_{km}^N \end{bmatrix}, \ D^N \stackrel{\text{def}}{=} \begin{bmatrix} D_{11}^N & \cdots & D_{1m}^N \\ \vdots & \ddots & \vdots \\ D_{k1}^N & \cdots & D_{km}^N \end{bmatrix},$$

with $C_{pq}^N \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^{(t)} (u_q^N)^\dagger A_i^\dagger$ and $D_{pq}^N \stackrel{\text{def}}{=} \langle u_p^{(t)}, u_q^N \rangle \mathbb{I}_{n\times n}$. Also,

$$S^N = \begin{bmatrix} \sigma_1^N I_n & \dots & 0_n \\ \vdots & \vdots & \vdots \\ 0_n & \dots & \sigma_N^N I_n \end{bmatrix}, \quad v^N = \begin{bmatrix} v_1^N \\ \vdots \\ v_k^N \end{bmatrix}.$$

Now, multiplying (13) with $V_\perp^*$, we get:

$$V_\perp^{*\dagger} V^{t+1} = (V_\perp^{*\dagger} V^N \Sigma^N U^{N\dagger} U^{(t)} - V_\perp^{*\dagger} F - V_\perp^{*\dagger} G) R^{(t+1)^{-1}}.$$

That is,

$$\begin{aligned}
&\text{dist}(V^*, V^{t+1}) = \|V_\perp^{*\dagger} V^{t+1}\|_2 \\
&\leq (\|V^N \Sigma^N (U^N)^\dagger U^{(t)}\|_2 + \|F|_2 + \|G\|_2) \|(R^{(t+1)})^{-1}\|_2, \\
&\leq \left(\sigma_1^N + \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^*\|_2 + \|G\|_2\right) \|(R^{(t+1)})^{-1}\|_2, \\
&\leq \left(\sigma_1^N + \frac{\sigma_1^* \delta_{2k} k}{1-\delta_{2k}} \text{dist}(U^t, U^*) + \|G\|_2\right) \|(R^{(t+1)})^{-1}\|_2, \qquad (15)
\end{aligned}$$

where the last inequality follows using Lemma 2.4.5.

Now, we break down the proof in the following two steps:

- Bound $\|G\|_2$ (Lemma A.2.4, analogous to Lemma 2.4.5)
- Bound $\|(R^{(t+1)})^{-1}\|_2$ (Lemma A.2.5, similar to Lemma 2.4.6)

Later in this section, we provide the above mentioned lemmas and their detailed proof.

Now, by assumption, $\sigma_1^N \leq \|N\|_{\text{F}} \leq \sigma_k^*$. Also, as $\delta_{2k} \leq 1/2$, $\frac{1}{1-\delta_{2k}} \leq 2$. Finally, assume $\text{dist}(V^*, V^{t+1}) \geq \max(10 \cdot \frac{\sigma_1^N}{\sigma_k^*}, 10\frac{\|N\|_F}{\sigma_1^*})$. Using these observations and lemmas A.2.4, A.2.5 along with (15), we get:

$$\text{dist}(V^*, V^{t+1}) \leq \frac{0.5\text{dist}(U^*, U^t)}{\sqrt{1 - \text{dist}(U^t, U^*)^2} - 0.5\text{dist}(U^*, U^t)}. \qquad (16)$$

As, U^0 is obtained using SVD of $\sum_i A_i b_i$. Hence, using Lemma A.1.1, we have:

$$\begin{aligned}
&\|\mathcal{A}(U^0\Sigma^0 V^0 - U^*\Sigma^*(V^*)^\dagger)\|_2^2 \leq 0.5\|\mathcal{A}(N)\|_2^2 + 4\delta_{2k}\|\mathcal{A}(U^*\Sigma^*(V^*)^\dagger)\|_2^2, \\
&\Rightarrow \|U^0\Sigma^0 V^0 - U^*\Sigma^*(V^*)^\dagger\|_F^2 \leq \|N\|_F^2 + 4\delta_{2k}(1+\delta_{2k})\|\Sigma^*\|_F^2, \\
&\Rightarrow (\sigma_k^*)^2 \|(U^0(U^0)^\dagger - I)U^*\|_F^2 \leq \|N\|_F^2 + 4\delta_{2k}(1+\delta_{2k})k(\sigma_1^*)^2, \\
&\Rightarrow \text{dist}(U^0, U^*) \leq \|(U^0(U^0)^\dagger - I)U^*\|_F^2 \leq \frac{\|N\|_F^2}{(\sigma_k^*)^2} + 6\delta_{2k}k\left(\frac{\sigma_1^*}{\sigma_k^*}\right)^2 < \frac{1}{2},
\end{aligned}$$

where last inequality follows using $\frac{\|N\|_F}{\sigma_k^*} < 1/100$.

Theorem now follows using above equation with (16). □

Lemma A.2.4. *Let linear measurement $\mathcal{A}$ satisfy RIP for all $2k$-rank matrices and let $b = \mathcal{A}(M+N)$ with $M \in \mathbb{R}^{m\times n}$ being a rank-k matrix and let $N = U^N\Sigma^N(V^N)^\dagger$. Let δ_{2k} be the RIP constant for rank $2k$-matrices. Then, we have the following bound on $\|G\|_2$:*

$$\|G\|_2 \leq \frac{\delta_{2k}\|N\|_F}{1-\delta_{2k}}. \tag{17}$$

Proof. Note that,

$$\begin{aligned}\|G\|_2 \leq \|G\|_F &= \|B^{-1}(BD^N - C^N)S^N v^N\|_2 \\ &\leq \|B^{-1}\|_2\|(BD^N - C^N)S^N\|_2\|S^N v^N\|_2 \\ &\leq \frac{\sqrt{k}}{1-\delta_{2k}}\|(BD^N - C^N)S^N\|_2, \end{aligned} \tag{18}$$

where the last inequality follows using Lemma A.2.2 and the fact that $\|V^N\|_F = \sqrt{k}$. Now let $w = [w_1^\dagger\ w_2^\dagger\ \dots\ w_k^\dagger]^\dagger \in \mathbb{R}^{nk}$ and $z = [z_1^\dagger\ z_2^\dagger\ \dots\ z_n^\dagger]^\dagger \in \mathbb{R}^{n^2}$ be any two arbitrary vectors such that $\|w\|_2 = \|z\|_2 = 1$. Then,

$$\begin{aligned}& w^\dagger\left(BD^N - C^N\right)S^N z \\ &= \sum_{p=1}^{k}\sum_{q=1}^{n} w_p^\dagger \sum_{i=1}^{d} A_i u_p^t {u_q^N}^\dagger \left(U^t(U^t)^\dagger - \mathbb{I}_{n\times n}\right)A_i^\dagger \sigma_q^N z_q \\ &= \sum_{i=1}^{d}\sum_{p=1}^{k}\sum_{q=1}^{n} w_p^\dagger A_i u_p^t {u_q^N}^\dagger \left(U^t(U^t)^\dagger - \mathbb{I}_{n\times n}\right)A_i^\dagger \sigma_q^N z_q \\ &= \sum_{i=1}^{d}\left(\sum_{p=1}^{k} w_p^\dagger A_i u_p^t\right)\left(\sum_{q=1}^{n}\sigma_q^N z_q^\dagger A_i\left(U^t(U^t)^\dagger - \mathbb{I}_{n\times n}\right)u_q^N\right) \\ &= \sum_{q=1}^{n}\left(\sum_{i=1}^{d}\mathrm{Tr}\left(A_i U W^\dagger\right)\mathrm{Tr}\left(A_i\left(U^t(U^t)^\dagger - \mathbb{I}_{n\times n}\right)u_q^N\sigma_q^N z_q^\dagger\right)\right).\end{aligned}$$

Now, using RIP, we get:

$$
\begin{aligned}
& w^{\dagger}\left(BD^{N}-C^{N}\right)z \\
& \leq \sum_{q=1}^{n} u_{q}^{N\dagger}\left(U^{t}U^{t\dagger}-\mathbb{I}_{n\times n}\right)U^{t}W^{\dagger}\sigma_{q}^{N}z_{q} \\
& \qquad\qquad + \delta_{2k}\left\|U^{t}W^{\dagger}\right\|_{\mathrm{F}}\left\|\left(U^{t}U^{t\dagger}-\mathbb{I}_{n\times n}\right)u_{q}^{N}\sigma_{q}^{N}z_{q}{}^{\dagger}\right\|_{\mathrm{F}} \\
& \leq \sum_{q=1}^{n}\delta_{2k}\|W^{\dagger}\|_{F}\|(U^{t}(U^{t})^{\dagger}-\mathbb{I}_{n\times n})u_{q}^{N}\|_{2}\|\sigma_{q}^{N}z_{q}\|_{2}, \\
& \leq \delta_{2k}\ \sum_{q=1}^{n}\left\|u_{q}^{N}\right\|_{2}\left\|\sigma_{q}^{N}z_{q}\right\|_{2}=\delta_{2k}\sum_{q=1}^{n}\sigma_{q}^{N}\left\|z_{q}\right\|_{2}, \\
& \leq \delta_{2k}\sqrt{\sum_{q=1}^{n}\left\|z_{q}\right\|_{2}^{2}}\sqrt{\sum_{q=1}^{n}(\sigma_{q}^{N})^{2}}\leq\delta_{2k}\left\|N\right\|_{\mathrm{F}}.
\end{aligned}
$$

This finishes the proof. □

Lemma A.2.5. *Assuming conditions of Lemma A.2.4, we have the following bound on the minimum singular value of $R^{(t)}$:*

$$\sigma_{min}\left(R^{(t+1)}\right)\geq\sigma_{k}^{*}\sqrt{1-\mathrm{dist}(U^{t},U^{*})^{2}}-\sigma_{1}^{N}-\left\|F\right\|_{2}-\left\|G\right\|_{2}.$$

Proof. Similar to the proof of Lemma 2.4.6, we have the following set of in-

equalities:

$$
\begin{aligned}
&\sigma_{\min}\left(R^{(t+1)}\right) \\
&= \min_{\|z\|_2=1} \left\|R^{(t+1)}z\right\|_2 \\
&= \min_{\|z\|_2=1} \left\|V^{(t)}R^{(t+1)}z\right\|_2 \\
&= \min_{\|z\|_2=1} \left\|V^*\Sigma^*U^{*\dagger}U^tz + V^N\Sigma^N(U^N)^\dagger U^{(t)}z - Fz - Gz\right\|_2 \\
&\geq \min_{\|z\|_2=1} \left\|V^*\Sigma^*U^{*\dagger}U^tz\right\|_2 - \left\|V^N\Sigma^N(U^N)^\dagger\right\|_2 - \|F\|_2 - \|G\|_2 \\
&\geq \sigma_k^* \min_{\|z\|_2=1} \left\|U^{*\dagger}U^tz\right\|_2 - \sigma_1^N - \|F\|_2 - \|G\|_2 \\
&\geq \sigma_k^*\sqrt{1 - \left\|U_\perp^{*\dagger}U\right\|_2^2} - \sigma_1^N - \|F\|_2 - \|G\|_2 \\
&= \sigma_k^*\sqrt{1 - \mathrm{dist}(U^t, U^*)^2} - \sigma_1^N - \|F\|_2 - \|G\|_2 .
\end{aligned}
$$

This proves the lemma. □

A.2.4 Stagewise Alternating Minimization for Matrix Sensing

Proof of Lemma 2.6.1. As the initial point of the i-th stage is obtained by one step of SVP [44], using Lemma A.1.1, we obtain:

$$\left\|M - \widehat{U}^0_{1:i}(\widehat{V}^0_{1:i})^\dagger\right\|_F^2 \leq \sum_{j=i+1}^{k}(\sigma_j^*)^2 + 2\delta_{2k}\|M - \widehat{U}^T_{1:i-1}V^T_{1:i-1}\|_F^2.$$

Now, by assumption over the $(i-1)$-th stage error (this assumption follows from the inductive hypothesis in proof of Theorem 2.3.2),

$$\left\|M - \widehat{U}^0_{1:i}(\widehat{V}^0_{1:i})^\dagger\right\|_F^2 \leq \sum_{j=i+1}^{k}(\sigma_j^*)^2 + 2\delta_{2k}16k(\sigma_i^*)^2.$$

Lemma now follows by setting $\delta_{2k} \leq \frac{1}{3200k}$. □

Proof of Lemma 2.6.2. For our proof, we consider two cases: a) $\frac{\sigma_i^*}{\sigma_{i+1}^*} < 5\sqrt{k}$, b) $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$.

Case (a): In this case, using monotonicity of the AltMin algorithm directly gives error bound. That is,

$$\begin{aligned}\|M - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\|_F^2 &\leq \|M - \widehat{U}_{1:i}^0 V_{1:i}^0\|_F^2 \\ &\leq k(\sigma_{i+1}^*)^2 + \frac{25k}{100}(\sigma_{i+1}^*)^2.\end{aligned}$$

Case (b): At a high level, if $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$ then $U_{1:i}^0$ is "close" to $U_{1:i}^*$ and hence the error bound follows by using an analysis similar to the noisy case. Note that σ_{i+1}^* being small implies that the "noise" is small. See Lemma A.2.6 for a formal proof of this case. □

Lemma A.2.6. *Assume conditions given in Theorem 2.3.2 are satisfied and let* $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$. *Also, let*

$$\left\|M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger\right\|_F^2 \leq \sum_{j=i+1}^{k}(\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2.$$

Then, $U_{1:i}^T$, $V_{1:i}^T$ *satisfy:*

$$\|M - \widehat{U}_{1:i}^T V_{1:i}^T\|_F^2 \leq \max(\epsilon, 16k(\sigma_{i+1}^*)^2),$$

Proof. We first show that if σ_i and σ_{i+1} have large gap then $\forall\, t$, the t^{th} iterate of the i-th stage, $\widehat{U}_{1:i}^t$ is close to $U_{1:i}^*$. Let $U_\perp^t$ be a basis of the subspace orthogonal to $\widehat{U}_{1:i}^t$.

$$\begin{aligned}&\|\left(U_\perp^t\right)^\dagger(M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger)\|_2 \\ &= \|\left(U_\perp^t\right)^\dagger M\|_2 \\ &\geq \|\left(U_\perp^t\right)^\dagger U_{1:i}^*\Sigma_{1:i}^*(V_{1:i}^*)^\dagger\|_2 - \|\left(U_\perp^t\right)^\dagger U_{i+1:k}^*\Sigma_{i+1:k}^*(V_{i+1:k}^*)^\dagger\|_2, \\ &\geq \sigma_i^*\|\left(U_\perp^t\right)^\dagger U_{1:i}^*\|_2 - \sigma_{i+1}^* \geq \sigma_i^*(\|\left(U_\perp^t\right)^\dagger U_{1:i}^*\|_2 - \frac{1}{5\sqrt{k}}).\end{aligned} \tag{19}$$

We also have:

$$\begin{aligned}
\| \left(U_\perp^t\right)^\dagger (M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger)\|_2^2 &\leq \|M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger\|_F^2 \\
&\leq \frac{1}{1-\delta_{2k}} \left\|\mathcal{A}\left(M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger\right)\right\|_2^2 \\
&\overset{(\zeta_1)}{\leq} \frac{1}{1-\delta_{2k}} \left\|\mathcal{A}\left(M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger\right)\right\|_2^2 \\
&\leq \frac{1+\delta_{2k}}{1-\delta_{2k}} \|M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger\|_F^2 \\
&\leq \frac{1+\delta_{2k}}{1-\delta_{2k}} \left(\sum_{j=i+1}^{k} (\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2\right) \\
&\leq \frac{1+\delta_{2k}}{1-\delta_{2k}} \left(k(\sigma_{i+1}^*)^2 + \frac{1}{100}(\sigma_i^*)^2\right), \qquad (20)
\end{aligned}$$

where (ζ_1) follows from the fact that lines $5-8$ of Algorithm 3 never increases $\left\|\mathcal{A}\left(M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger\right)\right\|_2$. Using (19), (20), and $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$, we obtain the following bound:

$$\| \left(U_\perp^t\right)^\dagger U_{1:i}^*\|_2 \leq \frac{1}{2} \; \forall \; t. \qquad (21)$$

Now, we consider the update equation for $\widehat{V}^{t+1}$:

$$\widehat{V}^{t+1} = \arg\min_{\widehat{V}} \|\mathcal{A}(\widehat{U}_{1:i}^t \widehat{V} - U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger)\|_2^2.$$

Note that, the update is same as noisy case with noise matrix $N = U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger$ from (13):

$$\widehat{V}^{t+1} = V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t - F + V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t - G, \qquad (22)$$

where F and G are given by (8), (14). Multiplying (22) from the left by

$V_\perp^\dagger = I - V^{t+1}(V^{t+1})^\dagger$, we obtain:

$$\begin{aligned}
&0 = V_\perp^\dagger \widehat{V}_{1:i}^{t+1} = V_\perp^\dagger \left(V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t - F + V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t - G \right) \\
\Rightarrow &V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t = V_\perp^\dagger \left(F - V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t + G \right) \\
\Rightarrow &\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t \right\|_{\mathrm{F}} \leq \|F\|_{\mathrm{F}} + \left\| V_\perp^\dagger V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t \right\|_{\mathrm{F}} + \|G\|_{\mathrm{F}} \\
\Rightarrow &\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} \\
&\quad \leq \frac{1}{\sigma_{\min}\left((U_{1:i}^*)^\dagger U_{1:i}^t \right)} \left(\|F\|_{\mathrm{F}} + \left\| V_\perp^\dagger V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t \right\|_{\mathrm{F}} + \|G\|_{\mathrm{F}} \right),
\end{aligned} \tag{23}$$

where the last inequality follows using the fact that $\sigma_{\min}(A)\|B\|_F \leq \|AB\|_F$. Using Lemma A.2.4, and a modification of Lemma 2.4.5, we get:

$$\|F\|_{\mathrm{F}} \leq \delta_{2k} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}}, \quad \|G\|_{\mathrm{F}} \leq \delta_{2k} \left\| U_\perp^\dagger U_{i+1:k}^* \Sigma_{i+1:k}^* \right\|_{\mathrm{F}} \leq \delta_{2k} \sqrt{k} \sigma_{i+1}. \tag{24}$$

Using (23), (24), and the fact that $\sigma_{\min}(U_\perp^\dagger U_{1:i}^*) = \sqrt{1 - \|U_\perp^\dagger U_{1:i}^*\|_2^2}$,

$$\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} \leq \frac{2}{\sqrt{3}} \left(\delta_{2k} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} + \sqrt{\sum_{j=i+1}^{k} (\sigma_j^*)^2} + \delta_{2k} \sqrt{k} \sigma_{i+1} \right).$$

Assuming $\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} > 2\sqrt{\sum_{j=i+1}^{k} \sigma_j^2}$, we obtain:

$$\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} \leq \frac{2}{3} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}}. \tag{25}$$

Using similar analysis, we can show that,

$$\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}} \leq \frac{2}{3} \left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}}.$$

So after $T \geq 8 \log(k\sigma_i^*)$ iterations, we have:

$$\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_{\mathrm{F}}^2 \leq 4 \sum_{j=i+1}^{k} (\sigma_j^*)^2.$$

Using the above inequality, we now bound the error after $T \geq 8\log(k\sigma_i^*)$ iterations of the i-th stage:

$$\left\|M - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_F \leq \left\|U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_F + \left\|U_{i+1:k}^*\Sigma_{i+1:k}^*\left(V_{i+1:k}^*\right)^\dagger\right\|_{\mathrm{F}}. \tag{26}$$

For the first term, we have:

$$\begin{aligned}
&\left\|U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_F^2 \\
&= \left\|U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger - U_{1:i}^T\left(U_{1:i}^T\right)^\dagger U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger\right. \\
&\quad \left. + U_{1:i}^T\left(U_{1:i}^T\right)^\dagger U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_F^2 \\
&= \left\|\left(I - U_{1:i}^T\left(U_{1:i}^T\right)^\dagger\right)U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger\right\|_{\mathrm{F}}^2 \\
&\quad + \left\|U_{1:i}^T\left(U_{1:i}^T\right)^\dagger U_{1:i}^*\Sigma_{1:i}^*\left(V_{1:i}^*\right)^\dagger - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_{\mathrm{F}}^2 \\
&\overset{(\zeta_1)}{\leq} \left\|U_\perp^\dagger U_{1:i}^*\Sigma_{1:i}^*\right\|_{\mathrm{F}}^2 + \left\|U_{1:i}^T\left(F + G - \left(U_{1:i}^t\right)^\dagger U_{i+1:k}^*\Sigma_{i+1:k}^*(V_{i+1:k}^*)^\dagger\right)\right\|_{\mathrm{F}}^2 \\
&= \left\|U_\perp^\dagger U_{1:i}^*\Sigma_{1:i}^*\right\|_{\mathrm{F}}^2 + \left\|F + G - \left(U_{1:i}^t\right)^\dagger U_{i+1:k}^*\Sigma_{i+1:k}^*(V_{i+1:k}^*)^\dagger\right\|_{\mathrm{F}}^2 \\
&= \left\|U_\perp^\dagger U_{1:i}^*\Sigma_{1:i}^*\right\|_{\mathrm{F}}^2 + 3\left\|F\right\|_{\mathrm{F}}^2 + 3\left\|G\right\|_{\mathrm{F}}^2 + 3\left\|U_{i+1:k}^*\Sigma_{i+1:k}^*(V_{i+1:k}^*)^\dagger\right\|_{\mathrm{F}}^2 \\
&\overset{(\zeta_2)}{\leq} (1+3\delta_{2k}^2)\left\|U_\perp^\dagger U_{1:i}^*\Sigma_{1:i}^*\right\|_{\mathrm{F}}^2 + 3(1+\delta_{2k}^2)\left\|U_{i+1:k}^*\Sigma_{i+1:k}^*(V_{i+1:k}^*)^\dagger\right\|_{\mathrm{F}}^2 \\
&\leq 8k(\sigma_{i+1}^*)^2,
\end{aligned} \tag{27}$$

where (ζ_1) follows from (22) and (ζ_2) follows from (24). Using (26) and (27), we obtain the following bound:

$$\left\|M - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\right\|_F \leq 4\sqrt{k}\sigma_{i+1}^*. \tag{28}$$

Hence Proved. □

A.3 Matrix Completion

Proof Of Theorem 2.3.3. Using Theorem 2.5.1, after $O(\log(1/\epsilon))$ iterations, we get:

$$\mathrm{dist}(U^t, U^*) \leq \epsilon,\ \mathrm{dist}(V^{t+1}, V^*) \leq \epsilon.$$

Now, using (19), the residual after t-th step is given by:

$$M - U^t(\widehat{V}^{t+1})^\dagger = (I - U^t(U^t)^\dagger)M - U^t F^\dagger.$$

That is,

$$\begin{aligned}\|M - U^t(\widehat{V}^{t+1})^\dagger\|_F &\leq \|(I - U^t(U^t)^\dagger)M\|_F + \|F\|_F \\ &\leq \sqrt{k}\|(I - U^t(U^t)^\dagger)U^*\Sigma^*\|_2 + \|F\|_F \\ &\leq \sqrt{k}\sigma_1^* \text{dist}(\widehat{U}^t, U^*).\end{aligned}$$

Now, using the fact that $\text{dist}(U^t, U^*) \leq \epsilon$ and the above equation, we get:

$$\|M - U^t(\widehat{V}^{t+1})^\dagger\|_F \leq \sqrt{k}\sigma_1^*\epsilon + \|F\|_F \overset{\zeta_1}{\leq} \sqrt{k}\sigma_1^*\epsilon + \sigma_1^*\sqrt{k}\epsilon \leq 2\sigma_1^*\sqrt{k}\epsilon,$$

where ζ_1 follows by Lemma 2.5.6 and setting δ_{2k} appropriately. Theorem 2.3.3 now follows by setting $\epsilon' = 2\sqrt{k}\|M\|_F\epsilon$. □

A.3.1 Initialization

Proof Of Lemma 2.5.2. From Lemma A.3.1, we see that U^0 obtained after step 3 of Algorithm 4 satisfies: $\text{dist}\,(U^0, U^*) \leq \frac{1}{64k}$. Lemma now follows by using the above mentioned observation with Lemma A.3.2. □

We now provide the two results used in the above lemma.

Lemma A.3.1. *After step* 3 *in Algorithm 4, whp we have:*

$$\text{dist}\left(U^0, U^*\right) \leq \frac{1}{64k}$$

Proof. From Theorem 3.1 in [50], we have the following result:

$$\|M - M_k\|_2 \leq C\left(\frac{k}{p\sqrt{mn}}\right)^{\frac{1}{2}}\|M\|_F.$$

Let $U^{(0)}\Sigma V^\dagger$ be the top k singular components of M_k. We also have:

$$\begin{aligned}
\|M - M_k\|_2^2 &= \left\|U^*\Sigma^*(V^*)^\dagger - U^{(0)}\Sigma V^\dagger\right\|_2^2 \\
&= \Big\|U^*\Sigma^*(V^*)^\dagger - U^{(0)}\left(U^{(0)}\right)^\dagger U^*\Sigma^*(V^*)^\dagger \\
&\qquad + U^{(0)}\left(U^{(0)}\right)^\dagger U^*\Sigma^*(V^*)^\dagger - U^{(0)}\Sigma V^\dagger\Big\|_2^2 \\
&= \Big\|\left(I - U^{(0)}\left(U^{(0)}\right)^\dagger\right)U^*\Sigma^*(V^*)^\dagger \\
&\qquad + U^{(0)}\left(\left(U^{(0)}\right)^\dagger U^*\Sigma^*(V^*)^\dagger - \Sigma V^\dagger\right)\Big\|_2^2 \\
&\overset{(\zeta_1)}{\geq} \left\|\left(I - U^{(0)}\left(U^{(0)}\right)^\dagger\right)U^*\Sigma^*(V^*)^\dagger\right\|_2^2 \\
&= \left\|\left(U_\perp^{(0)}\right)^\dagger U^*\Sigma^*\right\|_2^2 \geq (\sigma_k^*)^2\left\|\left(U_\perp^{(0)}\right)^\dagger U^*\right\|_2^2,
\end{aligned}$$

where (ζ_1) follows from the fact that the column space of the first two terms in the equation is $U_\perp^{(0)}$ where as the column space of the last two terms is $U^{(0)}$. Using the above two inequalities, we get:

$$\left\|\left(U_\perp^{(0)}\right)^\dagger U^*\right\|_2 \leq C\cdot\frac{\sigma_1^*}{\sigma_k^*}\cdot\frac{k}{\sqrt{mp}} \leq \frac{1}{10^4 k},$$

if $p > \frac{C'k^4\log n}{m}\cdot\frac{(\sigma_1^*)^2}{(\sigma_k^*)^2}$ for a large enough constant C'. □

Lemma A.3.2. *(Analysis of step 4 of Algorithm 4) Suppose U^* is incoherent with parameter μ and U is an orthonormal column matrix such that $\operatorname{dist}(U, U^*) \leq \frac{1}{64k}$. Let U^c be obtained from U by setting all entries greater than $\frac{2\mu\sqrt{k}}{\sqrt{n}}$ to zero. Let $\widetilde{U}$ be an orthonormal basis of U^c. Then,*

- $\operatorname{dist}\left(\widetilde{U}, U^*\right) \leq 1/2$ *and*
- $\widetilde{U}$ *is incoherent with parameter* $4\mu\sqrt{k}$.

Proof. Since $\operatorname{dist}(U, U^*) \leq d$, we have that for every i, $\exists \breve{u}_i \in \operatorname{Span}(U^*)$, $\|\breve{u}_i\|_2 = 1$ such that $\langle u_i, \breve{u}_i\rangle \geq \sqrt{1-d^2}$. Also, since $\breve{u}_i \in \operatorname{Span}(U^*)$, we have that $\breve{u}_i$ is incoherent with parameter $\mu\sqrt{k}$:

$$\|\breve{u}_i\|_2 = 1 \text{ and } \|\breve{u}_i\|_\infty \leq \frac{\mu\sqrt{k}}{\sqrt{m}}.$$

Let u_i^c be the vector obtained by setting all the elements of u_i with magnitude greater than $\frac{2\mu\sqrt{k}}{\sqrt{m}}$ to zero and let $u_i^{\bar{c}} \stackrel{\text{def}}{=} u_i - u_i^c$. Now, note that if for element j of u_i we have $\left|u_i^j\right| > \frac{2\mu\sqrt{k}}{\sqrt{m}}$, then, $|(u_i^c)^j - \breve{u}_i^j| = \left|\breve{u}_i^j\right| \leq \frac{\mu\sqrt{k}}{\sqrt{m}} \leq \left|u_i^j - \breve{u}_i^j\right|$. Hence,

$$\|u_i^c - \breve{u}_i\|_2 \stackrel{(\zeta_1)}{\leq} \|u_i - \breve{u}_i\|_2 = \left(\|u_i\|_2^2 + \|\breve{u}_i\|_2^2 - 2\langle u_i, \breve{u}_i\rangle\right)^{\frac{1}{2}} \leq \sqrt{2}d,$$

This also implies the following:

$$\begin{aligned}
&\|u_i^c\|_2 \geq \|\breve{u}_i\|_2 - \sqrt{2}d = 1 - \sqrt{2}d \quad \text{, and} \\
&\left\|u_i^{\bar{c}}\right\|_2 \leq \sqrt{1 - \|u_i^c\|_2^2} \leq \sqrt{2d(\sqrt{2} - d)} \leq 2\sqrt{d}, \quad \text{for } d < \frac{1}{\sqrt{2}}.
\end{aligned}$$

Let $U^c = \widetilde{U}\Lambda^{-1}$ (QR decomposition). Then, for any $u_\perp^* \in \text{Span}(U_\perp^*)$ we have:

$$\begin{aligned}
\left\|(u_\perp^*)^\dagger \widetilde{U}\right\|_2 &= \left\|(u_\perp^*)^\dagger U^c \Lambda\right\|_2 \leq \left\|(u_\perp^*)^\dagger U^c\right\|_2 \|\Lambda\|_2 \\
&\leq \left(\left\|(u_\perp^*)^\dagger U\right\|_2 + \left\|(u_\perp^*)^\dagger U^{\bar{c}}\right\|_2\right) \|\Lambda\|_2 \\
&\leq \left(d + \left\|U^{\bar{c}}\right\|_2\right) \|\Lambda\|_2 \leq \left(d + \left\|U^{\bar{c}}\right\|_{\text{F}}\right) \|\Lambda\|_2 \leq \left(d + 2\sqrt{kd}\right) \|\Lambda\|_2 \\
&\leq 3\sqrt{kd}\, \|\Lambda\|_2 .
\end{aligned}$$

We now bound $\|\Lambda\|_2$ as follows:

$$\begin{aligned}
\|\Lambda\|_2^2 = \frac{1}{\sigma_{\min}\left(\Lambda^{-1}\right)^2} = \frac{1}{\sigma_{\min}\left(\widetilde{U}\Lambda^{-1}\right)^2} = \frac{1}{\sigma_{\min}\left(U^c\right)^2} \leq \frac{1}{1 - \|U^{\bar{c}}\|_2^2} \leq \frac{1}{1 - 4kd} \\
\leq 4/3,
\end{aligned}$$

where we used the fact that $d < \frac{1}{16k}$. So we have:

$$\left\|(u_\perp^*)^\dagger \widetilde{U}\right\|_2 \leq 3\sqrt{kd} \cdot 4/3 = 4\sqrt{kd}.$$

This proves the first part of the lemma.

Incoherence of $\widetilde{U}$ follows using the following set of inequalities:

$$\mu(\widetilde{U}) = \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^\dagger \widetilde{U}\|_2 \leq \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^\dagger U^c \Lambda\| \leq \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^\dagger U^c\|_2 \|\Lambda\|_2 \leq 4\mu\sqrt{k}.$$

□

A.3.2 Rank-1 Matrix Completion

Proof Of Lemma 2.5.3. Using the definition of spectral norm,

$$\|B^{-1}\left(\langle u^*, u^t\rangle B - C\right)v^*\|_2 \leq \|B^{-1}\|_2 \|(\langle u^*, u^t\rangle B - C)v^*\|_2.$$

As B is a diagonal matrix, $\|B^{-1}\|_2 = \frac{1}{\min_i B_{ii}} \leq \frac{1}{1-\delta_2}$, where the last inequality follows using Lemma A.3.3. The lemma now follows using the above observation and Lemma A.3.4. □

Lemma A.3.3. *Let $M = \sigma^* u^* (v^*)^\dagger$, p, Ω, u^t be as defined in Lemma 2.5.3. Then, w.p. at least $1 - \frac{1}{n^3}$,*

$$\left|\frac{\sum_{i:(i,j)\in\Omega}(u_i^t)^2}{p} - 1\right| \leq \delta_2, \ \left|\frac{\sum_{i:(i,j)\in\Omega} u_i^t u_i^*}{p} - \langle u^t, u^*\rangle\right| \leq \delta_2.$$

Proof. Since the first part of the lemma is a direct consequence of the second part, we will prove only the second part. Let δ_{ij} be a Bernoulli random variable that indicates membership of index $(i,j) \in \Omega$. That is, $\delta_{ij} = 1$ w.p. p and 0 otherwise. Define $Z_j = \frac{1}{p}\sum_i \delta_{ij} u_i^t u_i^*$. Note that $\mathbb{E}[Z_j] = \langle u^t, u^*\rangle$. Furthermore, $\mathbb{E}[Z_j^2] = \left(\frac{1}{p} - 1\right)\sum_i (u_i^t u_i^*)^2 \leq \frac{\mu_1^2}{mp}$ and $\max_i |u_i^t u_i^*| \leq \frac{\mu_1^2}{m}$. Using Bernstein's inequality, we get:

$$\Pr(\left|Z_j - \langle u^t, u^*\rangle\right| > \delta_2) \leq \exp\left(-\frac{\delta_2^2 mp/2}{\mu_1^2 + \mu_1^2\delta_2/3}\right). \tag{29}$$

Using union bound (for all j) and for $p \geq \frac{9\mu_1^2 \log n}{m\delta_2^2}$, w.p. $1-\frac{1}{n^3}$: $\forall j, \langle u^t, u^*\rangle - \delta_2 \leq Z_j \leq \langle u^t, u^*\rangle + \delta_2$. □

Lemma A.3.4. *Let $M = \sigma^* u^* (v^*)^\dagger$, p, Ω, u^t be as defined in Lemma 2.5.3. Then, w.p. at least $1 - \frac{1}{n^3}$,*

$$\|(\langle u^*, u^t\rangle B - C)v^*\|_2 \leq \delta_2\sqrt{1 - \langle u^*, u^t\rangle^2}.$$

Proof. Let $x \in \mathbb{R}^n$ be a unit vector. Then, $\forall x$:

$$\begin{aligned} x^\dagger(\langle u^*, u^t\rangle B - C)v^* &= \frac{1}{p}\sum_{ij\in\Omega} x_j v_j^*(\langle u^*, u^t\rangle (u_i^t)^2 - u_i^t u_i^*) \\ &\overset{(\zeta_1)}{\leq} \frac{1}{p} C\sqrt{mp}\sqrt{\sum_j x_j^2 (v_j^*)^2}\sqrt{\sum_i (\langle u^*, u^t\rangle (u_i^t)^2 - u_i^t u_i^*)^2}, \\ &\overset{(\zeta_2)}{\leq} \frac{1}{p} C\frac{\sqrt{mp}\mu_1^2}{n}\sqrt{1 - \langle u^*, u^t\rangle^2}, \end{aligned} \tag{30}$$

where $C > 0$ is a global constant and (ζ_1) follows by using a modified version of Lemma 6.1 by [50] (see Lemma A.3.5) and (ζ_2) follows by using incoherence of v^* and u^t. Lemma now follows by observing that $\max_{x,\|x\|_2=1} x^\dagger(\langle u^*, u^t\rangle B - C)v^* = \|(\langle u^*, u^t\rangle B - C)v^*\|_2$ and $p > \frac{C\mu_1^2 \log n}{m\delta_2^2}$. □

Proof of Lemma 2.5.4. Using (15) and using the fact that B, C are diagonal matrices:

$$\widehat{v}_j^{t+1} = \sigma^*\langle u^t, u^*\rangle v_j^* - \frac{\sigma^*}{B_{jj}}\left(\langle u^t, u^*\rangle B_{jj} - C_{jj}\right) v_j^*.$$

We bound the largest magnitude of elements in $\widehat{v}^{t+1}$ as follows. For every $j \in [n]$, we have:

$$\begin{aligned} \left|\widehat{v}_j^{t+1}\right| &\leq \left|\sigma^*\langle u^t, u^*\rangle v_j^*\right| + \left|\frac{\sigma^*}{B_{jj}}\left(\langle u^t, u^*\rangle B_{jj} - C_{jj}\right) v_j^*\right| \\ &\overset{(\zeta_1)}{\leq} \sigma^*\langle u^t, u^*\rangle \frac{\mu}{\sqrt{n}} + \frac{\sigma^*}{1-\delta_2}\left(\langle u^t, u^*\rangle (1+\delta_2) + \left(\langle u^t, u^*\rangle + \delta_2\right)\right)\frac{\mu}{\sqrt{n}} \\ &\leq \frac{\frac{3\sigma^*(1+\delta_2)\mu}{1-\delta_2}}{\sqrt{n}} \leq \frac{\sigma^*\mu_1}{2\sqrt{n}}, \end{aligned}$$

where (ζ_1) follows from the fact that $1 - \delta_2 \leq B_{jj} \leq 1 + \delta_2$ and $|C_{jj}| \leq (|\langle u^t, u^*\rangle| + \delta_2)$ (please refer Lemma A.3.3).

Also, from (17) we see that:

$$\begin{aligned} \left\|\widehat{v}^{t+1}\right\|_2 \geq \langle \widehat{v}^{t+1}, v^*\rangle &\geq \sigma^*\langle u^t, u^*\rangle - 2\sigma^*\delta_2\sqrt{1 - \langle u^t, u^*\rangle^2} \\ &\geq \sigma^*\langle u^0, u^*\rangle - 2\sigma^*\delta_2\sqrt{1 - \langle u^0, u^*\rangle^2} \\ &\overset{(\zeta_1)}{\geq} \frac{\sigma^*}{2}, \end{aligned}$$

where (ζ_1) follows from the fact that $\text{dist}\,(u^0, u^*) \leq \frac{3}{50}$ (please refer Lemma 2.5.2). Using the above two inequalities, we obtain:

$$\left\|v^{t+1}\right\|_\infty = \frac{\|\widehat{v}^{t+1}\|_\infty}{\|\widehat{v}^{t+1}\|_2} \leq \frac{\left(\frac{\sigma^*\mu_1}{2\sqrt{n}}\right)}{\left(\frac{\sigma^*}{2}\right)} = \frac{\mu_1}{\sqrt{n}}.$$

This finishes the proof. □

Lemma A.3.5 (Modified version of Lemma 6.1 of [50])**.** *Let Ω be a set of indices sampled uniformly at random from $[m]\times[n]$ with each element of $[m]\times[n]$ sampled independently with probability $p \geq \frac{C\log n}{m}$. Then, w.p. at least $1-\frac{1}{n^3}$, $\forall x \in \mathbb{R}^m, y \in \mathbb{R}^n$ s.t. $\sum_i x_i = 0$, we have: $\sum_{ij\in\Omega} x_i y_j \leq C\sqrt{\sqrt{mn}p}\|x\|_2\|y\|_2$, where $C > 0$ is a global constant.*

A.3.3 General Rank-k Matrix Completion

Proof of Lemma 2.5.5. From the decoupled update equation, (20), we obtain:

$$(V^{t+1})^{(j)} = (R^{(t+1)})^{-1}(D^j - (B^j)^{-1}(B^jD^j - C^j))\Sigma^*(V^*)^{(j)}, \qquad 1 \leq j \leq n.$$

We bound the two norm of the $(V^{t+1})^{(j)}$ as follows:

$$\begin{aligned}
&\left\|(V^{t+1})^{(j)}\right\|_2 \\
&\leq \frac{\sigma_1\left\|(V^*)^{(j)}\right\|_2}{\sigma_{\min}\left(R^{(t+1)}\right)}\left(\left\|D^j\right\|_2 + \left\|(B^j)^{-1}(B^jD^j - C^j)\right\|_2\right) \\
&\leq \frac{\sigma_1\left\|(V^*)^{(j)}\right\|_2}{\sigma_{\min}\left(R^{(t+1)}\right)}\left(\left\|D^j\right\|_2 + \frac{\|B^jD^j\|_2 + \|C^j\|_2}{\sigma_{\min}\left(B^j\right)}\right) \\
&\overset{(\zeta_1)}{\leq} \frac{\sigma_1\frac{\mu\sqrt{k}}{\sqrt{n}}}{\sigma_k^*\sqrt{1-\text{dist}^2\left(U^{(t)},U^*\right)} - \frac{\sigma_1^*\delta_{2k}k\text{dist}(U^{(t)},U^*)}{1-\delta_{2k}}}\left(1 + \frac{(1+\delta_{2k})+(1+\delta_{2k})}{1-\delta_{2k}}\right) \\
&\leq \frac{4\sigma_1\frac{\mu\sqrt{k}}{\sqrt{n}}}{\sigma_k^*\sqrt{1-\text{dist}^2\left(U^{(0)},U^*\right)} - \frac{\sigma_1^*\delta_{2k}k\text{dist}(U^{(0)},U^*)}{1-\delta_{2k}}} \leq \frac{\left(\frac{16\sigma_1^*\mu}{\sigma_k^*}\right)\sqrt{k}}{\sqrt{n}},
\end{aligned}$$

where we used the following inequalities in (ζ_1):

$$\left\|(V^*)^j\right\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}}, \tag{31}$$

$$\sigma_{\min}\left(R^{(t+1)}\right) \geq \sigma_k^*\sqrt{1-\mathrm{dist}^2\left(U^{(t)},U^*\right)} - \sigma_1^*\delta_{2k}k\mathrm{dist}(U^{(t)},U^*), \tag{32}$$

$$\sigma_{\min}\left(B^j\right) \geq 1-\delta_{2k} \text{ and } \sigma_{\max}\left(B^j\right) \leq 1+\delta_{2k}, \tag{33}$$

$$\sigma_{\max}\left(C^j\right) \leq 1+\delta_{2k} \text{ and} \tag{34}$$

$$\sigma_{\max}\left(D^j\right) \leq 1, \tag{35}$$

where (31) follows from the incoherence of V^*, (32) follows from from an analysis similar to the proof of Lemma 2.4.6, (33) follows from (the proof of) Lemma A.3.6, (34) follows from Lemma A.3.7 and finally (35) follows from the fact that $D^j = \left(U^t\right)^\dagger U^*$ with U^t and U^* being orthonormal column matrices. □

Proof of Lemma 2.5.6. Note that,

$$\begin{aligned}\left\|F(\Sigma^*)^{-1}\right\|_2 \leq \left\|F(\Sigma^*)^{-1}\right\|_{\mathrm{F}} &= \left\|B^{-1}\left(BD-C\right)v^*\right\|_2 \\ &\leq \left\|B^{-1}\right\|_2 \left\|(BD-C)v^*\right\|_2 \\ &\leq \frac{\delta_{2k}}{1-\delta_{2k}}\mathrm{dist}(U^t,U^*),\end{aligned} \tag{36}$$

where the last inequality follows using Lemma A.3.6 and Lemma A.3.8. □

We now bound $\|B^{-1}\|_2$ and $\|C^j\|_2$, which is required by our bound for F as well as for our incoherence proof.

Lemma A.3.6. *Let M, Ω, p, and U^t be as defined in Theorem 2.3.3 and Lemma 2.5.6. Then, w.p. at least $1-\frac{1}{n^3}$:*

$$\|B^{-1}\|_2 \leq \frac{1}{1-\delta_{2k}}. \tag{37}$$

Proof of Lemma A.3.6. We have:

$$\|B^{-1}\|_2 = \frac{1}{\sigma_{min}(B)} = \frac{1}{\min_{x,\|x\|=1} x^\dagger Bx},$$

where $x \in \mathbb{R}^{nk}$. Let $x = vec(X)$, i.e., x_p is the p-th column of X and x^j is the j-th row of X. Now, $\forall x$,

$$x^\dagger B x = \sum_j (x^j)^\dagger B^j (x^j) \geq min_j \sigma_{min}(B^j).$$

Lemma would follow using the bound on $\sigma_{min}(B^j), \forall j$ that we show below.

Lower bound on $\sigma_{min}(B^j)$: Consider any $w \in \mathbb{R}^k$ such that $\|w\|_2 = 1$. We have:

$$Z = w^\dagger B^j w = \frac{1}{p} \sum_{i:(i,j)\in\Omega} \langle w, (U^t)^{(i)} \rangle^2 = \frac{1}{p} \sum_i \delta_{ij} \langle w, (U^t)^{(i)} \rangle^2.$$

Note that, $\mathbb{E}[Z] = w^\dagger U U^\dagger w = w^\dagger w = 1$ and $\mathbb{E}[Z^2] = \frac{1}{p} \sum_i \langle w, (U^t)^{(i)} \rangle^4 \leq \frac{\mu_1^2 k}{mp} \sum_i \langle w, (U^t)^{(i)} \rangle^2 = \frac{\mu_1^2 k}{mp}$, where the second last inequality follows using incoherence of U^t. Similarly, $\max_i |\langle w, (U^t)^{(i)} \rangle^2| \leq \frac{\mu_1^2 k}{mp}$. Hence, using Bernstein's inequality:

$$\Pr(|Z - \mathbb{E}[Z]| \geq \delta_{2k}) \leq \exp(-\frac{\delta_{2k}^2/2}{1 + \delta_{2k}/3} \frac{mp}{\mu_1^2 k}).$$

That is, by using p as in the statement of the lemma with the above equation and using union bound, we get (w.p. $> 1 - 1/n^3$): $\forall w, j \quad w^\dagger B^j w \geq 1 - \delta_{2k}$. That is, $\forall j, \sigma_{min}(B^j) \geq (1 - \delta_{2k})$. □

Lemma A.3.7. *Let M, Ω, p, and U^t be as defined in Theorem 2.3.3 and Lemma 2.5.6. Also, let $C^j \in \mathbb{R}^{k \times k}$ be defined as: $C^j = \frac{1}{p} \sum_{i:(i,j)\in\Omega} (U^t)^{(i)} (U^*)^{(i)^\dagger}$. Then, w.p. at least $1 - \frac{1}{n^3}$:*

$$\|C^j\|_2 \leq 1 + \delta_{2k}, \forall j \tag{38}$$

Proof of Lemma A.3.7. Let $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^k$ be two arbitrary unit vectors. Then,

$$x^T C^j y = \frac{1}{p} \sum_{i:(i,j)\in\Omega} (x^\dagger (U^t)^{(i)})(y^\dagger (U^*)^{(i)}).$$

That is, $Z = x^T C^j y = \frac{1}{p} \sum_i \delta_{ij} (x^\dagger (U^t)^{(i)})(y^\dagger (U^*)^{(i)})$. Note that, $\mathbb{E}[Z] = x^\dagger (U^t)^\dagger U^* y$, $\mathbb{E}[Z^2] = \frac{1}{p} \sum_i (x^\dagger (U^t)^{(i)})^2 (y^\dagger (U^*)^{(i)})^2 \leq \frac{\mu^2}{mp} x^\dagger (U^t)^\dagger U^t x = \frac{\mu^2 k}{mp}$ and $\max_i |(x^\dagger (U^t)^{(i)})(y^\dagger (U^*)^{(i)})| \leq \frac{\mu_1^2 k}{m}$. Lemma now follows using Bernstein's inequality and using bound for p given in the lemma statement. □

Finally, we provide a lemma to bound the second part of the error term (F).

Lemma A.3.8. *Let M, Ω, p, and U^t be as defined in Theorem 2.3.3 and Lemma 2.5.6. Then, w.p. at least $1 - \frac{1}{n^3}$:*

$$\|(BD - C)v^*\|_2 \leq \delta_{2k}\text{dist}(V^{t+1}, V^*), \tag{39}$$

where $v^* = vec(V^*)$, i.e. $v^* = \begin{bmatrix} V_1^* \\ \vdots \\ V_k^* \end{bmatrix}$.

Proof of Lemma A.3.8. Let $X \in \mathbb{R}^{n\times k}$ and let $x = vec(X) \in \mathbb{R}^{nk}$ s.t. $\|x\|_2 = 1$. Also, let x_p be the p-th column of X and x^j be the j-th column of X.

Let $u^i = (U^t)^{(i)}$ and $u^{*(i)} = (U^*)^{(i)}$. Also, let $H^j = (B^j D - C^j)$, i.e.,

$$H^j = \frac{1}{p}\sum_{i:(i,j)\in\Omega} u^i(u^i)^\dagger(U^t)^\dagger U^* - u^i(u^{*(i)})^\dagger = \frac{1}{p}\sum_{i:(i,j)\in\Omega} H_i^j,$$

where $H_i^j \in \mathbb{R}^{k\times k}$. Note that,

$$\sum_i H_i^j = (U^t)^\dagger U^t (U^t)^\dagger U^* - (U^t)^\dagger U^* = 0. \tag{40}$$

Now, $x^\dagger(BD-C)v^* = \sum_j (x^j)^\dagger(B^jD-C^j)(V^*)^{(j)} = \frac{1}{p}\sum_{pq}\sum_{(i,j)\in\Omega} x_p^j V_{jq}^*(H_i^j)_{pq}$. Also, using (40), $\forall(p,q)$:

$$\sum_i (H_i^j)_{pq} = 0.$$

Hence, applying Lemma A.3.5, we get w.p. at least $1 - \frac{1}{n^3}$:

$$x^\dagger(BD-C)v^* = \sum_j (x^j)^\dagger(B^jD-C^j)(V^*)^{(j)} \leq \frac{1}{p}\sum_{pq}\sqrt{\sum_j (x_p^j)^2(V_{jq}^*)^2}\sqrt{\sum_i (H_i^j)_{pq}^2}. \tag{41}$$

Also,

$$\begin{aligned}\sum_i (H_i^j)_{pq}^2 &= \sum_i (u_p^i)^2((u^i)^\dagger(U^t)^\dagger U_q^* - U_{iq}^*)^2 \leq \max_i (u_p^i)^2 \sum_i ((u^i)^\dagger(U^t)^\dagger U_q^* - U_{iq}^*)^2 \\ &= \max_i (u_p^i)^2(1 - \|U^t U_q^*\|_2^2) \leq \frac{\mu_1^2 k}{m}\text{dist}(U^t, U^*)^2. \end{aligned} \tag{42}$$

Using (41), (42) and incoherence of V^*, we get (w.p. $1 - 1/n^3$), $\forall x$:

$$x^\dagger(BD - C)v^* \leq \sum_{pq} \frac{\mu_1^2 k}{mp} \text{dist}(U^t, U^*)\|x_p\|_2 \leq \delta_{2k}\text{dist}(U^t, U^*),$$

where we used the fact that $\sum_p \|x_p\|_2 \leq \sqrt{k}\,\|x\|_2 = \sqrt{k}$ in the last step. Lemma now follows by observing $\max_{x,\|x\|=1} x^\dagger(BD-C)v^* = \|(BD-C)v^*\|_2$. □

Appendix B

Proofs for Phase Retrieval using Alternating Minimization

B.1 Proofs for Section 3.5

B.1.1 Proof of the Initialization Step

Proof of Theorem 3.5.1. Recall that x^0 is the top singular vector of $S = \frac{1}{n}\sum_\ell |{a_\ell}^T x^*|^2 a_\ell {a_\ell}^T$. As a_ℓ are rotationally invariant random variables, wlog, we can assume that $x^* = e_1$ where e_1 is the first canonical basis vector. Also note that $\mathbb{E}\left[|\langle a, e_1\rangle|^2 aa^T\right] = D$, where D is a diagonal matrix with $D_{11} = \mathbb{E}_{a\sim\mathcal{N}_C(0,1)}[|a|^4] = 8$ and $D_{ii} = \mathbb{E}_{a\sim\mathcal{N}_C(0,1),b\sim\mathcal{N}_C(0,1)}[|a|^2|b|^2] = 1, \forall i > 1$.

We break our proof of the theorem into two steps:

(1): Show that, with probability $> 1 - \frac{4}{m^2}$: $\|S - D\|_2 < c/4$.

(2): Use (1) to prove the theorem.

Proof of Step (2): We have $\langle x^0, Sx^0\rangle \leq c/4 + 3\left((x^0)^T e_1\right)^2 + \sum_{i=2}^n (x^0{}_i)^2 = c/4 + \langle x^0, Sx^0\rangle > 3 - c/4$. Hence, $\langle x^0, e_1\rangle^2 > 1 - c/2$. This yields $\|x^0 - x^*\|_2^2 = 2 - 2\langle x^0, e_1\rangle^2 < c$.

Proof of Step (1): We now complete our proof by proving (1). To this end, we use the following matrix concentration result from [85]:

Theorem B.1.1 (Theorem 1.5 of [85])**.** *Consider a finite sequence X_i of self-adjoint independent random matrices with dimensions $n \times n$. Assume that $\mathbb{E}[X_i] = 0$ and $\|X_i\|_2 \leq R, \forall i$, almost surely. Let $\sigma^2 := \|\sum_i \mathbb{E}[X_i]\|_2$. Then the following holds $\forall \nu \geq 0$:*

$$P\left(\|\frac{1}{m}\sum_{i=1}^m X_i\|_2 \geq \nu\right) \leq 2n\exp\left(\frac{-m^2\nu^2}{\sigma^2 + Rm\nu/3}\right).$$

Note that Theorem B.1.1 assumes $\max_\ell |a_{1\ell}|^2\|a_\ell\|^2$ to be bounded, where $a_{1\ell}$ is the first component of a_ℓ. However, a_ℓ is a normal random variable and hence can be unbounded. We address this issue by observing that probability that $\Pr(\|a_\ell\|^2 \geq 2n \; OR \; |a_{1\ell}|^2 \geq 2\log m) \leq 2\exp(-n/2) + \frac{1}{m^2}$. Hence, for large enough $n, \widehat{c}$ and $m > \widehat{c}n$, w.p. $1 - \frac{3}{m^2}$,

$$\max_\ell |a_{1\ell}|^2\|a_\ell\|^2 \leq 4n\log(m). \tag{1}$$

Now, consider truncated random variable $\widetilde{a}_\ell$ s.t. $\widetilde{a}_\ell = a_\ell \quad if \quad |a_{1\ell}|^2 \leq 2\log(m) \& \|a_\ell\|^2 \leq 2n$ and $\widetilde{a}_\ell = 0$ otherwise. Now, note that $\widetilde{a}_\ell$ is symmetric around origin and also $\mathbb{E}[\widetilde{a}_{i\ell}\widetilde{a}_{j\ell}] = 0, \forall i \neq j$. Also, $\mathbb{E}[|\widetilde{a}_{i\ell}|^2] \leq 1$. Hence, $\|\mathbb{E}[|\widetilde{a}_{1\ell}|^2\|\widetilde{a}_\ell\|^2\widetilde{a}_\ell\widetilde{a}_\ell^\dagger]\|_2 \leq 4n\log(m)$. Now, applying Theorem B.1.1 given above, we get (w.p. $\geq 1 - 1/m^2$)

$$\|\frac{1}{m}\sum_\ell |\widetilde{a}_{1\ell}|^2\widetilde{a}_\ell\widetilde{a}_\ell^\dagger - \mathbb{E}[|\widetilde{a}_{1\ell}|^2\widetilde{a}_\ell\widetilde{a}_\ell^\dagger]\|_2 \leq \frac{4n\log^{3/2}(m)}{\sqrt{m}}.$$

Furthermore, $a_\ell = \widetilde{a}_\ell$ with probability larger than $1 - \frac{3}{m^2}$. Hence, w.p. $\geq 1 - \frac{4}{m^2}$:

$$\|S - \mathbb{E}[|\widetilde{a}_\ell^1|^2\widetilde{a}_\ell\widetilde{a}_\ell^\dagger]\|_2 \leq \frac{4n\log^{3/2}(m)}{\sqrt{m}}.$$

Now, the remaining task is to show that $\|\mathbb{E}[|\widetilde{a}_\ell^1|^2\widetilde{a}_\ell\widetilde{a}_\ell^\dagger] - \mathbb{E}[|a_\ell^1|^2 a_\ell a_\ell^\dagger]\|_2 \leq \frac{1}{m}$. This follows easily by observing that $\mathbb{E}[\widetilde{a}_\ell^i\widetilde{a}_\ell^j] = 0$ and by bounding $\mathbb{E}[|\widetilde{a}_\ell^1|^2|\widetilde{a}_\ell^i|^2 - |a_\ell^1|^2|a_\ell^i|^2 \leq 1/m$ by using a simple second and fourth moment calculations for the normal distribution.

□

B.1.2 Proof of per step reduction in error

In all the lemmas in this section, δ is a small numerical constant (can be taken to be 0.01).

Lemma B.1.2. *Assume the hypothesis of Theorem 3.5.2 and let x^+ be as defined in* (3)*. Then, there exists an absolute numerical constant c such that the following holds (w.p. $\geq 1 - \frac{\eta}{4}$): $\left\| \left(AA^T\right)^{-1} A\left(D - I\right)A^T x^* \right\|_2 < c dist\left(x^*, x\right).$*

Proof. Using (4) and the fact that $\|x^*\|_2 = 1$, and

$$x^{*T}x^+ = 1 + x^{*T}\left(AA^T\right)^{-1} A\left(D-I\right)A^T x^*,$$

we have, $|x^{*T}x^+| \geq 1 - \|\left(\frac{1}{2m}AA^T\right)^{-1}\|_2 \|\frac{1}{\sqrt{2m}}A\|_2 \|\frac{1}{\sqrt{2m}}\left(D-I\right)A^T x^*\|_2$. Now, using standard bounds on the singular values of Gaussian matrices [89] and assuming $m > \widehat{c}\log\frac{1}{\eta}n$, we have (w.p. $\geq 1 - \frac{\eta}{4}$): $\|\left(\frac{1}{2m}AA^T\right)^{-1}\|_2 \leq 1/(1 - 2/\sqrt{\widehat{c}})^2$ and $\|A\|_2 \leq 1 + 2/\sqrt{\widehat{c}}$. Note that both the quantities can be bounded by constants that are close to 1 by selecting a large enough $\widehat{c}$. Also note that $\frac{1}{2m}AA^T$ converges to I (the identity matrix), or equivalently $\frac{1}{m}AA^T$ converges to $2I$ since the elements of A are standard normal complex random variables and not standard normal real random variables.

The key challenge now is to bound $\left\|\left(D-I\right)A^T x^*\right\|_2$ by $c\sqrt{m}dist\left(x^*, x^t\right)$ for a global constant $c > 0$. Note that since (4) is invariant with respect to $\|x^t\|_2$, we can assume that $\|x^t\|_2 = 1$. Note further that, since the distribution of A is rotationally invariant and is independent of x^* and x^t, wlog, we can assume that $x^* = e_1$ and $x^t = \alpha e_1 + \sqrt{1-\alpha^2}e_2$, where $\alpha = \langle x^t, x^*\rangle$.

$$\left\|\left(D-I\right)A^T e_1\right\|_2^2 = \sum_{l=1}^{m}|a_{1l}|^2\left|\mathrm{Ph}\left(\left(\alpha\overline{a}_{1l} + \sqrt{1-\alpha^2}\overline{a}_{2l}\right)a_{1l}\right) - 1\right|^2 = \sum_{l=1}^{m}U_\ell,$$

where U_l is given by,

$$U_l \stackrel{\text{def}}{=} |a_{1l}|^2\left|\mathrm{Ph}\left(\left(\alpha\overline{a}_{1l} + \sqrt{1-\alpha^2}\overline{a}_{2l}\right)a_{1l}\right) - 1\right|^2. \tag{2}$$

Using Lemma B.1.3 finishes the proof. □

The following lemma, Lemma B.1.3 shows that if U_ℓ are as defined in Lemma B.1.2 then, the sum of $U_\ell, 1 \leq \ell \leq m$ concentrates well around $\mathbb{E}\left[U_\ell\right]$ and also $\mathbb{E}\left[U_\ell\right] \leq c\sqrt{m}dist\left(x^*, x^t\right)$. The proof of Lemma B.1.3 requires careful analysis as it provides tail bound and expectation bound of a random variable that is a product of correlated sub-exponential complex random variables.

Lemma B.1.3. *Assume the hypothesis of Lemma B.1.2. Let U_ℓ be as defined in (2) and let each $a_{1l}, a_{2l}, \forall 1 \leq l \leq m$ be sampled from standard normal distribution for complex numbers. Then, with probability greater than $1 - \frac{\eta}{4}$, we have: $\sum_{l=1}^{m}U_l \leq c^2 m(1-\alpha^2)$, for a global constant $c > 0$.*

Proof of Lemma B.1.3. We first estimate $\mathbb{P}[U_l > t]$ so as to:

1. Calculate $\mathbb{E}[U_l]$ and,
2. Show that U_l is a subexponential random variable and use that fact to derive concentration bounds.

Now, $\mathbb{P}[U_l > t] = \int_{\frac{\sqrt{t}}{2}}^{\infty} p_{|a_{1l}|}(s)\mathbb{P}\left[W_l > \frac{\sqrt{t}}{s}\middle||a_{1l}| = s\right] ds$, where,

$$W_l \stackrel{\text{def}}{=} \left|\text{Ph}\left(\left(\alpha\overline{a}_{1l} + \sqrt{1-\alpha^2}\overline{a}_{2l}\right)a_{1l}\right) - 1\right|.$$

$$\begin{aligned}
&\mathbb{P}\left[W_l > \frac{\sqrt{t}}{s}\middle||a_{1l}| = s\right] \\
&= \mathbb{P}\left[\left|\text{Ph}\left(\left(\alpha\overline{a}_{1l} + \sqrt{1-\alpha^2}\overline{a}_{2l}\right)a_{1l}\right) - 1\right| > \frac{\sqrt{t}}{s}\middle||a_{1l}| = s\right] \\
&= \mathbb{P}\left[\left|\text{Ph}\left(1 + \frac{\sqrt{1-\alpha^2}\overline{a}_{2l}}{\alpha\overline{a}_{1l}}\right) - 1\right| > \frac{\sqrt{t}}{s}\middle||a_{1l}| = s\right] \\
&\stackrel{(\zeta_1)}{\leq} \mathbb{P}\left[\frac{\sqrt{1-\alpha^2}\,|a_{2l}|}{\alpha\,|a_{2l}|} > \frac{c\sqrt{t}}{s}\middle||a_{1l}| = s\right] \\
&= \mathbb{P}\left[|a_{2l}| > \frac{c\alpha\sqrt{t}}{\sqrt{1-\alpha^2}}\right] \\
&\stackrel{(\zeta_2)}{\leq} \exp\left(1 - \frac{c\alpha^2 t}{1-\alpha^2}\right),
\end{aligned}$$

where (ζ_1) follows from Lemma B.1.8 and (ζ_2) follows from the fact that a_{2l} is a sub-gaussian random variable. So we have:

$$\begin{aligned}
\mathbb{P}[U_l > t] \leq \int_{\frac{\sqrt{t}}{2}}^{\infty} \exp\left(1 - \frac{c\alpha^2 t}{1-\alpha^2}\right) ds \quad &= \exp\left(1 - \frac{c\alpha^2 t}{1-\alpha^2}\right) \int_{\frac{\sqrt{t}}{2}}^{\infty} s e^{-\frac{s^2}{2}} ds \\
&= \exp\left(1 - \frac{ct}{1-\alpha^2}\right). \qquad (3)
\end{aligned}$$

Using this, we have the following bound on the expected value of U_l:

$$\mathbb{E}[U_l] = \int_0^{\infty} \mathbb{P}[U_l > t]\, dt \leq \int_0^{\infty} \exp\left(1 - \frac{ct}{1-\alpha^2}\right) dt \leq c\left(1-\alpha^2\right). \qquad (4)$$

From (3), we see that U_l is a subexponential random variable with parameter $c\left(1-\alpha^2\right)$. Using Proposition 5.16 from [89], we obtain:

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{l=1}^{m} U_l - \mathbb{E}\left[U_l\right]\right| > \delta m\left(1-\alpha^2\right)\right] \\
&\leq 2\exp\left(-\min\left(\frac{c\delta^2 m^2\left(1-\alpha^2\right)^2}{\left(1-\alpha^2\right)^2 m}, \frac{c\delta m\left(1-\alpha^2\right)}{1-\alpha^2}\right)\right) \\
&\leq 2\exp\left(-c\delta^2 m\right) \leq \frac{\eta}{4}.
\end{aligned}$$

So, with probability greater than $1-\frac{\eta}{4}$, we have:

$$\sum_{l=1}^{m} U_l \leq c^2 m(1-\alpha^2).$$

This proves the lemma. □

Lemma B.1.4. *Assume the hypothesis of Theorem 3.5.2 and let x^+ be as defined in* (3)*. Then, $\forall z$ s.t. $\langle z, x^*\rangle = 0$, the following holds (w.p. $\geq 1-\frac{\eta}{4}e^{-n}$): $|\langle z, x^+\rangle| \leq \frac{5}{9} dist\left(x^*, x\right)$.*

Proof. Fix z such that $\langle z, x^*\rangle = 0$. Since the distribution of A is rotationally invariant, wlog we can assume that: a) $x^* = e_1$, b) $x = \alpha e_1 + \sqrt{1-\alpha^2}e_2$ where $\alpha \in \mathbb{R}$ and $\alpha \geq 0$ and c) $z = \beta e_2 + \sqrt{1-|\beta|^2}e_3$ for some $\beta \in \mathfrak{C}$. Note that we first prove the lemma for a *fixed* z and then using union bound, we obtain the result $\forall z \in \mathbb{C}^n$. We have:

$$\left|\langle z, x^+\rangle\right| \leq |\beta|\,|\langle e_2, x^+\rangle| + \sqrt{1-|\beta|^2}|\langle e_3, x^+\rangle|. \tag{5}$$

Now,

$$
\begin{aligned}
&\left|{e_2}^T x^+\right| \\
&= \left|{e_2}^T \left(AA^T\right)^{-1} A\left(D-I\right)A^T e_1\right| \\
&\leq \frac{1}{2m}\left|{e_2}^T\left(\left(\frac{1}{2m}AA^T\right)^{-1}-I\right)A\left(D-I\right)A^T e_1\right| + \frac{1}{2m}\left|{e_2}^T A\left(D-I\right)A^T e_1\right| \\
&\leq \frac{1}{2m}\left\|\left(\frac{1}{2m}AA^T\right)^{-1}-I\right\|_2 \|A\|_2 \left\|\left(D-I\right)A^T e_1\right\|_2 + \frac{1}{2m}\left|{e_2}^T A\left(D-I\right)A^T e_1\right|, \\
&\leq \frac{4c}{\sqrt{\hat{c}}} dist\left(x^t, x^*\right) + \frac{1}{2m}\left|{e_2}^T A\left(D-I\right)A^T e_1\right|, \qquad (6)
\end{aligned}
$$

where the last inequality follows from the proof of Lemma B.1.2. Similarly,

$$
\begin{aligned}
&\left|{e_3}^T x^+\right| \\
&= \left|{e_3}^T \left(AA^T\right)^{-1} A\left(D-I\right)A^T e_1\right| \\
&\leq \frac{1}{2m}\left|{e_3}^T\left(\left(\frac{1}{2m}AA^T\right)^{-1}-I\right)A\left(D-I\right)A^T e_1\right| + \frac{1}{2m}\left|{e_3}^T A\left(D-I\right)A^T e_1\right| \\
&\leq \frac{1}{2m}\left\|\left(\frac{1}{2m}AA^T\right)^{-1}-I\right\|_2 \|A\|_2 \left\|\left(D-I\right)A^T e_1\right\|_2 + \frac{1}{2m}\left|{e_3}^T A\left(D-I\right)A^T e_1\right| \\
&\leq \frac{4c}{\sqrt{\hat{c}}} dist\left(x^t, x^*\right) + \frac{1}{2m}\left|{e_3}^T A\left(D-I\right)A^T e_1\right|, \qquad (7)
\end{aligned}
$$

Again, the last inequality follows from the proof of Lemma B.1.2. The lemma now follows by using (5), (6), (7) along with Lemmas B.1.5 and B.1.7. □

Lemma B.1.5. *Assume the hypothesis of Theorem 3.5.2 and the notation therein. Then,*

$$\left|{e_2}^T A\left(D-I\right)A^T e_1\right| \leq \frac{100}{99} m\sqrt{1-\alpha^2},$$

with probability greater than $1-\frac{\eta}{10}e^{-n}$.

Proof. We have:

$$\begin{aligned} {e_2}^T A \left(D - I\right) A^T e_1 &= \sum_{l=1}^{m} \overline{a}_{1l} a_{2l} \left(\mathrm{Ph}\left(\left(\alpha \overline{a}_{1l} + \sqrt{1-\alpha^2}\overline{a}_{2l} \right) a_{1l} \right) - 1 \right) \\ &= \sum_{l=1}^{m} \left|a_{1l}\right| a'_{2l} \left(\mathrm{Ph}\left(\alpha \left|a_{1l}\right| + \sqrt{1-\alpha^2}\overline{a'_{2l}} \right) - 1 \right), \end{aligned}$$

where $a'_{2l} \stackrel{\text{def}}{=} a_{2l}\mathrm{Ph}\left(\overline{a}_{1l}\right)$ is identically distributed to a_{2l} and is independent of $|a_{1l}|$. Define the random variable U_l as:

$$U_l \stackrel{\text{def}}{=} \left|a_{1l}\right| a'_{2l} \left(\mathrm{Ph}\left(1 + \frac{\sqrt{1-\alpha^2}\overline{a'_{2l}}}{\alpha \left|a_{1l}\right|} \right) - 1 \right).$$

Similar to Lemma B.1.2, we will calculate $\mathbb{P}\left[U_l > t\right]$ to show that U_l is subexponential and use it to derive concentration bounds. However, using the above estimate to bound $\mathbb{E}\left[U_l\right]$ will result in a weak bound that we will not be able to use. Lemma B.1.6 bounds $\mathbb{E}\left[U_l\right]$ using a different technique carefully.

$$\begin{aligned} \mathbb{P}\left[\left|U_l\right| > t\right] &\leq \mathbb{P}\left[\left|a_{1l}\right| \left|a'_{2l}\right| \frac{c\sqrt{1-\alpha^2}\left|a'_{2l}\right|}{\alpha \left|a_{1l}\right|} > t \right] \\ &= \mathbb{P}\left[\left|a'_{2l}\right|^2 > \frac{c\alpha t}{\sqrt{1-\alpha^2}} \right] \leq \exp\left(1 - \frac{c\alpha t}{\sqrt{1-\alpha^2}} \right), \end{aligned}$$

where the last step follows from the fact that a'_{2l} is a subgaussian random variable and hence $\left|a'_{2l}\right|^2$ is a subexponential random variable. Using Proposition 5.16 from [89], we obtain:

$$\begin{aligned} &\mathbb{P}\left[\left| \sum_{l=1}^{m} U_l - \mathbb{E}\left[U_l\right] \right| > \delta m \sqrt{1-\alpha^2} \right] \\ &\leq 2\exp\left(-\min\left(\frac{c\delta^2 m^2 \left(1-\alpha^2\right)}{\left(1-\alpha^2\right) m}, \frac{c\delta m\sqrt{1-\alpha^2}}{\sqrt{1-\alpha^2}} \right) \right) \\ &\leq 2\exp\left(-c\delta^2 m\right) \leq \frac{\eta}{10}\exp\left(-n\right). \end{aligned}$$

Using Lemma B.1.6, we obtain:

$$\left|{e_2}^T A \left(D - I\right) A^T e_1\right| = \left| \sum_{l=1}^{m} U_l \right| \leq \left(1+\delta\right) m\sqrt{1-\alpha^2},$$

with probability greater than $1 - \frac{\eta}{10}\exp(-n)$. This proves the lemma. □

Lemma B.1.6. *Let w_1 and w_2 be two independent standard complex Gaussian random variables*[1]*. Let $U = |w_1|\, w_2 \left(\mathrm{Ph}\left(1 + \frac{\sqrt{1-\alpha^2}\overline{w_2}}{\alpha|w_1|}\right) - 1\right)$. Fix $\delta > 0$. Then, there exists a constant $\gamma > 0$ such that if $\sqrt{1-\alpha^2} < \gamma$, then: $\mathbb{E}[U] \leq (1+\delta)\sqrt{1-\alpha^2}$.*

Proof. Let $w_2 = |w_2|\, e^{i\theta}$. Then $|w_1|, |w_2|$ and θ are all independent random variables. θ is a uniform random variable over $[-\pi, \pi]$ and $|w_1|$ and $|w_2|$ are identically distributed with probability distribution function:

$$p(x) = x \exp\left(-\frac{x^2}{2}\right) \mathbb{1}_{\{x \geq 0\}}.$$

We have:

$$\begin{aligned}
\mathbb{E}[U] &= \mathbb{E}\left[|w_1|\,|w_2|\, e^{i\theta}\left(\mathrm{Ph}\left(1 + \frac{\sqrt{1-\alpha^2}\,|w_2|\, e^{-i\theta}}{\alpha\,|w_1|}\right) - 1\right)\right] \\
&= \mathbb{E}\left[|w_1|\,|w_2|\,\mathbb{E}\left[e^{i\theta}\left(\mathrm{Ph}\left(1 + \frac{\sqrt{1-\alpha^2}\,|w_2|\, e^{-i\theta}}{\alpha\,|w_1|}\right) - 1\right)\right]\middle| |w_1|, |w_2|\right]
\end{aligned}$$

Let $\beta \stackrel{\text{def}}{=} \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|}$. We will first calculate $\mathbb{E}\left[e^{i\theta}\mathrm{Ph}\left(1+\beta e^{-i\theta}\right)\middle||w_1|, |w_2|\right]$. Note that the above expectation is taken only over the randomness in θ. For simplicity of notation, we will drop the conditioning variables, and calculate the above expectation in terms of β.

$$\begin{aligned}
e^{i\theta}\mathrm{Ph}\left(1+\beta e^{-i\theta}\right) &= (\cos\theta + i\sin\theta)\,\frac{1+\beta\cos\theta - i\beta\sin\theta}{\left[(1+\beta\cos\theta)^2 + \beta^2\sin^2\theta\right]^{\frac{1}{2}}} \\
&= \frac{\cos\theta + \beta + i\sin\theta}{(1+\beta^2+2\beta\cos\theta)^{\frac{1}{2}}}.
\end{aligned}$$

We will first calculate the imaginary part of the above expectation:

$$\mathrm{Im}\left(\mathbb{E}\left[e^{i\theta}\mathrm{Ph}\left(1+\beta e^{-i\theta}\right)\right]\right) = \mathbb{E}\left[\frac{\sin\theta}{(1+\beta^2+2\beta\cos\theta)^{\frac{1}{2}}}\right] = 0, \tag{8}$$

[1] z is standard complex Gaussian if $z = z_1 + iz_2$ where z_1 and z_2 are independent standard normal random variables.

where the last step follows because we are taking the expectation of an odd function. Focusing on the real part, we let:

$$F(\beta) \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{\cos\theta+\beta}{(1+\beta^2+2\beta\cos\theta)^{\frac{1}{2}}}\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\cos\theta+\beta}{(1+\beta^2+2\beta\cos\theta)^{\frac{1}{2}}}d\theta.$$

Note that $F(\beta):\mathbb{R}\to\mathbb{R}$ and $F(0)=0$. We will show that there is a small absolute numerical constant γ (depending on δ) such that:

$$0<\beta<\gamma \Rightarrow |F(\beta)| \leq (\frac{1}{2}+\delta)\beta. \tag{9}$$

We show this by calculating $F'(0)$ and using the continuity of $F'(\beta)$ at $\beta=0$. We first calculate $F'(\beta)$ as follows:

$$\begin{aligned} F'(\beta) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{(1+\beta^2+2\beta\cos\theta)^{\frac{1}{2}}} - \frac{(\cos\theta+\beta)(\beta+\cos\theta)}{(1+\beta^2+2\beta\cos\theta)^{\frac{3}{2}}}d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sin^2\theta}{(1+\beta^2+2\beta\cos\theta)^{\frac{3}{2}}}d\theta \end{aligned}$$

From the above, we see that $F'(0)=\frac{1}{2}$ and (9) then follows from the continuity

of $F'(\beta)$ at $\beta = 0$. Getting back to the expected value of U, we have:

$$\begin{aligned}
|\mathbb{E}[U]| &= \left| \mathbb{E}\left[|w_1|\,|w_2|\, F\left(\frac{\sqrt{1-\alpha^2}\,|w_2|}{\alpha\,|w_1|} \right) \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} < \gamma \right\}} \right] \right. \\
&\quad \left. + \mathbb{E}\left[|w_1|\,|w_2|\, F\left(\frac{\sqrt{1-\alpha^2}\,|w_2|}{\alpha\,|w_1|} \right) \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} \geq \gamma \right\}} \right] \right| \\
&= \left| \mathbb{E}\left[|w_1|\,|w_2|\, F\left(\frac{\sqrt{1-\alpha^2}\,|w_2|}{\alpha\,|w_1|} \right) \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} < \gamma \right\}} \right] \right| \\
&\quad + \left| \mathbb{E}\left[|w_1|\,|w_2|\, F\left(\frac{\sqrt{1-\alpha^2}\,|w_2|}{\alpha\,|w_1|} \right) \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} \geq \gamma \right\}} \right] \right| \\
&\overset{(\zeta_1)}{\leq} \left(\frac{1}{2} + \delta \right) \mathbb{E}\left[|w_1|\,|w_2|\, \frac{\sqrt{1-\alpha^2}\,|w_2|}{\alpha\,|w_1|} \right] + \mathbb{E}\left[|w_1|\,|w_2|\, \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} \geq \gamma \right\}} \right], \\
&= \left(\frac{1}{2} + \delta \right) \left(\frac{\sqrt{1-\alpha^2}}{\alpha} \right) \mathbb{E}\left[|w_2|^2 \right] + \mathbb{E}\left[|w_1|\,|w_2|\, \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} \geq \gamma \right\}} \right], \\
&\overset{(\zeta_2)}{=} (1 + 2\delta) \left(\frac{\sqrt{1-\alpha^2}}{\alpha} \right) + \mathbb{E}\left[|w_1|\,|w_2|\, \mathbb{1}_{\left\{ \frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|} \geq \gamma \right\}} \right], \qquad (10)
\end{aligned}$$

where (ζ_1) follows from (9) and the fact that $|F(\beta)| \leq 1$ for every β and (ζ_2) follows from the fact that $\mathbb{E}\left[|z_2|^2 \right] = 2$. We will now bound the second term in the above inequality. We start with the following integral:

$$\begin{aligned}
\int_t^\infty s^2 e^{-\frac{s^2}{2}} ds &= -\int_t^\infty s d\left(e^{-\frac{s^2}{2}} \right) \\
&= t e^{-\frac{t^2}{2}} + \int_t^\infty e^{-\frac{s^2}{2}} ds \leq (t + e) e^{-\frac{t^2}{c}}, \qquad (11)
\end{aligned}$$

where c is some constant. The last step follows from standard bounds on the tail probabilities of gaussian random variables. We now bound the second

term of (10) as follows:

$$\begin{aligned}
\mathbb{E}\left[|w_1|\,|w_2|\,\mathbb{1}_{\left\{\frac{\sqrt{1-\alpha^2}|w_2|}{\alpha|w_1|}\geq\gamma\right\}}\right] &= \int_0^\infty t^2 e^{-\frac{t^2}{2}} \int_{\frac{\alpha t}{\sqrt{1-\alpha^2}}}^\infty s^2 e^{-\frac{s^2}{2}} ds dt \\
&\overset{(\zeta_1)}{\leq} \int_0^\infty t^2 e^{-\frac{t^2}{2}} \left(\frac{\alpha t}{\sqrt{1-\alpha^2}} + e\right) e^{-\frac{\alpha^2 t^2}{c(1-\alpha^2)}} dt \\
&\leq \int_0^\infty \left(\frac{\alpha t^3}{\sqrt{1-\alpha^2}} + e t^2\right) e^{-\frac{t^2}{c(1-\alpha^2)}} dt \\
&= \frac{\alpha}{\sqrt{1-\alpha^2}} \int_0^\infty t^3 e^{-\frac{t^2}{c(1-\alpha^2)}} dt + e \int_0^\infty t^2 e^{-\frac{t^2}{c(1-\alpha^2)}} dt \\
&\overset{(\zeta_2)}{\leq} c\left(1-\alpha^2\right)^{\frac{3}{2}} \overset{(\zeta_3)}{\leq} \delta\sqrt{1-\alpha^2}
\end{aligned}$$

where (ζ_1) follows from (11), (ζ_2) follows from the formulae for second and third absolute moments of gaussian random variables and (ζ_3) follows from the fact that $1-\alpha^2 < \delta$. Plugging the above inequality in (10), we obtain:

$$|\mathbb{E}[U]| \leq (1+2\delta)\left(\frac{\sqrt{1-\alpha^2}}{\alpha}\right) + \delta\sqrt{1-\alpha^2} \leq (1+4\delta)\sqrt{1-\alpha^2},$$

where we used the fact that $\alpha \geq 1 - \frac{\delta}{2}$. This proves the lemma. □

Lemma B.1.7. *Assume the hypothesis of Theorem 3.5.2 and the notation therein. Then,*

$$\left|{e_3}^T A\left(D-I\right)A^T e_1\right| \leq \delta m\sqrt{1-\alpha^2},$$

with probability greater than $1 - \frac{\eta}{10}e^{-n}$.

Proof. The proof of this lemma is very similar to that of Lemma B.1.5. We have:

$$\begin{aligned}
{e_3}^T A\left(D-I\right)A^T e_1 &= \sum_{l=1}^m \overline{a}_{1l} a_{3l} \left(\mathrm{Ph}\left(\left(\alpha\overline{a}_{1l} + \overline{a}_{2l}\sqrt{1-\alpha^2}\overline{a}_{3l}\right) a_{1l}\right) - 1\right) \\
&= \sum_{l=1}^m |a_{1l}|\, a'_{3l} \left(\mathrm{Ph}\left(\alpha\,|a_{1l}| + \overline{a'_{2l}}\sqrt{1-\alpha^2}\right) - 1\right),
\end{aligned}$$

where $a'_{3l} \stackrel{\text{def}}{=} a_{3l}\text{Ph}\left(\overline{a}_{1l}\right)$ is identically distributed to a_{3l} and is independent of $|a_{1l}|$ and a'_{2l}. Define the random variable U_l as:

$$U_l \stackrel{\text{def}}{=} |a_{1l}|\, a'_{3l} \left(\text{Ph}\left(1 + \frac{\overline{a'_{2l}}\sqrt{1-\alpha^2}}{\alpha\,|a_{1l}|}\right) - 1\right).$$

Since a'_{3l} has mean zero and is independent of everything else, we have:

$$\mathbb{E}\left[U_l\right] = 0.$$

Similar to Lemma B.1.5, we will calculate $\mathbb{P}\left[U_l > t\right]$ to show that U_l is subexponential and use it to derive concentration bounds.

$$\begin{aligned}\mathbb{P}\left[|U_l| > t\right] &\leq \mathbb{P}\left[|a_{1l}|\,|a'_{3l}|\,\frac{c\sqrt{1-\alpha^2}\,|a'_{2l}|}{\alpha\,|a_{1l}|} > t\right] \\ &= \mathbb{P}\left[|a'_{2l}a'_{3l}| > \frac{c\alpha t}{\sqrt{1-\alpha^2}}\right] \leq \exp\left(1 - \frac{c\alpha t}{\sqrt{1-\alpha^2}}\right),\end{aligned}$$

where the last step follows from the fact that a'_{2l} and a'_{3l} are independent subgaussian random variables and hence $|a'_{2l}a'_{3l}|$ is a subexponential random variable. Using Proposition 5.16 from [89], we obtain:

$$\begin{aligned}&\mathbb{P}\left[\left|\sum_{l=1}^{m} U_l - \mathbb{E}\left[U_l\right]\right| > \delta m\sqrt{1-\alpha^2}\right] \\ &\leq 2\exp\left(-\min\left(\frac{c\delta^2 m^2\left(1-\alpha^2\right)}{\left(1-\alpha^2\right)m}, \frac{c\delta m\sqrt{1-\alpha^2}}{\sqrt{1-\alpha^2}}\right)\right) \\ &\leq 2\exp\left(-c\delta^2 m\right) \leq \frac{\eta}{10}\exp\left(-n\right).\end{aligned}$$

Hence, we have:

$$\left|{e_3}^T A\left(D-I\right)A^T e_1\right| = \left|\sum_{l=1}^{m} U_l\right| \leq \delta m\sqrt{1-\alpha^2},$$

with probability greater than $1 - \frac{\eta}{10}\exp(-n)$. This proves the lemma. □

Lemma B.1.8. *For every $w \in \mathfrak{C}$, we have:*

$$\left|\text{Ph}\left(1+w\right) - 1\right| \leq 2\left|w\right|.$$

Proof. The proof is straight forward:

$$|\mathrm{Ph}\,(1+w)-1| \leq |\mathrm{Ph}\,(1+w)-(1+w)| + |w| = |1-|1+w|| + |w| \leq 2\,|w|\,.$$

□

B.2 Proofs for Section 3.6

Proof of Lemma 3.6.1. For every $j \in [n]$ and $i \in [m]$, consider the random variable $Z_{ij} \stackrel{\text{def}}{=} |a_{ij} y_i|$. We have the following:

- if $j \in S$, then

$$\begin{aligned}\mathbb{E}\,[Z_{ij}] &= \frac{2}{\pi}\left(\sqrt{1-\left(x_j^*\right)^2} + x_j^* \arcsin x_j^*\right) \\ &\geq \frac{2}{\pi}\left(1 - \frac{5}{6}\left(x_j^*\right)^2 - \frac{1}{6}\left(x_j^*\right)^4 + x_j^*\left(x_j^* + \frac{1}{6}\left(x_j^*\right)^3\right)\right) \\ &\geq \frac{2}{\pi} + \frac{1}{6}\left(x_{\min}^*\right)^2,\end{aligned}$$

  where the first step follows from Corollary 3.1 in [59] and the second step follows from the Taylor series expansions of $\sqrt{1-x^2}$ and $\arcsin(x)$,

- if $j \notin S$, then $\mathbb{E}\,[Z_{ij}] = \mathbb{E}\,[|a_{ij}|]\,\mathbb{E}\,[|y_i|] = \frac{2}{\pi}$ and finally,

- for every $j \in [n]$, Z_{ij} is a sub-exponential random variable with parameter $c = O(1)$ (since it is a product of two standard normal random variables).

Using the hypothesis of the theorem about m, we have:

- for any $j \in S$, $\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m Z_{ij} - \left(\frac{2}{\pi} + \frac{1}{12}\left(x_{\min}^*\right)^2\right) < 0\right] \leq \exp\left(-c\left(x_{\min}^*\right)^4 m\right) \leq \delta n^{-c}$, and

- for any $j \notin S$, $\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m Z_{ij} - \left(\frac{2}{\pi} + \frac{1}{12}\left(x_{\min}^*\right)^2\right) > 0\right] \leq \exp\left(-c\left(x_{\min}^*\right)^4 m\right) \leq \delta n^{-c}$.

Applying a union bound to the above, we see that with probability greater than $1-\delta$, there is a separation in the values of $\frac{1}{m}\sum_{i=1}^m Z_{ij}$ for $j \in S$ and $j \notin S$. This proves the theorem. □

Appendix C

Proofs for Learning Sparsely used Dictionaries using Alternating Minimization

C.1 Proofs of the main theorems

We first present the proof of Theorem 4.3.2. All the required lemmas for the proof of this theorem can be found in Appendix C.2.

Proof of Theorem 4.3.2:

Consider a particular iteration of Algorithm 8. Procedure 1 returns Uniq-intersect(Y_{i^*}, Y_{j^*}) with probability greater than $1 - 2\exp(-\gamma^2 \left|\widehat{S}\right| /2)$. If $\neg$ Uniq-intersect(Y_{i^*}, Y_{j^*}), then Algorithm 8 proceeds to the next iteration. Consider the case of Uniq-intersect(Y_{i^*}, Y_{j^*}) and suppose $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_l\}$. Using Proposition 4.3.5, with probability greater than $1 - d\exp\left(-c\alpha^2 \left|\widehat{S}\right|\right)$, we have:

$$\|A^*{}_l - \widehat{a}\|_2^2 < 32sM^2 \left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}}\right).$$

Using Lemma C.2.5 and Lemma C.2.1, we see that $\left|\widehat{S}\right| \geq \frac{ns}{4r}$ with probability greater than $1 - \exp\left(\frac{-ns}{16r}\right)$. Using a union bound over all the iterations (which are at most n^2), the above claims hold for all iterations with probability greater than $1 - n^2 d\exp\left(\frac{-c\alpha^2 ns}{r}\right) - 2n^2\exp\left(\frac{-\gamma^2 ns}{8r}\right) - n^2\exp\left(\frac{-ns}{16r}\right)$.

Using Lemma C.2.5 and Lemma C.2.1, with probability greater than $1 - r\exp\left(\frac{-ns}{64r}\right)$, for every $l \in [r]$, there are at least $\frac{ns}{8r}$ pairs (i^*, j^*) such that $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_l\}$ and $(i^*, j^*) \in G_{\mathrm{corr}(\rho)}$. Lines 9-11 of the algorithm then ensure that there is a unique copy of the approximation to $A^*{}_l$ dictionary element. Using a union bound now gives the result.

□

We now prove our second main result - Theorem 4.3.3. All the auxiliary lemmas and definitions we make use of in the proof of this theorem can be found in Appendix C.3.

In order to keep the notation less cumbersome, we will track the progress made in one iteration of Algorithm 2. For any iteration t, we denote $A(t)$ as $\widetilde{A}$ and $A(t+1)$ as A. Similarly we denote $X(t)$ and $X(t+1)$ as $\widetilde{X}$ and X respectively. Then the goal will be to show that A is closer to A^* than $\widetilde{A}$.

In order to establish Theorem 4.3.3, it suffices to establish a recurrence relation of the form

$$dist\left(A, A^*\right) < c \cdot dist\left(\widetilde{A}, A^*\right),$$

for some $c < 1$.

Proof of Theorem 4.3.3: As an induction hypothesis, we have $dist\left(\widetilde{A}, A^*\right) < \epsilon_t$. We will show that for every $p \in [r]$, we will have:

$$dist\left(A_p, {A^*}_p\right) \leq \frac{23616\mu_1 s^3}{\sqrt{d}}\epsilon_t < \epsilon_{t+1}. \tag{1}$$

This suffices to prove the theorem by appealing to Lemma C.3.1.

Now fix any $w \perp {A^*}_p$ such that $\|w\|_2 = 1$. We first provide a bound on $\langle w, A_p \rangle$. We have with high probability

$$\begin{aligned}
\langle w, A_p \rangle = w^\top A^* X^* X^+{}_p &\overset{(\zeta_1)}{\leq} \left\|w^\top A^*\right\|_2 \left\|\left(X^* X^+\right)_p^{\setminus p}\right\|_2 \\
&\overset{(\zeta_2)}{\leq} \mu_1\sqrt{\frac{r}{d}} \cdot \frac{1968 s^2 \left\|\triangle X\right\|_\infty}{\sqrt{r}} \\
&= \frac{17712\mu_1 s^3}{\sqrt{d}}\epsilon_t,
\end{aligned} \tag{2}$$

where (ζ_1) follows from the fact that $w^\top {A^*}_p = 0$ and (ζ_2) follows from Assumption $(A2)$ and Lemma 4.3.8.

In order to bound $dist\,(A, A^*)$, it remains to show a lower bound on $\|A\|_2$. This is again just algebraic given our main lemmas.

$$\begin{aligned}\|A_p\|_2 = \left\|A^* X^* X^+{}_p\right\|_2 &= \left\|A^*\left(X - \triangle X\right) X^+{}_p\right\|_2 \\ &\overset{(\zeta_1)}{=} \left\|A^*{}_p - A^* \triangle X X^+{}_p\right\|_2 \\ &\geq \|A^*{}_p\|_2 - \left\|A^*\left(\triangle X X^+\right)_p\right\|_2,\end{aligned}$$

where (ζ_1) follows from the fact that $XX^+ = \mathbb{I}$. We decompose the second term into diagonal and off-diagonal terms of $\triangle X X^+$, followed by triangle inequality and obtain

$$\begin{aligned}\|A_p\|_2 &\geq 1 - \left\|A^*{}_p\left(\triangle X X^+\right)_p^p + A^*{}_{\backslash p}\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - \|A^*{}_p\|_2 \left|\left(\triangle X X^+\right)_p^p\right| - \left\|A^*{}_{\backslash p}\right\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - 1 \cdot \left\|\triangle X X^\top \left(X X^\top\right)^{-1}\right\|_2 - \left\|A^*{}_{\backslash p}\right\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - \underbrace{\|\triangle X\|_2 \left\|X^\top\right\|_2 \left\|\left(X X^\top\right)^{-1}\right\|_2}_{\mathcal{T}_1} - \underbrace{\left\|A^*{}_{\backslash p}\right\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2}_{\mathcal{T}_2}\end{aligned}$$

It remains to control $\mathcal{T}_1$ and $\mathcal{T}_2$ at an appropriate level. We start from $\mathcal{T}_1$. Note that $\|\triangle X\|_2$ is bounded by Lemmas 4.3.6 and 4.3.7, while $\left\|X^\top\right\|_2$ is controlled by Lemma C.3.6 (recall $\|\triangle X\|_\infty \leq 1/(64s)$). Invoking Lemma C.3.6 to control $\left\|XX^{\top -1}\right\|_2$, we obtain the following bound on $\mathcal{T}_1$ with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$

$$\mathcal{T}_1 \leq 18\epsilon_t s^2 \sqrt{\frac{n}{r}} \cdot 3s\sqrt{\frac{n}{r}} \cdot \frac{8r}{sn} = 432 s^2 \epsilon_t.$$

The second term $\mathcal{T}_2$ is directly controlled by Lemma 4.3.8, yielding with probability at least
$1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$

$$\mathcal{T}_2 \leq \mu_1 \sqrt{\frac{r}{d}} \frac{1968 s^3 \epsilon_t}{\sqrt{r}}$$

Putting all the terms together, we obtain that

$$\|A_p\|_2 \geq 1 - 9s^2 \left(48 + \frac{1968s\mu_1}{\sqrt{d}}\right)\epsilon_t \geq \frac{3}{4}. \tag{3}$$

Combining the bounds (2) and (3) yields the desired recursion (1). Appealing to Lemma C.3.1 along with our setting of ϵ_t (3) completes the proof of the claim (1). Finally note that the error probability in the theorem is obtained by using the fact that $M \geq 1$, and that the failure probability is purely incurred from the structure of the non-zero entries of X^*, so that it is incurred only once and not at each round. This avoids the need of a union bound over all the rounds, yielding the result. □

C.2 Proofs for initialization

In this section we will present the proof of Theorem 4.3.2, which is our main result for Algorithm 8. We will start by presenting a host of useful lemmas, and sketch out how they fit together to yield the main results before moving on to the proofs.

C.2.1 Correlation graph properties

In this section we will present some useful properties of the correlation graph $G_{\text{corr}(\rho)}$ described in Section 4.3.4. Recall that $G_{\text{corr}(\rho)}$, where the nodes are samples $\{Y_1, Y_2, \dots Y_n\}$ and an edge $(Y_i, Y_j) \in G_{\text{corr}(\rho)}$ implies that $|\langle Y_i, Y_j \rangle| > \rho$, for some $\rho > 0$. This is employed by Algorithm 8 as a proxy for identifying samples which have common dictionary elements. We now make this connection concrete in the next few lemmas. For this we also recall our notation $\mathcal{N}_B(y)$ which is the neighborhood of a sample y in the coefficient bipartite graph (see Figure 4.2), that is, the set of dictionary elements that combine to yield y.

Lemma C.2.1 (Correlation graph)**.** *Under the incoherence assumption* (*A*1) *and the threshold ρ in the hypothesis of Theorem 4.3.2, the following is true for the edges in the correlation graph $G_{\text{corr}(\rho)}$:*

$$|\,\mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l)| = 1 \Rightarrow (Y_k, Y_l) \in G_{\mathrm{corr}(\rho)}, \quad \forall\, i \in [r], \tag{4}$$

$$(Y_k, Y_l) \in G_{\mathrm{corr}(\rho)} \Rightarrow |\,\mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l)| \geq 1, \tag{5}$$

for all $k, l \in \{1, 2, , \ldots, n\}, k \neq l$.

Lemma C.2.1 suggests that nodes which intersect in *exactly one* dictionary element are special, in that they are guaranteed to have an edge between them in $G_{\mathrm{corr}(\rho)}$. Our next lemma works towards establishing something even stronger. We will next establish that there are large cliques in the correlation graph where any two samples in the clique intersect in the same unique dictionary element. In order to state the lemma, we need some additional notation.

For each dictionary element $A^*{}_i$, consider a set of samples[1] $\{Y_k, k \in S\}$, for some $S \subset \{1, 2, \ldots, n\}$, such that they only have $A^*{}_i$ in common, and denote such a set by $\mathcal{C}_i$ i.e.

$$\mathcal{C}_i := \{Y_k, k \in S : \mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l) = \{A^*{}_i, \}\ \forall\, k, l \in S\}. \tag{6}$$

Lemma C.2.1 implies that in the correlation graph, the set of nodes in $\mathcal{C}_i$ form a clique (not necessarily maximal), for each $i \in \{1, 2, \ldots, r\}$, as shown in Figure 4.1. The above implication can be exploited for recovery of dictionary elements: if we find the set $\mathcal{C}_i$, then we can hope to recover the element $A^*{}_i$, since that is the only element in common to the samples in $\mathcal{C}_i$.

For ease of stating the next lemma, we further define two shorthand notations.

$$\text{Uniq-intersect}(Y_i, Y_j) := \{(Y_i, Y_j) \in G_{\mathrm{corr}(\rho)} \quad \text{and} \quad |\,\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j)| = 1\}, \tag{7}$$

[1]Note that such a set need not be unique.

Intuitively, the samples satisfying Uniq-intersect(Y_i, Y_j) are guaranteed to have an edge between them by Lemma C.2.1. In order to guarantee large cliques, we will also need to measure the number of triangles in $G_{\text{corr}(\rho)}$.

In order to do this, given anchor samples Y_{i^*} and Y_{j^*} have a unique intersection, we now bound the probability that a randomly chosen sample Y_i, among the neighborhood set of Y_{i^*} and Y_{j^*} in the correlation graph also has a unique intersection. Now define unique intersection event for a new sample Y_i with respect to anchor samples Y_{i^*} and Y_{j^*} as follows

$$\text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) := \{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{i^*}) = \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_k\}\}, \tag{8}$$

where $\{A^*{}_k\} = \mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})$ is the unique intersection of the anchor samples Y_{i^*} and Y_{j^*}. In other words, Uniq-intersect$(Y_i; Y_{i^*}, Y_{j^*})$ indicates the event that the pairwise intersections of the new sample Y_i with each of the anchors Y_{i^*} and Y_{j^*} is unique and equal to the unique intersection of Y_{i^*} and Y_{j^*}.

Lemma C.2.2 (Formation of clique under good anchor samples)**.**

$$\begin{aligned}&\mathbb{P}\left[\text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*}),\ \textit{and} \right. \\ &\qquad\qquad \left. (Y_i, Y_{i^*}), (Y_i, Y_{j^*}) \in G_{\text{corr}(\rho)}\right] \\ &\geq 1 - \frac{s^3}{r}.\end{aligned}$$

Lemma C.2.2 is crucial for our algorithm. It guarantees that given a pair of good anchor elements—one satisfying unique intersection property—a large fraction of their neighbors also contain this common dictionary element. Some further arguments can then be made to establish that a large fraction of the neighbors of Y_{i^*} and Y_{j^*} also have edges amongst themselves and hence form cliques as defined in Equation 6.

C.2.2 Correctness of Procedure 1

A key component in our analysis is the correctness of Procedure 1. As we saw in the previous lemmas, it is crucial for a chosen pair of anchor elements to have a unique intersection in order to use them for identifying large cliques $\mathcal{C}_i$ in $G_{\text{corr}(\rho)}$. Procedure 1 plays a crucial role by providing a verifiable test for whether a pair of anchor elements have a unique intersection or not. Our next two lemmas help us establish that this test is sound with high probability. We first show that two neighbors of a bad anchor pair do not have an edge amongst them with high probability.

Denote the event

$$\Delta(Y_i, Y_j, Y_k) := \{(Y_i, Y_j), (Y_j, Y_k), (Y_i, Y_k) \in G_{\text{corr}(\rho)}\},$$

i.e., the samples Y_i, Y_j, Y_k form a triangle in the correlation graph.

Lemma C.2.3 (Detection of bad anchor samples)**.** *For randomly chosen samples Y_i, Y_j*

$$\mathbb{P}\left[(Y_i, Y_j) \notin G_{\text{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] > \frac{1}{16}.$$

Intuitively, this means that the number of sets S_i which will be edges in $G_{\text{corr}(\rho)}$ is rather small for an anchor pair with multiple dictionary elements in common. In order for correctness of the procedure, we will in fact need this number to be substantially smaller than that for a good anchor pair. This is indeed the case as we next establish.

Lemma C.2.4 (Detection of good anchor samples)**.** *For randomly chosen samples Y_i, Y_j*

$$\mathbb{P}\left[(Y_i, Y_j) \notin G_{\text{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \leq \frac{24s^3}{r}.$$

Combining the above two lemmas, the correctness of Procedure 1 naturally follows. In particular, we note that the above two lemmas prove Proposition 4.3.4.

Proposition 4.3.4 (Correctness of Procedure 1)**.** Suppose $(Y_{i^*}, Y_{j^*}) \in G_{\text{corr}(\rho)}$. Suppose that $s^3 \leq r/1536$ and $\gamma \leq 1/64$. Then Algorithm 8 returns the value of Uniq-intersect(Y_{i^*}, Y_{j^*}) correctly with probability greater than $1 - 2\exp(-\gamma^2 m)$.

C.2.3 Estimation of the Dictionary Elements via SVD

In this section we will put all the pieces together and establish Theorem 4.3.2. We start by establishing that given a pair of good anchor elements, the SVD step in Algorithm 8 approximately recovers the unique dictionary element in the intersection of the two anchors. In this context, we recall Proposition 4.3.5.

Proposition 4.3.5 (Accuracy of SVD)**.** Consider anchor samples Y_{i^*} and Y_{j^*} such that Uniq-intersect(Y_{i^*}, Y_{j^*}) is satisfied, and wlog, let $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_1\}$. Recall the definition of $\widehat{S}$ (16), and further define $\widehat{Q} := \sum_{i \in \widehat{S}} Y_i Y_i^\top$ and $|\widehat{S}| = m$. If $\widehat{a}$ is the top singular vector of $\widehat{Q}$, then there exists a universal constant c such that we have:

$$\min_{z \in \{-1,1\}} \|\widehat{a} - zA^*{}_1\|_2^2 < 32sM^2 \left(\frac{\mu_1}{\sqrt{ds}} + \frac{\mu_1^2}{d} + \frac{s^3}{r} + \alpha^2 + \frac{\alpha}{\sqrt{s}} \right),$$

with probability greater than $1 - d\exp\left(-c\alpha^2 m\right)$ for $\alpha < 1/20$.

The key missing piece from using Proposition 4.3.5 to prove Theorem 4.3.2 is the dependence on the random quantity $|\widehat{S}|$ in the error probability in Proposition 4.3.5. The following lemma bounds the size of this set.

Lemma C.2.5. *In each iteration of Algorithm 8, the size of the set $\widehat{S}$ satisfies:*

$$\left|\widehat{S}\right| \geq \frac{ns}{4r},$$

with probability greater than $1 - \exp\left(\frac{-ns}{16r}\right)$.

C.2.4 Proofs of correlation graph properties

We start by proving Lemmas C.2.1 and C.2.2 in Section C.2.1.

Proof of Lemma C.2.1:

We first prove (5) via contradiction. Suppose $\mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l) = \emptyset$, we then have

$$
\begin{aligned}
|\langle Y_k, Y_l \rangle| &= |\sum_{i,j} {X^*}^i_k {X^*}^j_l \langle {A^*}_i, {A^*}_j \rangle| \leq \sum_{i,j} |{X^*}^i_k {X^*}^j_l \langle {A^*}_i, {A^*}_j \rangle| \\
&\leq |\mathcal{N}_B(Y_k)| \cdot |\mathcal{N}_B(Y_l)| \cdot \max_{i,j,k,l} |{X^*}^i_k {X^*}^j_l| \cdot \max_{i \neq j} |\langle {A^*}_i, {A^*}_j \rangle| \leq \frac{s^2 \mu_0}{\sqrt{d}} M^2
\end{aligned}
$$

For (4), let $\{{A^*}_{i^*}\} = \mathcal{N}_B(Y_k) \cap \mathcal{N}_B(Y_l)$

$$
\begin{aligned}
|\langle Y_k, Y_l \rangle| &= |\sum_{i,j} {X^*}^i_k {X^*}^j_l \langle {A^*}_i, {A^*}_j \rangle| \\
&\geq |{X^*}^{i^*}_k {X^*}^{i^*}_l| \langle {A^*}_{i^*}, {A^*}_{i^*} \rangle - \sum_{i \neq j} |{X^*}^i_k {X^*}^j_l \langle {A^*}_i, {A^*}_j \rangle| \\
&\geq 1 - \frac{s^2 \mu_0}{\sqrt{d}} M^2,
\end{aligned}
$$

using the above analysis. The claims now follow from the setting of ρ. $\square$

We next establish Lemma C.2.2.

Proof of Lemma C.2.2: Define the event

$$
\mathcal{A} := \{|\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{i^*})| \geq 1\} \cap \{|\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{j^*})| \geq 1\}.
$$

From Lemma C.2.1, we have that

$$
\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*}),\ \text{and} \right. \\
&\qquad\qquad \left. (Y_i, Y_{i^*}), (Y_i, Y_{j^*}) \in G_{\text{corr}(\rho)}\right] \\
&\geq \mathbb{P}\left[\text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*}), \mathcal{A}\right]
\end{aligned}
$$

In order to lower bound $\mathbb{P}\left[\text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*}), \mathcal{A}\right]$, we instead upper bound the probability of the complementary event $\mathbb{P}\left[\neg\, \text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*}), \mathcal{A}\right]$

In order to do so, we first bound the following

$$\mathbb{P}\left[\mathcal{A} \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \geq \frac{s}{r}, \tag{9}$$

since $\mathcal{A}$ holds when the unique element in $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})$ is chosen and its probability is s/r. We also have

$$\mathbb{P}\left[\neg\, \text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*}) \cap \mathcal{A} \,\middle|\, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \leq \frac{(s-1)^2\binom{r-3}{s-2}}{\binom{r}{s}},$$

since for $\neg\, \text{Uniq-intersect}(Y_i; Y_{i^*}, Y_{j^*})$ to hold, we need to choose at least one of the $s-1$ elements in $\mathcal{N}_B(Y_{i^*}) / \mathcal{N}_B(Y_{j^*})$, and similarly one from the $s-1$ elements of $\mathcal{N}_B(Y_{j^*}) / \mathcal{N}_B(Y_{i^*})$. The rest of the $s-2$ elements can be picked arbitrarily from the $r-3$ dictionary atoms that remain after excluding the two already picked and the unique intersection $\mathcal{N}_B(Y_{j^*}) \cap \mathcal{N}_B(Y_{i^*})$.

It is easy to check that

$$\begin{aligned}\frac{(s-1)^2\binom{r-3}{s-2}}{\binom{r}{s}} &= \frac{(s-1)^2(r-s)s(s-1)}{r(r-1)(r-2)} \\ &\leq \frac{s^4}{r^2}.\end{aligned} \tag{10}$$

Taking the ratio of the two bounds in (9) and (10) completes the proof. □

C.2.5 Proofs of Lemmas C.2.3 and C.2.4

We now prove the two lemmas that are crucial to establishing the correctness of Procedure 1.

Proof of Lemma C.2.3: Let $\mathcal{A}_1$ and $\mathcal{A}_2$ denote the following events:

$$\begin{aligned}\mathcal{A}_1 :=& \{|\, \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{i^*})| \geq 1\} \cap \{|\, \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{j^*})| \geq 1\} \\ & \cap \{|\, \mathcal{N}_B(Y_j) \cap \mathcal{N}_B(Y_{i^*})| \geq 1\} \cap \{|\, \mathcal{N}_B(Y_j) \cap \mathcal{N}_B(Y_{j^*})| \geq 1\} \\ \mathcal{A}_2 :=& \{|\, \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{i^*})| = 1\} \cap \{|\, \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{j^*})| = 1\} \\ & \cap \{|\, \mathcal{N}_B(Y_j) \cap \mathcal{N}_B(Y_{i^*})| = 1\} \cap \{|\, \mathcal{N}_B(Y_j) \cap \mathcal{N}_B(Y_{j^*})| = 1\}\end{aligned} \tag{11}$$

In words, both Y_i and Y_j have at least dictionary element in common with each of Y_{i^*} and Y_{j^*} under the event $\mathcal{A}_1$, while the number of common elements is *exactly one* under the event $\mathcal{A}_2$. We have

$$
\begin{aligned}
&\mathbb{P}\left[(Y_i, Y_j) \notin G_{\mathrm{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\overset{(a)}{=} \mathbb{P}\left[(Y_i, Y_j) \notin G_{\mathrm{corr}(\rho)} \mid \mathcal{A}_1, \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}),\right. \\
&\qquad\qquad \left.\neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&= \mathbb{P}\left[(Y_i, Y_j) \notin G_{\mathrm{corr}(\rho)}, \Delta(Y_j, Y_{i^*}, Y_{j^*}) \mid \mathcal{A}_1, \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}),\right. \\
&\qquad\qquad\qquad \left.\neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \mathbb{P}\left[(Y_i, Y_j) \notin G_{\mathrm{corr}(\rho)}, \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}) \mid \mathcal{A}_1,\right. \\
&\qquad\qquad \left.\neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*}), (Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}\right] \\
&\overset{(b)}{\geq} \mathbb{P}\left[(Y_i, Y_j) \notin G_{\mathrm{corr}(\rho)}, \mathcal{A}_2 \mid \mathcal{A}_1, \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*}), (Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}\right] \\
&\overset{(c)}{\geq} \mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \mathcal{A}_1, \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*}),\right. \\
&\qquad\qquad \left.(Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}\right], \qquad (12)
\end{aligned}
$$

where the inequalities (a), (b) and (c) follow from Lemma C.2.1. We will now work on lower bounding this resulting probability.

We first lower bound the numerator in writing the above conditional probability as the ratio of a joint to marginal probability. We begin by noting that

$$
\begin{aligned}
&\mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*}),\right. \\
&\qquad\qquad \left.(Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)})\right] \\
&= \mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*}),\right. \\
&\qquad\qquad \left.(Y_{i^*}, Y_{j^*}) \in G_{\mathrm{corr}(\rho)}\right]
\end{aligned}
$$

Let us define $m = |\,\mathcal{N}_B(Y_{i^*}) \cup \mathcal{N}_B(Y_{j^*})| \in [s, 2s]$ and $l = |\,\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})| \geq 2$[2]. The event in the probability above, that is $\mathcal{A}_2$ holds while Y_i

[2]the intersection is at least 1 by Lemma C.2.1

and Y_j do not share a dictionary element, can be arranged by choosing two of the l elements, and assigning a unique element to each Y_i and Y_j. Similarly the remaining elements can be chosen outside $\mathcal{N}_B(Y_{i^*}) \cup \mathcal{N}_B(Y_{j^*})$ in a non-overlapping manner: for Y_i assign $s-1$ elements among $r-m$ elements, and then for Y_j assign from remaining $r-m-s+1$ elements. This logic yields the following lower bound on the probability

$$\begin{aligned}
&\mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \frac{2\binom{l}{2}\binom{r-m}{s-1}\binom{r-m-s+1}{s-1}}{\binom{r}{s}^2} \geq \frac{2\binom{l}{2}\binom{r-2s}{s-1}\binom{r-3s+1}{s-1}}{\binom{r}{s}^2},
\end{aligned}$$

where the second inequality uses $m \leq 2s$. Now with some straightforward algebra, we can further lower bound this expression as

$$\begin{aligned}
&\mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{3s-3}{r-s}\right)^{s-1}\left(1 - \frac{2s-1}{r-s}\right)^{s-1} \\
&\geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{3s}{r-s}\right)^{s}\left(1 - \frac{2s}{r-s}\right)^{s}.
\end{aligned}$$

Now we invoke Lemma C.2.10 to further lower bound the RHS and obtain

$$\begin{aligned}
&\mathbb{P}\left[\{\mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_j) = \emptyset\} \cap \mathcal{A}_2 \mid \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \frac{s^2(l-1)^2}{r^2}\exp\left(-\frac{3s^2}{r-s}\right)\exp\left(-\frac{2s^2}{r-s}\right) \geq \frac{s^2(l-1)^2}{r^2}\left(1 - \frac{10s^2}{r-s}\right) \\
&\geq \frac{s^2(l-1)^2}{2r^2},
\end{aligned}$$

where the final inequality holds since $s^2 \leq r/40$.

In order to lower bound the conditional probability in Equation 12, we need to further upper bound the marginal probability in the denominator. To this end, we observe that we have to upper bound $\mathbb{P}\left[\mathcal{A}_1 | \neg\,\text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right]$.

Now conditioned on $\neg$ Uniq-intersect(Y_{i^*}, Y_{j^*}), for each Y_i and Y_j, $\mathcal{A}_1$ can be satisfied in two ways: choose at least one element from l elements in $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})$ or choose at least two elements from $m - l$ elements in $\mathcal{N}_B(Y_{i^*}) \cup \mathcal{N}_B(Y_{j^*})$. Making this precise, we obtain

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_1 \mid \neg \text{ Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] &\leq \left(\frac{ls}{r} + \frac{(m-l)^2\binom{r-2}{s-2}}{\binom{r}{s}}\right)^2 \\
&\leq \left(\frac{ls}{r} + \frac{s^2(m-l)^2}{(r-1)^2}\right)^2 \\
&\leq \left(\frac{ls}{r} + \frac{s^2(2s-2)^2}{(r-1)^2}\right)^2 \\
&\leq \frac{2l^2s^2}{r^2}, \text{ (since } 4s^3 < r-1)
\end{aligned}$$

The result follows by using the fact that $l \geq 2$. □

The proof of Lemma C.2.4 is similar, but involves controlling slightly different events.

Proof of Lemma C.2.4:

We will establish the lemma by lower bounding the probability of the complementary event. We recall the events $\mathcal{A}_1$ and $\mathcal{A}_2$ defined in Equation 11 in the proof of Lemma C.2.3. We can mimick the initial arguments in the proof of Lemma C.2.3 to conclude that

$$\begin{aligned}
&\mathbb{P}\left[(Y_i, Y_j) \in G_{\text{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \mathbb{P}\left[\text{Uniq-intersect}(Y_i, Y_j) \cap \mathcal{A}_2 \mid \mathcal{A}_1, \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right],
\end{aligned}$$

and we provide a lower bound for this. Once again, we express the conditional probability as the ratio of a joint to a marginal and then lower bound the numerator and upper bound the denominator. In the numerator, we have the event

We have

$$\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(Y_i, Y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&= \mathbb{P}\left[\text{Uniq-intersect}(Y_i, Y_j) \cap \mathcal{A}_2 \mid \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right]
\end{aligned}$$

The event Uniq-intersect$(Y_i, Y_j) \cap \mathcal{A}_2$ is guaranteed to occur if we choose Y_i and Y_j so that they have the only element in $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})$ in common. This yields the lower bound

$$\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(Y_i, Y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \frac{\binom{r-2s+1}{s-1}\binom{r-3s+2}{s-1}}{\binom{r}{s}^2}.
\end{aligned}$$

It is easy to further conclude that

$$\begin{aligned}
&\mathbb{P}\left[\text{Uniq-intersect}(Y_i, Y_j) \cap \mathcal{A}_2 \cap \mathcal{A}_1 \mid \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \frac{s^2}{r^2}\left(1 - \frac{3s-3}{r-s+1}\right)^{(s-1)}\left(1 - \frac{2s-2}{r-s+1}\right)^{s-1} \\
&\geq \frac{s^2}{r^2}\exp(-5(s-1)^2/(r-s+1)) \\
&\geq \frac{s^2}{r^2}\left(1 - \frac{10s^2}{r-s}\right) \geq \frac{s^2}{r^2}\left(1 - \frac{20s^2}{r}\right),
\end{aligned}$$

where we again invoked Lemma C.2.10 as well as the fact that $s \leq r/2$. As for the marginal probability in the denominator, we need to upper bound

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_1 \mid \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] &\leq \left(\frac{s}{r} + \frac{(2s-1)^2\binom{r-2}{s-2}}{\binom{r}{s}}\right)^2 \\
&\leq \left(\frac{s}{r} + \frac{(2s-1)^2(s-1)^2}{(r-1)^2}\right)^2 \leq \frac{s^2}{r^2}\left(1 + \frac{4s^3}{r}\right)^2,
\end{aligned}$$

since for each Y_i and Y_j, $\mathcal{A}_1$ can be satisfied in two ways: choose the unique element from $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})$ or choose at least two elements from $2s-1$ elements in $\mathcal{N}_B(Y_{i^*}) \cup \mathcal{N}_B(Y_{j^*})$.

Using the above two inequalities, we have:

$$\mathbb{P}\left[(Y_i, Y_j) \in G_{\mathrm{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \geq \frac{1 - \frac{20s^2}{r}}{\left(1 + \frac{4s^3}{r}\right)^2}.$$

It is easy to verify that $1/(1+x)^2 \leq 1 - x$ for $0 \leq x \leq (\sqrt{2}-1)/2$. Since $s^3 \leq r/5$, we obtain

$$\begin{aligned}
&\mathbb{P}\left[(Y_i, Y_j) \in G_{\mathrm{corr}(\rho)} \mid \Delta(Y_i, Y_{i^*}, Y_{j^*}), \Delta(Y_j, Y_{i^*}, Y_{j^*}), \text{Uniq-intersect}(Y_{i^*}, Y_{j^*})\right] \\
&\geq \left(1 - \frac{20s^2}{r}\right)\left(1 - \frac{4s^3}{r}\right) \\
&\geq 1 - \frac{24s^3}{r}.
\end{aligned}$$

□

C.2.6 Proof of Proposition 4.3.4

Let us start with the case when Uniq-intersect$(Y_{i^*}, Y_{j^*}) = 1$. For any pair (Y_i, Y_j) where Y_i and Y_j are taken from $\mathcal{N}_{G_{\mathrm{corr}(\rho)}}(Y_{i^*}) \cap \mathcal{N}_{G_{\mathrm{corr}(\rho)}}(Y_{j^*})$, let E_{ij} be the random variable which is 1 if $(Y_i, Y_j) \in G_{\mathrm{corr}(\rho)}$. Then Lemma C.2.4 guarantees $\mathbb{P}(E_{ij} = 1) \geq 1 - 24s^3/r$. Let

$$S := \{(i,j) \ : \ (Y_i, Y_j) \in G_{\mathrm{corr}(\rho)}, \text{ and } Y_i, Y_j \in \mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*})\}.$$

The size of the set constructed in Algorithm 1 is equal to $\sum_{(i,j)\in S} E_{ij}$. Recalling that $|S| = m$, Hoeffding's inequality guarantees that with probability at least $1 - 2\exp(-2m\gamma^2)$

$$\left|\frac{1}{m}\sum_{(i,j)\in S}(E_{ij} - \mathbb{P}(E_{ij} = 1))\right| \leq \gamma.$$

Combining with the lower bound on $\mathbb{P}(E_{ij} = 1)$, we obtain that with probability at least $1 - 2\exp(-2m\gamma^2)$,

$$\sum_{(i,j)\in S} E_{ij} \geq m\left(1 - 24\frac{s^3}{r}\right) - m\gamma. \tag{13}$$

Using $\gamma \leq 1/64$, we see that this quantity is at least $62m/64$ under the conditions of the lemma, which means that Algorithm 1 returns 1.

Now let us consider the case when Uniq-intersect$(Y_{i^*}, Y_{j^*}) = 0$. Defining E_{ij} the same way as above, we see that by Lemma C.2.3, $\mathbb{P}(E_{ij} = 1) \leq 15/16$. Then, a similar application of Hoeffding's inequality yields this time

$$\sum_{(i,j)\in S} E_{ij} \leq \frac{m}{16} + m\gamma, \tag{14}$$

which is at most $61m/64$ for $\gamma \leq 1/64$. Hence Algorithm 1 returns 0 in this case.

C.2.7 Proof of Proposition 4.3.5

We now prove Proposition 4.3.5. We need a couple of auxilliary results for the proof. We first restate a theorem from [88], which we will heavily use in the sequel.

Theorem C.2.6 (Restatement of Theorem 5.44 from [88])**.** *Consider a $d \times n$ matrix W where each column W_i of W is an independent random vector with covariance matrix Σ. Suppose further that $\|W_i\|_2 \leq \sqrt{u}$ a.s. for all i. Then for any $t \geq 0$, the following inequality holds with probability at least $1 - d\exp(-ct^2)$:*

$$\left\|\frac{1}{n}WW^T - \Sigma\right\|_2 \leq \max\left(\|\Sigma\|_2^{1/2}\,\delta, \delta^2\right) \text{ where } \delta = t\sqrt{\frac{u}{n}}.$$

Here $c > 0$ is an absolute numerical constant. In particular, this inequality yields:

$$\|W\|_2 \leq \|\Sigma\|_2^{\frac{1}{2}}\sqrt{n} + t\sqrt{u}.$$

In order to bound the errors made in Algorithm 8, we need some additional notation and auxilliary results. For now, let us consider a fixed pair of anchor samples Y_{i^*} and Y_{j^*} such that Uniq-intersect(Y_{i^*}, Y_{j^*}) is satisfied, and wlog, let $\mathcal{N}_B(Y_{i^*}) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_1\}$. We define the following sets of interest

$$\widehat{S} = \mathcal{N}_{\text{corr}}(Y_{i^*}) \cap \mathcal{N}_{\text{corr}}(Y_{j^*}),$$
$$S = \{Y_i \in \widehat{S} : \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{i^*}) = \mathcal{N}_B(Y_i) \cap \mathcal{N}_B(Y_{j^*}) = \{A^*{}_1\}\}, \text{ and} \quad (15)$$
$$\widetilde{S} = \widehat{S} \setminus S. \quad (16)$$

For the purposes of understanding the errors in Algorithm 8, it would be helpful to decompose each vector $Y_i \in S$ as

$$\breve{Y}_i := Y_i - X^{*1}{}_i A^*{}_1, \quad (17)$$

and accordingly define $\breve{Y}_S$ to be the $d \times |S|$ matrix of all such vectors in S. Intuitively, if all the vectors $\breve{y}$ were 0, then Algorithm 8 can recover $A^*{}_1$ via SVD in a relatively straightforward manner. We start by controlling the norm of the vectors Y_i and $\breve{Y}_i$.

Lemma C.2.7. *Given assumptions* $(B1)$ *and* $(B2)$*, we have for all* $i = 1, 2, \ldots, n$

$$\|Y_i\|_2 \leq M\sqrt{2s} \quad \textit{and} \quad \left\|\breve{Y}_i\right\|_2 \leq 2M\sqrt{s}.$$

Proof:

The proof is relatively straightforward consequence of our model and the assumptions. The model allows us to write

$$
\begin{aligned}
\|Y_i\|_2^2 = \langle Y_i, Y_i \rangle &= \sum_{A^*{}_p, A^*{}_q \in \mathcal{N}_B(Y_i)} X^{*p}_i X^{*q}_i \langle A^*{}_p, A^*{}_q \rangle \\
&\leq \sum_{A^*{}_p, A^*{}_q \in \mathcal{N}_B(Y_i)} |X^{*p}_i X^{*q}_i| \, |\langle A^*{}_p, A^*{}_q \rangle| \\
&= \sum_{A^*{}_p \in \mathcal{N}_B(Y_i)} (X^{*p}_i)^2 \, \|A^*{}_p\|_2^2 + \sum_{A^*{}_p \neq A^*{}_q \in \mathcal{N}_B(Y_i)} |X^{*p}_i X^{*q}_i| \, |\langle A^*{}_p, A^*{}_q \rangle| \\
&\leq M^2 \left(s + s^2 \frac{\mu_0}{\sqrt{d}} \right) \\
&\leq M^2 \left(s + \frac{1}{2} \right) \leq \frac{3sM^2}{2}.
\end{aligned}
$$

Finally, by triangle inequality we further have that $\left\| \breve{Y}_i \right\|_2 \leq \|Y_i\|_2 + M$.

□

Given this result, we would next like to control the amount of contribution the $\breve{Y}_i$ directions can have in the SVD step of Algorithm 8. Our next result shows that while these vectors are not zero, their random support along with the incoherence of our dictionary elements ensures that these vectors are not strongly aligned with any one direction. We do so by bounding the spectral norm of the matrix $\breve{Y}_S$.

Lemma C.2.8. *With the vectors $\breve{Y}_i$ defined in Equation 17, we have the following bound with probability greater than $1 - d \exp\left(-c\alpha^2 |S|\right)$ for any $\alpha > 0$*

$$
\left\| \breve{Y}_S \right\|_2 \leq M \sqrt{s\,|S|} \left(\frac{\mu_1}{\sqrt{d}} + 2\alpha \right),
$$

where c is a universal constant.

Proof:

In order to prove the lemma, we first calculate the spectral norm of the covariance matrix of $\breve{Y}_i$ and then use Theorem C.2.6. Note that from Lemma C.2.7, we have $\left\| \breve{Y}_i \right\|_2 \leq 2M\sqrt{s}$. We first bound the spectral norm of the

covariance matrix of $\breve{Y}_i \in S$ i.e., we bound $\left\|\mathbb{E}\left(\breve{Y}_i \breve{Y}_i^T\right)\right\|_2$. In order to do this, we first fix $w \in \mathbb{R}^d$ and calculate:

$$w^T \mathbb{E}\left[\breve{Y}_i \breve{Y}_i^T\right] w = \mathbb{E}\left[\left(w^T \breve{Y}_i\right)^2\right] = \mathbb{E}\left[\left(w^T A^* \breve{X}_i\right)^2\right] = \mathbb{E}\left[\left(z^T \breve{X}_i\right)^2\right],$$

where we use the notation $z := {A^*}^\top w$ and $\breve{X}_i$ is the same as ${X^*}_i$ but with ${X^*}_i^1$ set to 0. We further simplify the above as

$$\begin{aligned}
w^T \mathbb{E}\left[\breve{Y}_i \breve{Y}_i^T\right] w &\leq \mathbb{E}\left[\left(\sum_{p=1}^{r} z_p \breve{X}_i^p\right)^2\right] \\
&= \mathbb{E}\left[\sum_{p=1}^{r} z_p^2 \left(\breve{X}_i^p\right)^2\right] + \mathbb{E}\left[\sum_{p\neq q=1}^{r} z_p z_q \breve{X}_i^p \breve{X}_i^q\right] \\
&\leq \sum_{p=1}^{r} z_p^2 \mathbb{E}\left[\left(\breve{X}_i^p\right)^2\right] + \sum_{p\neq q=1}^{r} |z_p z_q| \left|\mathbb{E}\left[\breve{X}_i^p \breve{X}_i^q\right]\right| \\
&\leq \sum_{p=1}^{r} z_p^2 \frac{M^2 s}{r} + 0,
\end{aligned}$$

where the last inequality uses the fact that the values of $\mathbb{E}[{X^*}_i^p {X^*}_i^q] = 0$, since they are independent and zero mean.

Then we can further simplify the upper bound to obtain

$$w^T \mathbb{E}\left[\breve{Y}_i \breve{Y}_i^T\right] w \leq \frac{M^2 s}{r} \|z\|_2^2 \overset{(\zeta)}{\leq} \frac{M^2 s}{r} \cdot \frac{\mu_1^2 r}{d} = \frac{\mu_1^2 M^2 s}{d},$$

where (ζ) follows from Assumption $(A3)$, since

$$\|z\|_2 = \left\|{A^*}^\top w\right\|_2 \leq \left\|{A^*}^\top\right\|_2 \|w\|_2 = \left\|A^* {A^*}^\top\right\|_2^{\frac{1}{2}} \|w\|_2 \leq \sqrt{\frac{\mu_1^2 r}{d}}.$$

Recalling that w was an arbitrary unit vector, this immediately yields a spectral norm bound on the expected covariance

$$\left\| \mathbb{E}\left[\breve{Y}_i \breve{Y}_i^T \right] \right\|_2 \leq \frac{\mu_1^2 M^2 s}{d}.$$

We are now in a position to apply Theorem C.2.6 with the matrix $W = \breve{Y}_S$ of size $d \times |S|$, where $u = 2M\sqrt{s}$ and $t = \alpha\sqrt{|S|}$ for some $\alpha > 0$. Doing so yields the inequality

$$\begin{aligned} \left\| \breve{Y}_S \right\|_2 &\leq M \left(\sqrt{\frac{\mu_1^2 s}{kd}} \cdot \sqrt{|S|} + \alpha\sqrt{|S|} \cdot 2\sqrt{s} \right) \\ &\leq M\sqrt{s\,|S|} \left(\sqrt{\frac{\mu_1^2}{d}} + 2\alpha \right), \end{aligned}$$

with probability greater than $1 - d\exp\left(-c\alpha^2\,|S|\right)$. □

Finally we are in a position to establish a bound on the accuracy of the SVD step in Algorithm 8. Having bounded the contribution from from the directions apart from $A^*{}_1$ in the previous lemma, we will now lower bound the contribution of the $A^*{}_1$ direction, which will ensure that the largest singular vector is close to $A^*{}_1$.

Proof of Proposition 4.3.5: Recall the definitions of the sets S and $\widetilde{S}$ (16). In order for a vector Y_i to end up in $\widetilde{S}$, the event in Lemma C.2.2 has to fail. Hence, if we define E_i to be the random variable which is 1 if $Y_i \in \widetilde{S}$, then we have from Hoeffding's inequality

$$\left| \frac{1}{m} \sum_{i=1}^{m} (E_i - \mathbb{P}[E_i = 1]) \right| \leq \sqrt{\frac{2\log(2/\delta)}{m}},$$

with probability at least $1 - \delta/2$. From Lemma C.2.2 we further know that $\mathbb{P}[E_i = 1] \leq s^3/r$ so that

$$\left| \widetilde{S} \right| \leq \frac{ms^3}{r} + \alpha m, \tag{18}$$

with probability at least $1 - \exp(-2\alpha^2 m)$. As a consequence, the size of S is at least

$$|S| \geq m(1 - s^3/r - \alpha) \geq 9m/10 \tag{19}$$

for $\alpha < 1/20$ by our assumption that $s^3 < r/384$.

In order to understand the singular vector $\hat{a}$, we now write the matrix $\widehat{Q}$ as the sum of two matrices Q and $\widetilde{Q}$ as follows:

$$\widehat{Q} = Q + \widetilde{Q}, \text{ where,}$$
$$Q := \sum_{Y_i \in S} Y_i Y_i^T \text{ and } \widetilde{Q} := \sum_{Y_i \in \widetilde{S}} Y_i Y_i^T.$$

Recalling our earlier notation $\breve{Y}_i$ (17), we expand M as follows:

$$\begin{aligned} M &= \sum_{Y_i \in S} Y_i Y_i^T \\ &= \sum_{Y_i \in \widetilde{S}} \left(X^{*1}{}_i\right)^2 A^*{}_1 A^*{}_1{}^T + \sum_{i:Y_i \in S} X^{*1}{}_i \left(A^*{}_1 \breve{Y}_i^T + \breve{Y}_i A^*{}_1{}^T\right) + \sum_{i:Y_i \in S} \breve{Y}_i \breve{Y}_i^T \end{aligned}$$

We wish to show that $A^*{}_1$ is close to the top singular vector of $\widehat{Q}$. In order to show this, we bound the spectral norms of the following matrices: $\sum_{i:Y_i \in S} X^{*1}{}_i \left(A^*{}_1 \breve{Y}_i^T + \breve{Y}_i A^*{}_1{}^T\right)$, $\sum_{i:Y_i \in S} \breve{Y}_i \breve{Y}_i^T$ and $\widetilde{Q}$.

Using Lemma C.2.8, we first obtain:

$$\left\|\sum_{i:Y_i\in S} X^{*1}{}_i A^*{}_1 \breve{Y}_i^T\right\|_2 \leq \|A^*{}_1\|_2 \left\|\breve{Y}_S\right\|_2 \left\|X^{*S}{}_1\right\|_2$$

$$\leq M^2\sqrt{s\,|S|}\left(\sqrt{\frac{\mu_1^2}{d}}+2\alpha\right)\cdot\sqrt{|S|}$$

$$= M^2 s\,|S|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{2\alpha}{\sqrt{s}}\right) \text{ and,} \tag{20}$$

$$\left\|\sum_{i:Y_i\in S}\breve{Y}_i\breve{Y}_i^T\right\|_2 = \left\|\breve{Y}_S\breve{Y}_S^T\right\|_2 \leq 2M^2 s\,|S|\left(\frac{\mu_1^2}{d}+4\alpha^2\right). \tag{21}$$

Finally, we have the following bound on the spectral norm of $\widetilde{M}$:

$$\left\|\widetilde{Q}\right\|_2 = \left\|\sum_{Y_i\in\widetilde{S}} Y_i Y_i^T\right\|_2 \leq \left|\widetilde{S}\right| \|Y_i\|_2^2 \leq \left|\widetilde{S}\right| 2M^2 s. \tag{22}$$

Using (20), (21) and (22), we now prove the statement of the lemma. Let $|\langle A^*{}_1, \widehat{a}\rangle| = \theta$. On one hand, we have:

$$\left\|\widehat{a}^T\widehat{Q}\widehat{a}\right\|_2$$

$$\leq \theta^2\sum_{Y_i\in\widetilde{S}}\left(X^{*1}{}_i\right)^2 + M^2\left(2\left\|\sum_{i:Y_i\in S} X^{*1}{}_i A^*{}_1\breve{Y}_i^T\right\|_2 + \left\|\sum_{i:Y_i\in S}\breve{Y}_i\breve{Y}_i^T\right\|_2 + \left\|\widetilde{M}\right\|_2\right)$$

$$\leq \theta^2\sum_{Y_i\in\widetilde{S}}\left(X^{*1}{}_i\right)^2 + M^2\left(2s\,|S|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{2\alpha}{\sqrt{s}}\right) + 2s\,|S|\left(\frac{\mu_1^2}{d}+4\alpha^2\right) + \left|\widetilde{S}\right| 2s\right)$$

$$\leq |S|\left[\theta^2\frac{\sum_{Y_i\in\widetilde{S}}\left(X^{*1}{}_i\right)^2}{|S|} + 8sM^2\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\alpha^2+\frac{\alpha}{\sqrt{s}}+\left(\frac{s^3}{r}+\alpha\right)\right)\right],$$

where the last step uses the bounds (18) and (19). On the other hand, we have

$$
\begin{aligned}
&\left\|\widehat{a}^T\widehat{Q}\widehat{a}\right\|_2 = \left\|\widehat{Q}\right\|_2 \\
&\geq \sum_{Y_i\in\widetilde{S}} \left(X^{*1}_i\right)^2 \cdot \|A^*{}_1\|_2^2 \\
&\qquad\qquad - M^2\left(2\left\|\sum_{i:Y_i\in S} X^{*1}_i A^*{}_1 \breve{Y}_i^T\right\|_2 - \left\|\sum_{i:Y_i\in S}\breve{Y}_i\breve{Y}_i^T\right\|_2 - \left\|\widetilde{Q}\right\|_2\right) \\
&\geq \sum_{Y_i\in\widetilde{S}} \left(X^{*1}_i\right)^2 - M^2\left(2s\left|S\right|\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\alpha}{\sqrt{s}}\right) - 2s\left|S\right|\left(\frac{\mu_1^2}{d}+4\alpha^2\right) - \left|\widetilde{S}\right|2s\right) \\
&\geq |S|\left[\frac{\sum_{Y_i\in\widetilde{S}}\left(X^{*1}_i\right)^2}{|S|} - 8sM^2\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\alpha^2+\frac{\alpha}{\sqrt{s}}+\left(\frac{s^3}{r}+\alpha\right)\right)\right].
\end{aligned}
$$

Using the above two inequalities, we obtain

$$
\begin{aligned}
\theta^2 &\geq 1 - M^2\frac{16s\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\frac{s^3}{r}\right) - 16s\left(\alpha^2+\frac{\alpha}{\sqrt{s}}\right)}{\frac{\sum_{Y_i\in\widetilde{S}}\left(X^{*1}_i\right)^2}{|S|}} \\
&\geq 1 - M^2\left(16s\left(\frac{\mu_1}{\sqrt{ds}}+\frac{\mu_1^2}{d}+\frac{s^3}{r}\right) - 16s\left(\alpha^2+\frac{\alpha}{\sqrt{s}}\right)\right)
\end{aligned}
$$

Now we observe that since $\|A^*{}_1\|_2 = \|\hat{a}\|_2 = 1$, we have

$$
\|\hat{a} - A^*{}_1\|_2^2 = 2(1-\theta) \leq 2(1-\theta^2),
$$

for $0 \leq \theta \leq 1$, which completes the proof.

□

C.2.8 Bounding the size of $\widehat{S}$

So far, we have established that the sub-procedure in Algorithm 1 correctly detects good anchor pairs with high probability. Conditioned on this,

Proposition 4.3.5 shows that we can recover the dictionary element in this intersection to a bounded error with high probability. In this section, we prove Lemma C.2.5. Before moving to its proof, we have the following useful lemma.

Lemma C.2.9 (Number of good anchor pairs)**.** *Suppose we have n examples. Then, we have:*

$$\mathbb{P}\left\{\cup_{l\in[r]}\left|\{(i,j):\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_j)=\{A^*{}_l\}\}\right|>\frac{ns}{8r}\right\}\geq 1-r\exp\left(\frac{-ns}{64r}\right).$$

Proof: Fix $l\in[r]$. Define the set $S\subseteq[n]$ as follows:

$$S:=\left\{i:A^*{}_l\in\mathcal{N}_B(Y_i)\right\}.$$

Since for every $i\in[n]$, the probability of $i\in S$ is $\frac{s}{r}$, using standard Chernoff bounds, we see that:

$$\mathbb{P}\left[|S|<\frac{ns}{2r}\right]<\exp\left(\frac{-ns}{8r}\right). \tag{23}$$

Consider any two examples $Y_i,Y_j\in S$. Then,

$$\mathbb{P}\left[\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_j)=\{A^*{}_l\}\right]\geq 1-\frac{s^2}{r}.$$

Dividing the set S into $\frac{|S|}{2}$ disjoint pairs and using Chernoff bounds, we see that

$$\begin{aligned}\mathbb{P}\left[\left|\{(i,j):\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_j)=\{A^*{}_l\}\}\right|<\frac{|S|}{4}\right]&\leq\exp\left(\frac{-\left(1-\frac{s^2}{r}\right)|S|}{16}\right)\\&\leq\exp\left(\frac{-|S|}{32}\right).\end{aligned}\tag{24}$$

Using (23) and (24), we have:

$$\mathbb{P}\left[\left|\{(i,j):\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_j)=\{A^*{}_l\}\}\right|>\frac{ns}{8r}\right]\geq 1-\exp\left(\frac{-ns}{64r}\right).$$

Using a union bound over different dictionary elements, we have:

$$\mathbb{P}\left[|\{(i,j)|\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_j)=\{A^*{}_l\}\}|>\frac{ns}{8r}\;\forall\, l\in[r]\right]\geq 1-r\exp\left(\frac{-ns}{64r}\right).$$

□

Proof of Lemma C.2.5: Since $(Y_{i^*},Y_{j^*})\in G_{\mathrm{corr}(\rho)}$, from Lemma C.2.1, we know that $\mathcal{N}_B(Y_{i^*})\cap\mathcal{N}_B(Y_{j^*})\neq\emptyset$. Wlog let $A^*{}_1\in\mathcal{N}_B(Y_{i^*})\cap\mathcal{N}_B(Y_{j^*})$. Since each sample Y_i has probability of at least

$$\frac{s}{r}\cdot\frac{\binom{r-2s+1}{s-1}}{\binom{r-1}{s-1}}\geq\frac{s}{r}\cdot\left(\frac{r-3s}{r-s}\right)^s\geq\frac{s}{r}\cdot\left(1-\frac{2s}{r-s}\right)^s\geq\frac{s}{r}\cdot\left(1-\frac{2s^2}{r-s}\right)\geq\frac{s}{2r},$$

of satisfying $\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_{i^*})=\mathcal{N}_B(Y_i)\cap\mathcal{N}_B(Y_{j^*})=\{A^*{}_1\}$, using Chernoff bounds, we have:

$$\mathbb{P}\left[|i:\text{Uniq-intersect}\,(Y_i,Y_{i^*})\,\&\,\text{Uniq-intersect}\,(Y_i,Y_{j^*})|<\frac{ns}{4r}\right]\leq\exp\left(\frac{-ns}{16r}\right).$$

Using Lemma C.2.1 now finishes the proof. □

Lemma C.2.10. *For $r>2, c>0$, let $0\leq x\leq r/(2c+1)$. Then $(1-cx/(r-x))^x\geq\exp(-cx^2/(r-x))\geq 1-\frac{2x^2}{r-x}$.*

Proof:

We start by observing that $x/(r-x)$ is an increasing function of x for $x<r$, so that $x<r/(2c+1)$ implies that $cx/(r-x)<1/2$. Additionally, we have the following fact for any $\theta>0$

$$1-\theta\leq e^{-\theta}\leq 1-\theta+\frac{\theta^2}{2}. \tag{25}$$

The first inequality is a consequence of the convexity of $e^{-\theta}$ while the second one follows since the second derivative of $e^{-\theta}$ is at most 1 when $\theta>0$. Since we have $x/(r-x)\leq 1/2$, it is easy to see that

$$1-\frac{cx}{r-x}\geq 1-2\frac{cx}{r-x}+2\frac{c^2x^2}{(r-x)^2}.$$

Now applying the inequalities (25) with $\theta = 2cx/(r-x)$, we obtain

$$\begin{aligned}\left(1 - \frac{cx}{r-x}\right)^x &\geq \left(1 - 2\frac{cx}{r-x} + 2\frac{cx^2}{(r-x)^2}\right)^x \\ &\geq (\exp(-2cx/(r-x)))^x = \exp(-2cx^2/(r-x) \\ &\geq 1 - \frac{2cx^2}{r-x},\end{aligned}$$

where the second inequality follows from again using (25), this time with $\theta = 2cx^2/(r-x)$. □

C.3 Proofs for alternating minimization

In this section, we will present our proof for the results on alternating minimization. We present the proofs for Theorem 4.3.3 and the other main lemmas in Section C.3.1. In Section C.3.2, we present the auxiliary lemmas and their proofs.

C.3.1 Proofs of main lemmas

In this section we will present the proof of the main lemmas used to prove Theorem 4.3.3. The proofs of some auxilliary lemmas and more technical arguments will be deferred to the next section.

We recall from Appendix C.1, the following notational simplification: for any iteration t, we denote $A(t)$ as $\widetilde{A}$ and $A(t+1)$ as A. Similarly we denote $X(t)$ and $X(t+1)$ as $\widetilde{X}$ and X respectively. Then the goal is to show that A is closer to A^* than $\widetilde{A}$. For the purposes of our analysis, we will find it more convenient to directly work with dot products instead of ℓ_2-distances (and hence avoid sign ambiguities). With this motivation, we define the following notion of distance between two vectors.

Definition C.3.1. For any two vectors $z, w \in \mathbb{R}^d$, we define the distance between them as follows:

$$dist\,(z,w) \stackrel{def}{=} \sup_{v \perp w} \frac{\langle v, z\rangle}{\|v\|_2 \|z\|_2} = \sup_{v \perp z} \frac{\langle v, w\rangle}{\|v\|_2 \|w\|_2}.$$

This definition of distance suffices for our purposes due to the following simple lemma

Lemma C.3.1. *For any two unit vectors $u, v \in \mathbb{R}^d$, we have*

$$\min_{z \in \{-1,1\}} \|zu - v\|_2 \leq \sqrt{2} dist\,(u, v)\,.$$

Proof: The proof is rather straightforward. Suppose that $\langle u, v \rangle > 0$ so that the minimum happens at $z = 1$. The other case is identical. We can easily rewrite

$$\|u - v\|_2^2 = (2 - 2\langle u, v \rangle) \leq 2(1 - \langle u, v \rangle^2),$$

where the final inequality follows since $0 \leq \langle u, v \rangle \leq 1$. Writing $u = \langle u, v \rangle v + v_\perp$, where $\langle v_\perp, v \rangle = 0$, we see that

$$1 = \|u\|_2^2 = \langle u, v \rangle^2 + \|v_\perp\|^2 = \langle u, v \rangle^2 + dist\,(u, v)^2\,.$$

Substituting this into our earlier bound completes the proof. □

The distance is naturally extended to matrices for our purposes by applying it columnwise.

Definition C.3.2. For any two $d \times r$ matrices Z and W, we define the distance between them as follows:

$$dist\,(Z, W) \stackrel{def}{=} \sup_{p \in [r]} dist\,(Z_p, W_p)\,.$$

Note that the normalization in the definition of $dist\,(z, w)$ ensures that we can apply the distance directly to the result of the least-squares step without worrying about the effects of normalization. This allows us to work with the closed-form expression for A

$$A = YX^+ = A^* X^* X^+. \tag{26}$$

We first recall Lemmas 4.3.6, 4.3.7 and 4.3.8 from Section 4.3.4.

Lemma 4.3.6 (Error in sparse recovery)**.** Let $\triangle X \stackrel{def}{=} X(t) - X^*$. Assume that $2\mu_0 s/\sqrt{d} \leq 0.1$ and $\sqrt{s\epsilon_t} \leq 0.1$ Then, we have:

1. $\text{Supp}(\triangle X) \subseteq \text{Supp}(X^*)$.
2. $\|\triangle X\|_\infty \leq 9s \cdot dist\left(A(t-1), A^*\right) \leq 9s\epsilon_t$.

Lemma 4.3.6 shows that if the initial estimate of A^* is good enough, then the error in the recovered coefficients from the sparse recovery step are small and have structured sparsity.

Lemma 4.3.7. For every $r \times n$ matrix W s.t. $\text{Supp}(W) \subseteq \text{Supp}(X^*)$, we have (w.p. $\geq 1 - r\exp\left(-\frac{Cn}{r}\right)$):

$$\|W\|_2 \leq 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}}.$$

Lemma 4.3.7 shows that all matrices with structured sparsity have bounded spectral norm.

Lemma 4.3.8 (Off-diagonal error bound)**.** Suppose $\|\triangle X\|_\infty < \frac{1}{288s}$. Then with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$, we have uniformly for every $p \in [r]$,

$$\left\|\left(\triangle X X^+\right)_p^{\setminus p}\right\|_2 = \left\|\left(X^* X^+\right)_p^{\setminus p}\right\|_2 \leq \frac{1968 s^2 \|\triangle X\|_\infty}{\sqrt{r}}.$$

Lemma 4.3.8 shows that the off-diagonal norm of each column of the matrix $X^* X^+$ is quite small.

We now prove Lemma 4.3.6, which follows from the robustness properties of lasso. We first need an auxilliary result on the RIP constant of the matrix A^*.

Lemma C.3.2. *The $2s$-RIP constant of A^*, δ_{2s} satisfies $\delta_{2s} < \frac{2\mu_0 s}{\sqrt{d}}$.*

Proof: Consider a $2s$-sparse unit vector $w \in \mathbb{R}^r$ with $\text{Supp}(w) = S$. We have:

$$\begin{aligned}\|Aw\|^2 = \left(\sum_{j\in S} w_j {A^*}_j\right)^2 &= \sum_j w_j^2 \|{A^*}_j\|^2 + \sum_{j,l\in S, j\neq l} w_j w_l \langle {A^*}_j, {A^*}_l\rangle \\ &\geq 1 - \sum_{j,l\in S, j\neq l} |w_j w_l| \, |\langle {A^*}_j, {A^*}_l\rangle| \\ &\geq 1 - \sum_{j,l\in S, j\neq l} |w_j w_l| \, \frac{\mu_0}{\sqrt{d}} \\ &\geq 1 - \frac{\mu_0}{\sqrt{d}} \|w\|_1^2 \\ &\geq 1 - \frac{\mu_0}{\sqrt{d}} 2s \cdot \|w\|^2 = 1 - \frac{2\mu_0 s}{\sqrt{d}}.\end{aligned}$$

Similarly, we have:

$$\|A^* w\|^2 \leq 1 + \frac{2\mu_0 s}{\sqrt{d}}.$$

This proves the lemma. □

Proof of Lemma 4.3.6: In order to establish the lemma, we use a result of Candes regarding the lasso estimator with deterministic noise for the recovery procedure:

$$\widehat{x}_i = \arg\min_{x\in\mathbb{R}^r} \|x\|_1 \quad \text{such that,} \quad \|Y_i - Ax\|_2 \leq \epsilon. \tag{27}$$

Theorem C.3.3 (Theorem 1.2 from [11])**.** *Suppose $Y_i = Ax_i + z_i$, where x_i is s-sparse and $\|z_i\|_2 \leq \epsilon$. Assume further that $\delta_{2s} \leq \sqrt{2} - 1$. Then the solution to Equation* (27) *obeys the following, for a universal constant C_1,*

$$\|\widehat{x}_i - x_i\|_2 \leq C_1 \epsilon$$

In particular, $C_1 = 8.5$ suffices for $\delta_{2s} \leq 0.2$.

In order to apply the theorem, we need to demonstrate that the RIP condition holds on $\widetilde{A}$. Consider any $2s$-sparse subset S of $[r]$. We have:

$$\sigma_{\min}(\widetilde{A}_S) \geq \sigma_{\min}(A^*_S) - \|A^*_S - \widetilde{A}_S\|_2 \overset{(\zeta_1)}{\geq} 1 - \frac{2\mu_0 s}{\sqrt{d}} - \left\|A^*_S - \widetilde{A}_S\right\|_{\mathrm{F}} \quad \text{and,}$$
$$\sigma_{\max}(\widetilde{A}_S) \leq \sigma_{\max}(A^*_S) + \|A^*_S - \widetilde{A}_S\|_2 \overset{(\zeta_2)}{\leq} 1 + \frac{2\mu_0 s}{\sqrt{d}} + \left\|A^*_S - \widetilde{A}_S\right\|_{\mathrm{F}},$$

where ζ_1 and ζ_2 follow from Lemma C.3.2. Recalling the assumption $\sqrt{s\epsilon_t} < 0.1$, we see that the maximum and minimum singular values of $\widetilde{A}_S$ are at least $6/7$ and at most $8/7$ respectively. Appealing to Theorem C.3.3, we see that this guarantees $\|\triangle X_i\|_2 \leq 9s\epsilon_t$. Since this is also an infinity norm error bound, we obtain the second part of the lemma. The proof of the first part is further implied by the choice of our threshold at a level of $9s\epsilon_t$, which ensures that any non-zero element in X has $\left|{X^*}^i_p\right| \geq 0$ (since we would have $\left|X^i_p\right| \leq 9s\epsilon_t$ by our infinity norm bound otherwise. $\square$

We now move on to the proof of Lemma 4.3.7, which is a surprising but rather straight forward consequence of random matrix concentration theory.

Proof of Lemma 4.3.7: Since the support of W is a subset of the support of X^*, $W^p_i = \delta^p_i W^p_i$. Now,

$$\begin{aligned}\|W\|_2 &= \max_{u,v \| u\|_2=1, \|v\|_2=1} \sum_{ip} W^p_i u^i v^p = \max_{u,v \| u\|_2=1, \|v\|_2=1} \sum_{ip} \delta^p_i W^p_i u^i v^p \\ &\leq \|W\|_\infty \cdot \max_{u,v \| u\|_2=1, \|v\|_2=1} \cdot \sum_{ip} \delta^p_i u^i v^p,\end{aligned}$$

where the inequality holds since the maximum inner product over the all pairs (u, v) from the unit sphere is larger than that over pairs with $u^i v^p \geq 0$ for all i, p. Note that the last expression is equal to $\|W\|_\infty u^\top S(X^*) v$, where we overload the notation $S(X^*)$ to also be the matrix with the non-zero pattern of the matrix X^*. It suffices to control the operator norm of this matrix for proving the lemma. This can indeed be done by applying Lemmas C.3.4 and C.3.5 with $\mu = M = 1$ and $\sigma = 0$. Doing so, yields with probability at least $\geq 1 - r \exp\left(-\frac{Cn}{r}\right)$

$$\|W\|_2 \leq 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}},$$

which completes the proof. $\square$

We now finally prove Lemma 4.3.8, which is our main lemma on the structure of $X^* X^+$. Specifically, the lemma will show how to control the off-diagonal elements of this matrix carefully.

Proof of Lemma 4.3.8: For simplicity, we will prove the statement for $p = 1$. We first relate $X^* X^+$ to $\triangle X X^+$.

$$\begin{aligned}
\left(X^* X^+\right)_1^{\backslash 1} &= \left(\left(X^* - X\right) X^+\right)_1^{\backslash 1} \\
&= -\left(\triangle X X^+\right)_1^{\backslash 1} \\
&= -\left(\triangle X X^\top \left(X X^\top\right)^{-1}\right)_1^{\backslash 1},
\end{aligned}$$

where the first step follows from the fact that $X X^+ = \mathbb{I}$. This proves the first part of the lemma. We now expand the above as follows:

$$\left(\triangle X X^\top \left(X X^\top\right)^{-1}\right)_1^{\backslash 1} = \left(\triangle X X^\top\right)_1^{\backslash 1} \left(\left(X X^\top\right)^{-1}\right)_1^1 + \left(\triangle X X^\top\right)_{\backslash 1}^{\backslash 1} \left(\left(X X^\top\right)^{-1}\right)_1^{\backslash 1}.$$

Using triangle inequality, we have:

$$\begin{aligned}
\left\| \left(\triangle X X^\top \left(X X^\top\right)^{-1}\right)_1^{\backslash 1} \right\|_2 \leq & \underbrace{\left| \left(\left(X X^\top\right)^{-1}\right)_1^1 \right|}_{\mathcal{T}_1} \underbrace{\left\| \left(\triangle X X^\top\right)_1^{\backslash 1} \right\|_2}_{\mathcal{T}_2} \\
& + \underbrace{\left\| \left(\triangle X X^\top\right)_{\backslash 1}^{\backslash 1} \right\|_2}_{\mathcal{T}_3} \underbrace{\left\| \left(\left(X X^\top\right)^{-1}\right)_1^{\backslash 1} \right\|_2}_{\mathcal{T}_4}.
\end{aligned} \tag{28}$$

We now bound each of the above four quantities. We can easily bound $\mathcal{T}_1$ via a spectral norm bound on $\left(X X^\top\right)^{-1}$. Doing so, we obtain with probability at least $1 - r \exp(-\frac{C\delta^2 n s}{r M^2})$

$$\mathcal{T}_1 = \left| \left(\left(XX^\top \right)^{-1} \right)_1^1 \right| \leq \left\| \left(XX^\top \right)^{-1} \right\|_2 \overset{(\zeta_1)}{\leq} \frac{8r}{ns}, \tag{29}$$

where (ζ_1) follows from Lemma C.3.5. To bound $\mathcal{T}_2$, we use Lemma C.3.9 and obtain with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$ $-\exp\left(-ns^2/(3r^2)\right)$

$$\mathcal{T}_2 = \left\| \left(\triangle X X^\top \right)_1^{\setminus 1} \right\|_2 \leq \frac{6 \left\| \triangle X \right\|_\infty s^2 n}{r^{\frac{3}{2}}}, \tag{30}$$

where we recall the assumption $\left\| \triangle X \right\|_\infty \leq 1/(64s)$. We now bound $\mathcal{T}_3$ as follows

$$\begin{aligned} \mathcal{T}_3 = \left\| \left(\triangle X X^\top \right)_{\setminus 1}^{\setminus 1} \right\|_2 &\leq \left\| \left(\triangle X \right)_{\setminus 1}^{\setminus 1} \right\|_2 \left\| \left(X \right)_{\setminus 1}^{\setminus 1} \right\|_2 \\ &\overset{(\zeta_1)}{\leq} 2 \left\| \triangle X \right\|_\infty s \sqrt{\frac{n}{r}} \cdot 2(1 + \left\| \triangle X \right\|_\infty) s \sqrt{\frac{n}{r}} \\ &< \frac{6 \left\| \triangle X \right\|_\infty s^2 n}{r}, \end{aligned} \tag{31}$$

where (ζ_1) follows from Lemmas 4.3.7 and C.3.6 (since $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X) \cup \mathrm{Supp}(X^*) = \mathrm{Supp}(X^*)$). Finally, to bound $\mathcal{T}_4$, we start by noting the following block decomposition of the matrix $XX^\top$

$$XX^\top = \begin{bmatrix} X^1 (X^1)^\top & X^1 (X^{\setminus 1})^\top \\ X^{\setminus 1} {X^1}^\top & X^{\setminus 1} (X^{\setminus 1})^\top \end{bmatrix}.$$

Given this block-structure, we can now invoke Lemma C.3.10 (Schur complement lemma) to obtain

$$\left(\left(XX^\top \right)^{-1} \right)_{\setminus 1}^1 = -\frac{1}{X^1 (X^1)^\top} B X^{\setminus 1} \left(X^1 \right)^\top,$$

where,

$$B \stackrel{def}{=} \left(\left(XX^\top\right)^{-1}\right)^{\setminus 1}_{\setminus 1}. \tag{32}$$

Using Lemma C.3.9 and Equation 38 we have with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$

$$\begin{aligned}\left\|\left(\left(XX^\top\right)^{-1}\right)^{\setminus 1}_{\setminus 1}\right\|_2 &\leq \frac{1}{\left|X^1(X^1)^\top\right|}\left\|B\right\|_2\left\|X^{\setminus 1}\left(X^1\right)^\top\right\|_2 \leq \frac{8r}{sn}\cdot\left\|B\right\|_2\cdot\frac{5s^2n}{r^{\frac{3}{2}}} \\ &= \frac{40s}{\sqrt{r}}\left\|M\right\|_2. \end{aligned} \tag{33}$$

Using the expression (32) and the lower bound on $\sigma_{\min}(X)$ from Lemma C.3.6, we also have the following bound for $\left\|M\right\|_2$ with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$,

$$\left\|B\right\|_2 = \left\|\left(\left(XX^\top\right)^{-1}\right)^{\setminus 1}_{\setminus 1}\right\|_2 \leq \left\|\left(XX^\top\right)^{-1}\right\|_2 \leq \frac{8r}{ns}.$$

Plugging the above into (33), gives us:

$$\left\|\left(\left(XX^\top\right)^{-1}\right)^{\setminus 1}_{\setminus 1}\right\|_2 \leq \frac{40s}{\sqrt{r}}\cdot\frac{8r}{ns} \leq \frac{320\sqrt{r}}{n}. \tag{34}$$

Combining (29), (30), (31) and (34), we obtain with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$,

$$\begin{aligned}\left\|\left(XX^{*+}\right)^{\setminus p}_{p}\right\|_2 &\leq \frac{48\left\|\triangle X\right\|_\infty s}{\sqrt{r}} + \frac{1920\left\|\triangle X\right\|_\infty s^2}{\sqrt{r}} \\ &\leq \frac{1968s^2\left\|\triangle X\right\|_\infty}{\sqrt{r}}.\end{aligned}$$

□

C.3.2 Main Technical Lemmas

In this section, we state and prove the main technical lemmas used in our results.

C.3.2.1 Assumptions

We first recall some notation and define additional shorthands before proving the lemmas. Denote ${X^*}_i^p = \delta_i^p M_i^p,\ \forall 1 \leq p \leq n,\ \forall 1 \leq i \leq r$ where $\delta_i^p = 1$ if $p \in \text{Supp}(X_i^*)$ and 0 otherwise and M_i^p are i.i.d. random variables with $\mathbb{E}\left[M_i^p\right] = \mu$ and $\mathbb{E}\left[(M_i^p)^2\right] = \sigma^2 + \mu^2$. Assumptions $(C3) - (C4)$ give us:

1. $\mu^2 + \sigma^2 = 1$, and
2. $|M_i^p| \leq M$ a.s.

C.3.2.2 Proofs of Technical Lemmas

We prove all of our technical lemmas under the assumption that X^* is sampled as described in Section C.3.2.1.

Lemma C.3.4. *We have:*

$$\Sigma \stackrel{def}{=} \mathbb{E}\left[{X^*}_i {X^*}_i^\top\right] = \left(\frac{s}{r} - \frac{s(s-1)\mu^2}{r(r-1)}\right)\mathbb{I} + \frac{s(s-1)\mu^2}{r(r-1)} 1 1^\top.$$

Proof:

Note that, $\delta_i^p, 1 \leq p \leq r$ all have same distribution. Hence, by symmetry and linearity of expectation, $\mathbb{E}\left[\delta_i^p\right] = \frac{1}{r}\mathbb{E}\left[\sum_{q=1}^r \delta_i^q\right] = \frac{s}{r}$. Similarly, $\mathbb{E}\left[(\delta_i^p)^2\right] = \frac{1}{r}\mathbb{E}\left[\sum_{q=1}^r (\delta_i^q)^2\right] = \frac{s}{r}$. Also, $\mathbb{E}\left[(\sum_{q=1}^r (\delta_i^q))^2\right] = \mathbb{E}\left[\sum_{p,q} \delta_i^p \delta_i^q\right] = r\mathbb{E}\left[(\delta_i^p)^2\right] + (r^2 - r)\mathbb{E}\left[\delta_i^p \delta_i^q\right]$. Hence, $\mathbb{E}\left[\delta_i^p \delta_i^q\right] = \frac{s(s-1)}{r(r-1)}$.

Now, recall that ${X^*}_i^p = \delta_i^p M_i^p$. Now, we first consider diagonal terms of Σ:

$$\Sigma_p^p = \mathbb{E}\left[({X^*}_i^p)^2\right] = \mathbb{E}\left[(\delta_i^p)^2\right]\mathbb{E}\left[(M_i^p)^2\right] = \frac{s}{r}(\mu^2 + \sigma^2) = \frac{s}{r}. \tag{35}$$

Similarly, using independence of X^{*p}_i and X^{*j}_p, off-diagonal terms of Σ are given by:

$$\Sigma^q_p = \mathbb{E}\left[\delta^p_i \delta^q_i\right] \mathbb{E}\left[M^p_i\right] \mathbb{E}\left[M^q_i\right] = \frac{s(s-1)}{r(r-1)}\mu^2. \tag{36}$$

Lemma now follows by using (35) and (36). □

In particular, two consequences of the lemma which will be particularly useful are about the extreme singular values of Σ. Recalling that $2s \leq r$ and $\mu^2 \leq 1$ by assumption, we obtain

$$\sigma_{\min}(\Sigma) \geq \frac{s}{2r}, \quad \text{and} \quad \sigma_{\max}(\Sigma) \leq \frac{2s^2}{r}. \tag{37}$$

For convenience of the reader, we again recall Theorem C.2.6.

Theorem C.2.6 (Restatement of Theorem 5.44 from [88])**.** Consider a $r \times n$ matrix W where each column w_i of W is an independent random vector with covariance matrix Σ. Suppose further that $\|w_i\|_2 \leq \sqrt{u}$ a.s. for all i. Then for any $t \geq 0$, the following inequality holds with probability at least $1 - r\exp\left(-ct^2\right)$:

$$\left\|\frac{1}{n}WW^T - \Sigma\right\|_2 \leq \max\left(\|\Sigma\|_2^{1/2}\gamma, \gamma^2\right) \text{ where } \gamma = t\sqrt{\frac{u}{n}}.$$

Here $c > 0$ is an absolute numerical constant. In particular, this inequality yields:

$$\|W\|_2 \leq \|\Sigma\|_2^{\frac{1}{2}}\sqrt{n} + t\sqrt{u}.$$

We need the following results on concentration of empirical covariance matrices.

Lemma C.3.5. *There exists a universal constant C such that w.p. $\geq 1 - r\exp(-\frac{C\delta^2 ns}{rM^2})$, we have:*

$$\left\|\frac{1}{n}X^*X^{*\top} - \Sigma\right\|_2 \leq \max\left(\sqrt{2}\delta, \delta^2\right)\frac{s^2}{r}.$$

In particular, w.p. $\geq 1 - r\exp(-\frac{Cn}{rM^2})$*, we have the bounds*

$$\|X^*\|_2 \leq 2\sqrt{\frac{ns^2}{r}} \quad \textit{and} \quad \sigma_{\min}(X^*) \geq \sqrt{\frac{ns}{4r}}.$$

Proof:

Note that, $\|X_i^*\|_2 \leq \sqrt{s}M$. Also, $\|\Sigma\|_2 \leq \frac{s}{r} + \frac{s(s-1)\mu^2}{r-1} \leq \frac{2s^2}{r}$. Using Theorem C.2.6 with $t = \delta\sqrt{\frac{ns}{rM^2}}$, we obtain:

$$\left\|\frac{1}{n}X^*X^{*\top} - \Sigma\right\|_2 \leq \max\left(\sqrt{2}\delta, \delta^2\right)\frac{s^2}{r},$$

w.p. greater than $1 - r\exp\left(-\frac{C\delta^2 ns}{rM^2}\right)$. In order to obtain the second part, we apply the first part of the lemma with $\delta = 1/4\sqrt{2}$ as well as Lemma C.3.4 to bound the largest and smallest singular values of $XX^\top/n$. Taking square roots completes the proof.

□

Using the convergence of the covariance matrix of the sparsity pattern, we obtain the following uniform convergence bound.

In particular the following consequence of the above lemma would be particularly useful in our proofs, where we apply the lemma to matrices of the form $\triangle X = X - X^*$.

Lemma C.3.6. *W.p.* $\geq 1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$*, for every* $r \times n$ *matrix* X *s.t.* $\text{Supp}(X) \subseteq \text{Supp}(X^*)$*, we have:*

$$\|X\|_2 \leq 2 \cdot (1 + \|X - X^*\|_\infty) \cdot s\sqrt{\frac{n}{r}}.$$

Proof:

Let $X = X^* + E_{X^*}$ where $\text{Supp}(E_{X^*}) \subseteq \text{Supp}(X^*)$. Hence, $\|X\|_2 \leq \|X^*\|_2 + \|X - X^*\|_2$. Lemma follows directly using Lemma C.3.5 and Lemma 4.3.7.

□

A useful version of the above lemma is when applied to matrices of the form $XX^\top$. We will need control over the upper and lower singular values of such matrices for our proofs, which we next provide.

Lemma C.3.7. *W.p.* $\geq 1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$*), for every* $r \times n$ *matrix* X *s.t.* $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X^*)$*, we have:*

$$\left\|XX^\top - X^*X^{*\top}\right\|_2 \leq 4\left(\|X - X^*\|_\infty + \|X - X^*\|_\infty^2\right) \cdot \frac{s^2 n}{r}.$$

Further assuming $\|X - X^*\|_\infty \leq 1/(64s)$*, we have with the same probability*

$$\sigma_{\min}(XX^\top) \geq \frac{ns}{8r}.$$

Proof:

Let $X = X^* + E_{X^*}$. Note that $\mathrm{Supp}(E_{X^*}) \subseteq \mathrm{Supp}(X^*)$. Now,

$$\|XX^\top - X^*X^{*\top}\|_2 \leq \|E_{X^*}\|_2(\|E_{X^*}{}^\top\|_2 + 2\|X^*\|_2).$$

By Lemma 4.3.7, $\|E_{X^*}\|_2 \leq 2s\sqrt{\frac{n}{r}}\|\mathbb{E}_{X^*}\|_\infty$ with probability at least $\geq 1 - r\exp\left(-\frac{Cn}{r}\right)$. Combining this with the bound on $\|X^*\|_2$ from Lemma C.3.5 completes the proof. The second statement now follows by combining the result with our earlier lower bound on the minimum singular value of X^* in Lemma C.3.5. □

A particular consequence of this lemma which will be useful is a lower bound on the diagonal entries of the matrix $XX^\top$. Indeed, we see that under the assumption $\|X - X^*\|_\infty \leq 1/(64s)$, with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right)$ we have the lower bound uniformly for all $p = 1, 2, \ldots, r$

$$X^p X^{p\top} \geq \frac{ns}{8r}. \tag{38}$$

We finally have the following concentration lemma.

Lemma C.3.8. *Let* δ_i^p *be as defined in Section C.3.2.1. Then, w.p.* $\geq 1 - \exp\left(-\frac{\delta^2}{3}ns^2/r^2\right)$*:*

1. $\sum_{i=1}^{n} \delta_i^p \delta_i^q \leq (1+\delta)\frac{s^2 n}{r^2},\ \forall\, p \neq q,$
2. $(1-\delta)\frac{sn}{r} \leq \sum_{i=1}^{n} \delta_i^p \leq (1+\delta)\frac{sn}{r},\ \forall\, p,$
3. $(1-\delta)\frac{sn}{r} \leq \sum_{i=1}^{n} \delta_i^p \left(M_i^p\right)^2 \leq (1+\delta)\frac{sn}{r},\ \forall\, p,$ *and*
4. $\sum_{i=1}^{n} \delta_i^p \left|M_i^p\right| \leq (1+\delta)\frac{sn}{r},\ \forall\, p.$

Proof:

Recall from the proof of Lemma C.3.4 that, $\mathbb{E}\left[\delta_i^p\right] = \frac{s}{r}$ and $\mathbb{E}\left[\delta_i^p \delta_i^q\right] = \frac{s(s-1)}{r(r-1)}, \forall p \neq q$. Also, these random variables are independent for each i. Using Chernoff bound, we get (w.p. $\geq 1 - \exp(-\frac{\delta^2}{3} ns^2/r^2)$):

$$(1-\delta)n\frac{s}{r} \leq \sum_{i=1}^{n} \delta_i^p \leq (1+\delta)n\frac{s}{r}, \qquad \sum_{i=1}^{n} \delta_i^p \delta_i^q \leq (1+\delta)n\frac{s^2}{r^2},\ \ \forall p \neq q.$$

The third part follows similarly using Chernoff bound. The fourth part follows from Chernoff bound as well after noting that

$$\mathbb{E}\left[\left|M_i^p\right|\right] \leq \left(\mathbb{E}\left[\left(M_i^p\right)^2\right]\right)^{\frac{1}{2}} = 1,$$

where the first step follows from Jensen's inequality. □

Lemma C.3.9. *W.p.* $\geq 1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$, *for every* $r \times n$ *matrix* X *s.t.* $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X^*)$, *we have the following bounds uniformly for all* $p = 1, 2, \ldots, r$

1. $\left\| \left(\triangle X X^\top\right)_p^{\setminus p} \right\|_2 \leq \left(1 + \|\triangle X\|_\infty\right) \frac{4\sqrt{2}\|\triangle X\|_\infty s^2 n}{r^{\frac{3}{2}}}$, *and*
2. $\left\| X^{\setminus p} (X^p)^\top \right\|_2 \leq \left(1 + \|\triangle X\|_\infty\right)^2 \frac{4s^2 n}{r^{\frac{3}{2}}}$,

where $\triangle X \stackrel{def}{=} X - X^*$.

Proof: Since X has the same sparsity pattern as X^*, we can rewrite it as $X_i^p = \delta_i^p X_i^p$. We start by proving the first part of the lemma.

Proof of Part 1: Wlog, we will prove the statement for $p = 1$. Let D denote the $n \times n$ diagonal matrix with

$$D_i^i = \begin{cases} 1, & \text{if } {X^*}_i^{\setminus 1} \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Using this notation, we have $\left(\triangle X X^\top\right)_1^{\setminus 1} = \left(\triangle X D X^\top\right)_1^{\setminus 1}$. So, we have:

$$\begin{aligned} \left\|\left(\triangle X X^\top\right)_1^{\setminus 1}\right\|_2 &= \left\|\left(\triangle X D X^\top\right)_1^{\setminus 1}\right\|_2 \\ &\leq \left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2 \left\|\left(X^\top\right)_1\right\|_2 \\ &\leq \left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2 \left\|\left(X^{*\top}\right)_1 + \left(\triangle X^\top\right)_1\right\|_2 \\ &\overset{(\zeta_1)}{\leq} \left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2 \cdot \left(\sqrt{\frac{2sn}{r}} + \|\triangle X\|_\infty \sqrt{\frac{2sn}{r}}\right), \end{aligned}$$

where (ζ_1) follows from Lemma C.3.8. In order to control $\left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2$, we observe that it is a matrix with a random number of columns selected by the matrix D. In particular, conditioned on $\{i : D_i^i = 1\}$, the support of ${X^*}_i^{\setminus 1}$ is independent over $s-1$ sparse vectors (and the support of $\triangle X$ is a subset of the support of X^*). Hence we can easily see that

$$\begin{aligned} \mathbb{P}\left[\left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2 > t\right] \leq \mathbb{P}\left[\left\|\left(\triangle X D\right)^{\setminus 1}\right\|_2 > t \cap \frac{sn}{2r} < \left|\left\{i : D_i^i = 1\right\}\right| < \frac{2sn}{r}\right] \\ + \mathbb{P}\left[\left|\left\{i : D_i^i = 1\right\}\right| \leq \frac{sn}{2r} \cup \left|\left\{i : D_i^i = 1\right\}\right| \geq \frac{2sn}{r}\right]. \end{aligned}$$

The first probability can be controlled by appealing to Lemma 4.3.7, while the second one is bounded through Lemma C.3.8 above. Doing so, we obtain with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - \exp\left(-ns^2/(3r^2)\right)$

$$\left\|(\triangle XD)^{\backslash 1}\right\|_2 \leq 2\left\|\triangle X\right\|_\infty s\sqrt{\frac{\left(\frac{2sn}{r}\right)}{r}}.$$

This proves part 1.

Proof of Part 2: The proof of this is similar to that of part 1. Wlog, assume $p = 1$. We have:

$$\begin{aligned}\left\|X^{\backslash 1}\left(X^1\right)^\top\right\|_2 &= \left\|X^{\backslash 1}D\left(X^1\right)^\top\right\|_2 \\ &\leq \left\|(XD)^{\backslash 1}\right\|_2 \left\|\left(X^\top\right)_1\right\|_2 \\ &\leq \left\|(XD)^{\backslash 1}\right\|_2 \cdot 2\left(1+\left\|\triangle X\right\|_\infty\right)\sqrt{\frac{sn}{r}}.\end{aligned}$$

For the first term above, we have:

$$\left\|(XD)^{\backslash 1}\right\|_2 \leq \left\|(X^*D)^{\backslash 1}\right\|_2 + \left\|(\triangle XD)^{\backslash 1}\right\|_2$$

The second term in this decomposition was controlled above and the second one can be similarly bounded. Doing so, we obtain with probability at least $1 - r\exp\left(-\frac{Cn}{r}\right) - r\exp\left(-\frac{Cn}{rM^2}\right) - \exp\left(-ns^2/(3r^2)\right)$

$$\left\|(XD)^{\backslash 1}\right\|_2 \leq 2s\left(1+\left\|\triangle X\right\|_\infty\right)\sqrt{\frac{2sn}{r^3}},$$

This proves the lemma. □

Lemma C.3.10. *We have the following formula for matrix inversion:*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BMCA^{-1} & -A^{-1}BM \\ -MCA^{-1} & M \end{bmatrix},$$

where $M \stackrel{def}{=} (D - CA^{-1}B)^{-1}$ *is the Schur complement of* D *in the above matrix.*

Bibliography

[1] JP Abrahams and AGW Leslie. Methods used in the structure determination of bovine mitochondrial f1 atpase. *Acta Crystallographica Section D: Biological Crystallography*, 52(1):30–42, 1996.

[2] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory*, pages 123–137, 2014.

[3] S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *ArXiv e-prints*, August 2013.

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[5] Krishnakumar Balasubramanian, Kai Yu, and Guy Lebanon. Smooth sparse coding via marginal regression for learning sparse representations. In *ICML*, 2013.

[6] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. Hybrid projection–reflection method for phase retrieval. *JOSA A*, 20(6):1025–1034, 2003.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

[8] LM Bregman. Finding the common point of convex sets by the method of successive projection.(russian). In *Dokl. Akad. Nauk SSSR*, volume 162, pages 487–490, 1965.

[9] Yu M Bruck and LG Sodin. On the ambiguity of the image reconstruction problem. *Optics Communications*, 30(3):304–308, 1979.

[10] 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.

[11] Emmanuel J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589 – 592, 2008.

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

[13] Emmanuel J Candes and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *arXiv preprint arXiv:1208.6247*, 2012.

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

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

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

[17] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 2012.

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

[19] Emmanuel J Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[20] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

[21] Anwei Chai, Miguel Moscoso, and George Papanicolaou. Array imaging using intensity-only measurements. *Inverse Problems*, 27(1):015005, 2011.

[22] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[23] C. Chen, B. He, and X. Yuan. Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis*, 32(1):227–245, 2012.

[24] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118:887, 2005.

[25] J. C. Dainty and J. R. Fienup. Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application, ed. byH. Stark, Academic Press, San Diego*, pages 231–275, 1987.

[26] Hamootal Duadi, Ofer Margalit, Vicente Mico, José A Rodrigo, Tatiana Alieva, Javier Garcia, and Zeev Zalevsky. Digital holography and phase retrieval. *Source: Holography, Research and Technologies. InTech*, 2011.

[27] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[28] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *arXiv preprint arXiv:1211.0872*, 2012.

[29] Veit Elser. Phase retrieval by iterated projections. *JOSA A*, 20(1):40–55, 2003.

[30] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[31] James R Fienup et al. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[32] JR Fienup, JC Marron, TJ Schulz, and JH Seldin. Hubble space telescope characterized by using phase-retrieval algorithms. *Applied optics*, 32(10):1747–1767, 1993.

[33] Dennis Gabor. A new microscopic principle. *Nature*, 161(4098):777–778, 1948.

[34] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[35] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[36] Quan Geng, Huan Wang, and John Wright. On the local correctness of ℓ_1 minimization for dictionary learning. *arXiv preprint arXiv:1101:5672*, 2011. Preprint, URL:http://arxiv.org/abs/1101.5672.

[37] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

[38] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

[39] M Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 30(2):140–154, 1982.

[40] Christopher J Hillar and Friedrich T Sommer. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *arXiv preprint arXiv:1106.3616*, 2011.

[41] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S. Dhillon. Low rank modeling of signed networks. In *KDD*, pages 507–515, 2012.

[42] Norman E Hurt. *Phase Retrieval and Zero Crossings: Mathematical Methods in Image Reconstruction*, volume 52. Kluwer Academic Print on Demand, 2001.

[43] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1473–1477. IEEE, 2012.

[44] Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, pages 937–945, 2010.

[45] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 665–674. ACM, 2013.

[46] Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. In *ICML*, pages 1001–1008, 2011.

[47] Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume R Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, 2010.

[48] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[49] Raghunandan H. Keshavan. Efficient algorithms for collaborative filtering. Phd Thesis, Stanford University, 2012.

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

[51] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.*, 30(2):713–730, July 2008.

[52] Jingu Kim and Haesun Park. Sparse nonnegative matrix factorization for clustering. Technical Report GT-CSE-08-01, Georgia Institute of Technology, 2008.

[53] Yehuda Koren. The BellKor solution to the Netflix grand prize, 2009.

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

[55] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[56] Kiryung Lee and Yoram Bresler. Admira: atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

[57] Emmett N Leith and Juris Upatnieks. Reconstructed wavefronts and communication theory. *JOSA*, 52(10):1123–1128, 1962.

[58] Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[59] Wenbo V Li and Ang Wei. Gaussian integrals involving absolute value functions. In *Proceedings of the Conference in Luminy*, 2009.

[60] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *arXiv preprint arXiv:1209.4785*, 2012.

[61] Stefano Marchesini. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(1):011301–011301, 2007.

[62] Stefano Marchesini. Phase retrieval and saddle-point optimization. *JOSA A*, 24(10):3289–3296, 2007.

[63] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. *arXiv preprint arXiv:1209.0738*, 2012.

[64] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013.

[65] Nishant Mehta and Alexander G Gray. Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 36–44, 2013.

[66] Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit S. Dhillon. Rank minimization via online learning. In *ICML*, pages 656–663, 2008.

[67] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999.

[68] Jianwei Miao, Tetsuya Ishikawa, Bart Johnson, Erik H Anderson, Barry Lai, and Keith O Hodgson. High resolution 3d x-ray diffraction microscopy. *Physical review letters*, 89(8):088303, 2002.

[69] RP Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[70] DL Misell. A method for the solution of the phase problem in electron microscopy. *Journal of Physics D: Applied Physics*, 6(1):L6, 1973.

[71] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *arXiv preprint arXiv:1306.0160*, 2013.

[72] Henrik Ohlsson, Allen Y Yang, Roy Dong, and S Shankar Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint arXiv:1111.6323*, 2011.

[73] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[74] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

[75] Holger Rauhut. Compressive sensing, structured random matrices and recovery of functions in high dimensions. In *Oberwolfach Reports*, volume 7, pages 1990–1993, 2010.

[76] Benjamin Recht. A simple approach to matrix completion, 2009.

[77] 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.

[78] Jorge LC Sanz. Mathematical considerations for the problem of fourier transform phase retrieval from magnitude. *SIAM Journal on Applied Mathematics*, 45(4):651–664, 1985.

[79] Yoav Shechtman, Amir Beck, and Yonina C Eldar. Gespar: Efficient phase retrieval of sparse signals. *arXiv preprint arXiv:1301.1018*, 2013.

[80] Yoav Shechtman, Yonina C Eldar, Alexander Szameit, and Mordechai Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *arXiv preprint arXiv:1104.4406*, 2011.

[81] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proc. of Conf. on Learning Theory*, 2012.

[82] Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. Learning stable multilevel dictionaries for sparse representation of images. *ArXiv 1303.0448*, 2013.

[83] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, Dec. 2007.

[84] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[85] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[86] H Trussell and M Civanlar. The feasible solution in signal restoration. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(2):201–212, 1984.

[87] Daniel Vainsencher, Shie Mannor, and Alfred M Bruckstein. The sample complexity of dictionary learning. *The Journal of Machine Learning Research*, 12:3259–3281, 2011.

[88] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[89] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[90] Irène Waldspurger, Alexandre d'Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *arXiv preprint arXiv:1206.0102*, 2012.

[91] Wikipedia.

[92] Dan C Youla and Heywood Webb. Image restoration by the method of convex projections: Part 1theory. *Medical Imaging, IEEE Transactions on*, 1(2):81–94, 1982.

[93] W. I. Zangwill. *Nonlinear Programming: A Unified Approach*. Englewood Cliffs: Prentice-Hall, 1969.

[94] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2008.

[95] Yunhong Zhou, Dennis M. Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *AAIM*, pages 337–348, 2008.

[96] Hui Zou, Trevor Hastie, and Rob Tibshirani. Sparse principal component analysis. *JCGS*, 15(2):262–286, 2006.

Vita

Praneeth Kumar Netrapalli is a graduate student in the ECE department at UT Austin. He obtained B-Tech from IIT Bombay and M.S. from UT Austin, both in Electrical Engineering. His research focuses on designing and understanding practical algorithms for machine learning problems. His prior experience includes two years as a quantitative analyst at Goldman Sachs where he worked on pricing and evaluating risk on complex financial products.

Permanent address: praneethn@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.