Efficient Matrix Sensing Using Rank-1 Gaussian Measurements

Kai Zhong[1], Prateek Jain[2], and Inderjit S. Dhillon[1]

[1] University of Texas at Austin, USA
[2] Microsoft Research, India
zhongkai@ices.utexas.edu, prajain@microsoft.com, inderjit@cs.utexas.edu

Abstract. In this paper, we study the problem of low-rank matrix sensing where the goal is to reconstruct a matrix *exactly* using a small number of linear measurements. Existing methods for the problem either rely on measurement operators such as random element-wise sampling which cannot recover arbitrary low-rank matrices or require the measurement operator to satisfy the Restricted Isometry Property (RIP). However, RIP based linear operators are generally full rank and require large computation/storage cost for both measurement (encoding) as well as reconstruction (decoding).
In this paper, we propose simple rank-one Gaussian measurement operators for matrix sensing that are significantly less expensive in terms of memory and computation for both encoding and decoding. Moreover, we show that the matrix can be reconstructed *exactly* using a simple alternating minimization method as well as a nuclear-norm minimization method. Finally, we demonstrate the effectiveness of the measurement scheme vis-a-vis existing RIP based methods.

Keywords: Matrix Sensing, Matrix Completion, Inductive Learning, Alternating Minimization

1 Introduction

In this paper, we consider the matrix sensing problem, where the goal is to recover a low-rank matrix using a small number of linear measurements. The matrix sensing process contains two phases: a) compression phase (encoding), and b) reconstruction phase (decoding).

In the compression phase, a sketch/measurement of the given low-rank matrix is obtained by applying a linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$. That is, given a rank-$k$ matrix, $W_* \in \mathbb{R}^{d_1 \times d_2}$, its linear measurements are computed by:

$$\boldsymbol{b} = \mathcal{A}(W_*) = [\mathrm{Tr}(A_1^T W_*) \ \ \mathrm{Tr}(A_2^T W_*) \ \ \ldots \ \ \mathrm{Tr}(A_m^T W_*)]^T \ , \tag{1}$$

where $\{A_l \in \mathbb{R}^{d_1 \times d_2}\}_{l=1,2,\ldots,m}$ parameterize the linear operator $\mathcal{A}$ and Tr denotes the trace operator. Then, in the reconstruction phase, the underlying low-rank

matrix is reconstructed using the given measurements $\boldsymbol{b}$. That is, W_* is obtained by solving the following optimization problem:

$$\min_{W} \ \text{rank}(W) \quad \text{s.t.} \ \ \mathcal{A}(W) = \boldsymbol{b}\,. \tag{2}$$

The matrix sensing problem is a matrix generalization of the popular compressive sensing problem and has several real-world applications in the areas of system identification and control, Euclidean embedding, and computer vision (see [21] for a detailed list of references).

Naturally, the design of the measurement operator $\mathcal{A}$ is critical for the success of matrix sensing as it dictates cost of both the compression as well as the reconstruction phase. Most popular operators for this task come from a family of operators that satisfy a certain Restricted Isometry Property (RIP). However, these operators require each A_l, that parameterizes $\mathcal{A}$, to be a full rank matrix. That is, cost of compression as well as storage of $\mathcal{A}$ is $O(md_1d_2)$, which is infeasible for large matrices. Reconstruction of the low-rank matrix W_* is also expensive, requiring $O(md_1d_2 + d_1^2d_2)$ computation steps. Moreover, m is typically at least $O(k \cdot d \log(d))$ where $d = d_1 + d_2$. But, these operators are universal, i.e., every rank-k W_* can be compressed and recovered using such RIP based operators.

In this paper, we seek to reduce the computational/storage cost of such operators but at the cost of the universality property. That is, we propose to use simple rank-one operators, i.e., where each A_l is a rank one matrix. We show that using similar number of measurements as the RIP operators, i.e., $m = O(k \cdot d \log(d))$, we can recover a fixed rank W_* exactly.

In particular, we propose two measurement schemes: a) *rank-one independent* measurements, b) *rank-one dependent* measurements. In rank-one independent measurement, we use $A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T$, where $\boldsymbol{x}_l \in \mathbb{R}^{d_1}$, $\boldsymbol{y}_l \in \mathbb{R}^{d_2}$ are both sampled from zero mean sub-Gaussian product distributions, i.e., each element of $\boldsymbol{x}_l$ and $\boldsymbol{y}_l$ is sampled from a fixed zero-mean univariate sub-Gaussian distribution. Rank-one dependent measurements combine the above rank-one measurements with element-wise sampling, i.e., $A_l = \boldsymbol{x}_{i_l} \boldsymbol{y}_{j_l}^T$ where $\boldsymbol{x}_{i_l}, \boldsymbol{y}_{j_l}$ are sampled as above. Also, $(i_l, j_l) \in [n_1] \times [n_2]$ is a randomly sampled index, where $n_1 \geq d_1$, $n_2 \geq d_2$. These measurements can also be viewed as the inductive version of the matrix completion problem (see Section 3), where $\boldsymbol{x}_i$ represents features of the i-th user (row) and $\boldsymbol{y}_j$ represents features of the j-th movie (column). In fact, an additional contribution of our work is that we can show that the inductive matrix completion problem can also be solved using only $O(k(d_1+d_2)\log(d_1+d_2)\log(n_1+n_2))$ samples, as long as X, Y, W_* satisfy certain incoherence style assumptions (see Section 5 for more details)[3].

Next, we provide two recovery algorithms for both of the above measurement operators: a) alternating minimization, b) nuclear-norm minimization. Note that, in general, the recovery problem (2) is NP-hard to solve. However, for

[3] A preliminary version of this work appeared in [11]. Since the above mentioned work, several similar rank-one measurement operators have been studied [15,2,26].

Table 1. Comparison of sample complexity and computational complexity for different approaches and different measurements

METHODS		SAMPLE COMPLEXITY	COMPUTATIONAL COMPLEXITY
ALS	RANK-1 INDEP.	$O(k^4\beta^2 d\log^2 d)$	$O(kdm)$
	RANK-1 DEP.	$O(k^4\beta^2 d\log d)$	$O(dm + knd)$
	RIP	$O(k^4 d\log d)$	$O(d^2 m)$
NUCLEAR	RANK-1 INDEP.	$O(kd)$	$O(\hat{k}dm)$
	RANK-1 DEP.	$O(kd\log n\log d)$	$O(\hat{k}^2 m + \hat{k}nd)$
	RIP	$O(kd\log d)$	$O(d^2 m)$

the RIP based operators, both alternating minimization as well as nuclear norm minimization methods are known to solve the problem exactly in polynomial time [13]. Note that the existing analysis of both the methods crucially uses RIP and hence does not extend to the proposed operators.

We show that if $m = O(k^4 \cdot \beta^2 \cdot (d_1 + d_2)\log^2(d_1 + d_2))$, where β is the condition number of W_* then alternating minimization for the rank-one independent measurements recovers W_* in time $O(kdm)$, where $d = d_1 + d_2$. Similarly, if $m = O(k \cdot (d_1 + d_2) \cdot \log^2(d_1 + d_2))$ then the nuclear norm minimization based method recovers W_* in time $O(d^2 m)$ in the worst case. Note that alternating minimization has slightly higher sample complexity requirement but is significantly faster than the nuclear norm minimization method. Due to this, most practical low-rank estimation systems in fact use alternating minimization method for recovery. We obtain similar results for the rank-one dependent measurements.

We summarize the sample complexity and computational complexity for different approaches and different measurements in Table 1. In the table, ALS refers to alternating least squares, i.e., alternating minimization. For the symbols, $d = d_1 + d_2$, $n = n_1 + n_2$, and $\hat{k}$ is the maximum of rank estimate used as an input in the nuclear-norm solver [10][9]. This number can be as large as $\min\{d_1, d_2\}$ when we have no prior knowledge of the rank.

Paper Organization: We summarize related work in Section 2. In Section 3 we formally introduce the matrix sensing problem and our proposed rank-one measurement operators. In Section 4, we present the alternating minimization method for matrix reconstruction. We then present a *generic* analysis for alternating minimization when applied to the proposed rank-one measurement operators. We present the nuclear-norm minimization based method in Section 5 and present its analysis for the rank-one dependent measurements. Finally, we provide empirical validation of our methods in Section 6.

2 Related Work

Matrix Sensing: Matrix sensing[21][12][16] is a generalization of the popular compressive sensing problem for the sparse vectors and has applications in several domains such as control, vision etc. [21] introduced measurement operators that

satisfy RIP and showed that using only $O(kd \log d)$ measurements, a rank-k $W_* \in \mathbb{R}^{d_1 \times d_2}$ can be recovered. Recently, a set of universal Pauli measurements, used in quantum state tomography, have been shown to satisfy the RIP condition [18]. These measurement operators are Kronecker products of 2×2 matrices, thus, they have appealing computation and memory efficiency. In concurrent work, [15][2] also proposed an independent rank-one measurement using nuclear-norm minimization. In contrast, we use two different measurement operators and show that the popular alternating minimization method also solves the problem exactly.

Matrix Completion and Inductive Matrix Completion: Matrix completion [3][14][13] is a special case of rank-one matrix sensing problem when the operator takes a subset of the entries. However, to guarantee exact recovery, the target matrix has to satisfy the incoherence condition. Using our rank-one Gaussian operators, we don't require any condition on the target matrix. For inductive matrix completion (IMC), which is a generalization of matrix completion utilizing movies' and users' features, the authors of [23] provided the theoretical recovery guarantee for nuclear-norm minimization. In this paper, we show that IMC is equivalent to the matrix sensing problem using dependent rank-one measurements, and provide a similar result for nuclear-norm based methods for IMC but eliminate a "joint" incoherent condition on the rank-one measurements and an upper bound condition on the sample complexity. Moreover, we give a theoretical guarantee using alternating minimization methods.

Alternating Minimization: Although nuclear-norm minimization enjoys nice recovery guarantees, it usually doesn't scale well. In practice, alternating minimization is employed to solve problem (2) by assuming the rank is known. Alternating minimization solves two least square problems alternatively in each iteration, thus is very computationally efficient[24]. Although widely used in practice, its theoretical guarantees are relatively less understood due to non-convexity. [13] first showed optimality of alternating minimization in the matrix sensing/low-rank estimation setting under the RIP setting. Subsequently, several other papers have also shown global convergence guarantees for alternating minimization, e.g. matrix completion [8][7], robust PCA [19] and dictionary learning [1]. In this paper, we provide a *generic* analysis for alternating minimization applied to the proposed rank-one measurement operators. Our results distill out certain key problem specific properties that would imply global optimality of alternating minimization. We then show that the rank-one Gaussian measurements satisfy those properties.

3 Problem Formulation

The goal of matrix sensing is to design a linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ and a recovery algorithm so that a low-rank matrix $W_* \in \mathbb{R}^{d_1 \times d_2}$ can be recovered exactly using $\mathcal{A}(W_*)$. In this paper, we focus on rank-one measurement operators, $A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T$, and call such problems as *Low-Rank matrix estimation using Rank One Measurements* (**LRROM**): recover the rank-k matrix $W_* \in \mathbb{R}^{d_1 \times d_2}$

by using rank-1 measurements of the form:

$$\boldsymbol{b} = [\boldsymbol{x}_1^T W_* \boldsymbol{y}_1 \quad \boldsymbol{x}_2^T W_* \boldsymbol{y}_2 \quad \dots \quad \boldsymbol{x}_m^T W_* \boldsymbol{y}_m]^T, \tag{3}$$

where $\boldsymbol{x}_l, \boldsymbol{y}_l$ are "feature" vectors and are provided along with the measurements $\boldsymbol{b}$.

We propose two different kinds of rank-one measurement operators based on Gaussian distribution.

3.1 Rank-one Independent Gaussian Operator

Our first measurement operator is a simple rank-one Gaussian operator, $\mathcal{A}_{GI} = [A_1 \dots A_m]$, where, $\{A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T\}_{l=1,2,\dots,m}$, and $\boldsymbol{x}_l$, $\boldsymbol{y}_l$ are sampled i.i.d. from spherical Gaussian distribution.

3.2 Rank-one Dependent Gaussian Operator

Our second operator can introduce certain "dependencies" in our measurement and has in fact interesting connections to the matrix completion problem. We provide the operator as well as the connection to matrix completion in this sub-section. To generate the rank-one dependent Gaussian operator, we first sample two Gaussian matrices $X \in \mathbb{R}^{n_1 \times d_1}$ and $Y \in \mathbb{R}^{n_2 \times d_2}$, where each entry of both X, Y is sampled independently from Gaussian distribution and $n_1 \geq Cd_1$, $n_2 \geq Cd_2$ for a global constant $C \geq 1$. Then, the Gaussian dependent operator $\mathcal{A}_{GD} = [A_1, \dots A_m]$ where $\{A_l = \boldsymbol{x}_{i_l} \boldsymbol{y}_{i_l}^T\}_{(i_l, j_l) \in \Omega}$. Here $\boldsymbol{x}_i^T$ is the i-th row of X and $\boldsymbol{y}_j^T$ is the j-th row of Y. Ω is a uniformly random subset of $[n_1] \times [n_2]$ such that $E[|\Omega|] = m$. For simplicity, we assume that each entry $(i_l, j_l) \in [n_1] \times [n_2]$ is sampled i.i.d. with probability $p = m/(n_1 \times n_2)$. Therefore, the measurements using the above operator are given by: $b_l = \boldsymbol{x}_{i_l}^T W \boldsymbol{y}_{j_l}, (i_l, j_l) \in \Omega$.

Connections to Inductive Matrix Completion (IMC): Note that the above measurements are inspired by matrix completion style sampling operator. However, here we first multiply W with X, Y and then sample the obtained matrix XWY^T. In the domain of recommendation systems (say user-movie system), the corresponding reconstruction problem can also be thought as the inductive matrix completion problem. That is, let there be n_1 users, n_2 movies, X represents user features, and Y represents the movie features. Then, the true ratings matrix is given by $R = XWY^T \in \mathbb{R}^{n_1 \times n_2}$.

That is, given the user/movie feature vectors $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ for $i = 1, 2, ..., n_1$ and $\boldsymbol{y}_j \in \mathbb{R}^{d_2}$ for $j = 1, 2, ..., n_2$, our goal is to recover a rank-k matrix W_* of size $d_1 \times d_2$ from a few observed entries $R_{ij} = \boldsymbol{x}_i^T W_* \boldsymbol{y}_j$, for $(i, j) \in \Omega \subset [n_1] \times [n_2]$. Because of the equivalence between the dependent rank-one measurements and the entries of the rating matrix, in the rest of the paper, we will use $\{R_{ij}\}_{(i,j) \in \Omega}$ as the dependent rank-one measurements.

Now, if we can reconstruct W_* from the above measurements, we can predict ratings *inductively* for new users/movies, provided their feature vectors are given.

Algorithm 1 AltMin-LRROM : Alternating Minimization for LRROM

1: **Input**: Measurements: $\boldsymbol{b}_{all}$, Measurement matrices: $\mathcal{A}_{all}$, Number of iterations: H
2: Divide $(\mathcal{A}_{all}, \boldsymbol{b}_{all})$ into $2H+1$ sets (each of size m) with h-th set being $\mathcal{A}^h = \{A_1^h, A_2^h, \ldots, A_m^h\}$ and $\boldsymbol{b}^h = [b_1^h \ b_2^h \ \ldots \ b_m^h]^T$
3: **Initialization**: U_0 =top-k left singular vectors of $\frac{1}{m}\sum_{l=1}^m b_l^0 A_l^0$
4: **for** $h = 0$ to $H-1$ **do**
5: $\quad b \leftarrow b^{2h+1}, \mathcal{A} \leftarrow \mathcal{A}^{2h+1}$
6: $\quad \widehat{V}_{h+1} \leftarrow \text{argmin}_{V \in \mathbb{R}^{d_2 \times k}} \sum_l (b_l - \boldsymbol{x}_l^T U_h V^T \boldsymbol{y}_l)^2$
7: $\quad V_{h+1} = QR(\widehat{V}_{h+1})$ //orthonormalization of $\widehat{V}_{h+1}$
8: $\quad b \leftarrow b^{2h+2}, \mathcal{A} \leftarrow \mathcal{A}^{2h+2}$
9: $\quad \widehat{U}_{h+1} \leftarrow \text{argmin}_{U \in \mathbb{R}^{d_1 \times k}} \ \sum_l (b_l - \boldsymbol{x}_l^T U V_{h+1}^T \boldsymbol{y}_l)^2$
10: $\quad U_{h+1} = QR(\widehat{U}_{h+1})$ //orthonormalization of $\widehat{U}_{h+1}$
11: **end for**
12: **Output**: $W_H = U_H (\widehat{V}_H)^T$

Hence, our reconstruction procedure also solves the IMC problem. However, there is a key difference: in matrix sensing, we can select X, Y according to our convenience, while in IMC, X and Y are provided a priori. But, for general X, Y one cannot solve the problem because if say $R = XW_*Y^T$ is a 1-sparse matrix, then W_* cannot be reconstructed even with a large number of samples.

Interestingly, our proof for reconstruction using nuclear-norm based method does not require Gaussian X, Y. Instead, we can distill out two key properties of R, X, Y ensuring that using only $O(k(d_1+d_2)\log(d_1+d_2)\log(n_1+n_2))$ samples, we can reconstruct W_*. Note that a direct application of matrix completion results [3][4] would require $O(k(n_1+n_2)\log(n_1+n_2))$ samples which can be much large if $n_1 \gg d_1$ or $n_2 \gg d_2$. See Section 5 for more details on the assumptions that we require for the nuclear-norm minimization method to solve the IMC problem exactly.

4 Rank-one Matrix Sensing via Alternating Minimization

We now present the alternating minimization approach for solving the reconstruction problem (2) with rank-one measurements (3). Since W to be recovered is restricted to have at most rank-k, (2) can be reformulated as the following non-convex optimization problem:

$$\min_{U \in \mathbb{R}^{d_1 \times k}, V \in \mathbb{R}^{d_2 \times k}} \sum_{l=1}^m (b_l - \boldsymbol{x}_l^T U V^T \boldsymbol{y}_l)^2 \ . \tag{4}$$

Alternating minimization is an iterative procedure that alternately optimizes for U and V while keeping the other factor fixed. As optimizing for U (or V) involves solving just a least squares problem, so each individual iteration of the algorithm is linear in matrix dimensions. For the rank-one measurement operator, we use a particular initialization method to initialize U (see line 3 of Algorithm 1). See Algorithm 1 for a pseudo-code of the algorithm.

4.1 General Theoretical Guarantee for Alternating Minimization

As mentioned above, (4) is non-convex in U, V and hence standard analysis would only ensure convergence to a local minima. However, [13] recently showed that the alternating minimization method in fact converges to the global minima of two low-rank estimation problems: matrix sensing with RIP matrices and matrix completion.

The rank-one operator given above does not satisfy RIP (see Definition 1), even when the vectors $\boldsymbol{x}_l, \boldsymbol{y}_l$ are sampled from the normal distribution (see Claim 4.2). Furthermore, each measurement need not reveal exactly one entry of W_* as in the case of matrix completion. Hence, the proof of [13] does not apply directly. However, inspired by the proof of [13], we distill out three key properties that the operator should satisfy, so that alternating minimization would converge to the global optimum.

Theorem 1. *Let $W_* = U_* \Sigma_* V_*^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-k matrix with k-singular values $\sigma_*^1 \geq \sigma_*^2 \cdots \geq \sigma_*^k$. Also, let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement operator parameterized by m matrices, i.e., $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ where $A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T$. Let $\mathcal{A}(W)$ be as given by* (1).

Now, let $\mathcal{A}$ satisfy the following properties with parameter $\delta = \frac{1}{k^{3/2} \cdot \beta \cdot 100}$ ($\beta = \sigma_*^1 / \sigma_*^k$):

1. **Initialization***: $\| \frac{1}{m} \sum_l b_l A_l - W_* \|_2 \leq \|W_*\|_2 \cdot \delta$.*
2. **Concentration of operators** *B_x, B_y: Let $B_x = \frac{1}{m} \sum_{l=1}^m (\boldsymbol{y}_l^T \boldsymbol{v})^2 \boldsymbol{x}_l \boldsymbol{x}_l^T$ and $B_y = \frac{1}{m} \sum_{l=1}^m (\boldsymbol{x}_l^T \boldsymbol{u})^2 \boldsymbol{y}_l \boldsymbol{y}_l^T$, where $\boldsymbol{u} \in \mathbb{R}^{d_1}, \boldsymbol{v} \in \mathbb{R}^{d_2}$ are two unit vectors that are* independent *of randomness in $\boldsymbol{x}_l, \boldsymbol{y}_l$, $\forall i$. Then the following holds: $\|B_x - I\|_2 \leq \delta$ and $\|B_y - I\|_2 \leq \delta$.*
3. **Concentration of operators** *G_x, G_y: Let $G_x = \frac{1}{m} \sum_l (\boldsymbol{y}_l^T \boldsymbol{v})(\boldsymbol{y}_l \boldsymbol{v}_\perp) \boldsymbol{x}_l \boldsymbol{x}_l^T$, $G_y = \frac{1}{m} \sum_l (\boldsymbol{x}_l^T \boldsymbol{u})(\boldsymbol{u}_\perp^T \boldsymbol{x}_l) \boldsymbol{y}_l \boldsymbol{y}_l^T$, where $\boldsymbol{u}, \boldsymbol{u}_\perp \in \mathbb{R}^{d_1}$, $\boldsymbol{v}, \boldsymbol{v}_\perp \in \mathbb{R}^{d_2}$ are unit vectors, s.t., $\boldsymbol{u}^T \boldsymbol{u}_\perp = 0$ and $\boldsymbol{v}^T \boldsymbol{v}_\perp = 0$. Furthermore, let $\boldsymbol{u}, \boldsymbol{u}_\perp, \boldsymbol{v}, \boldsymbol{v}_\perp$ be independent of randomness in $\boldsymbol{x}_l, \boldsymbol{y}_l, \forall i$. Then, $\|G_x\|_2 \leq \delta$ and $\|G_y\|_2 \leq \delta$.*

Then, after H-iterations of the alternating minimization method (Algorithm 1), we obtain $W_H = U_H V_H^T$ s.t., $\|W_H - W_*\|_2 \leq \epsilon$, where $H \leq 100 \log(\|W_*\|_F / \epsilon)$.

See Appendix A for a detailed proof. Note that we require intermediate vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{u}_\perp, \boldsymbol{v}_\perp$ to be independent of randomness in A_l's. Hence, we partition $\mathcal{A}_{all}$ into $2H+1$ partitions and at each step $(\mathcal{A}^h, \boldsymbol{b}^h)$ and $(\mathcal{A}^{h+1}, \boldsymbol{b}^{h+1})$ are supplied to the algorithm. This implies that the measurement complexity of the algorithm is given by $m \cdot H = m \log(\|W_*\|_F / \epsilon)$. That is, given $O(m \log(\|(d_1 + d_2) W_*\|_F))$ samples, we can estimate matrix W_H, s.t., $\|W_H - W_*\|_2 \leq \frac{1}{(d_1+d_2)^c}$, where $c > 0$ is any constant.

4.2 Independent Gaussian Measurements

In this subsection, we consider the rank-one independent measurement operator $\mathcal{A}_{GI}$ specified in Section 3. Now, for this operator $\mathcal{A}_{GI}$, we show that if $m =$

$O(k^4\beta^2 \cdot (d_1+d_2) \cdot \log^2(d_1+d_2))$, then w.p. $\geq 1 - 1/(d_1+d_2)^{100}$, *any* fixed rank-$k$ matrix W_* can be recovered by AltMin-LRROM (Algorithm 1). Here $\beta = \sigma_*^1/\sigma_*^k$ is the condition number of W_*. That is, using nearly linear number of measurements in d_1, d_2, one can exactly recover the $d_1 \times d_2$ rank-k matrix W_*.

As mentioned in the previous section, the existing matrix sensing results typically assume that the measurement operator $\mathcal{A}$ satisfies the Restricted Isometry Property (RIP) defined below:

Definition 1. *A linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ satisfies RIP iff, for $\forall W$ s.t. $rank(W) \leq k$, the following holds:*

$$(1-\delta_k)\|W\|_F^2 \leq \|\mathcal{A}(W)\|_F^2 \leq (1+\delta_k)\|W\|_F^2 \ ,$$

where $\delta_k > 0$ is a constant dependent only on k.

Naturally, this begs the question whether we can show that our rank-1 measurement operator $\mathcal{A}_{GI}$ satisfies RIP, so that the existing analysis for RIP based low-rank matrix sensing can be used [13]. We answer this question in the negative, i.e., for $m = O((d_1+d_2)\log(d_1+d_2))$, $\mathcal{A}_{GI}$ does not satisfy RIP even for rank-1 matrices (with high probability):

Claim. 4.2. Let $\mathcal{A}_{GI} = \{A_1, A_2, \dots A_m\}$ be a measurement operator with each $A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T$, where $\boldsymbol{x}_l \in \mathbb{R}^{d_1} \sim \mathcal{N}(0, I)$, $\boldsymbol{y}_l \in \mathbb{R}^{d_2} \sim \mathcal{N}(0, I), 1 \leq l \leq m$. Let $m = O((d_1+d_2)\log^c(d_1+d_2))$, for any constant $c > 0$. Then, with probability at least $1 - 1/m^{10}$, $\mathcal{A}_{GI}$ does not satisfy RIP for rank-1 matrices with a constant δ.

See Appendix B for a detailed proof of the above claim. Now, even though $\mathcal{A}_{GI}$ does not satisfy RIP, we can still show that $\mathcal{A}_{GI}$ satisfies the three properties mentioned in Theorem 1. and hence we can use Theorem 1 to obtain the exact recovery result.

Theorem 2 (Rank-One Independent Gaussian Measurements using ALS). *Let $\mathcal{A}_{GI} = \{A_1, A_2, \dots A_m\}$ be a measurement operator with each $A_l = \boldsymbol{x}_l \boldsymbol{y}_l^T$, where $\boldsymbol{x}_l \in \mathbb{R}^{d_1} \sim \mathcal{N}(0, I)$, $\boldsymbol{y}_l \in \mathbb{R}^{d_2} \sim \mathcal{N}(0, I), 1 \leq l \leq m$. Let $m = O(k^4\beta^2(d_1+d_2)\log^2(d_1+d_2))$. Then, Property 1, 2, 3 required by Theorem 1 are satisfied with probability at least $1 - 1/(d_1+d_2)^{100}$.*

Proof. Here, we provide a brief proof sketch. See Appendix B for a detailed proof.

Initialization: Note that,

$$\frac{1}{m}\sum_{l=1}^m b_l \boldsymbol{x}_l \boldsymbol{y}_l^T = \frac{1}{m}\sum_{l=1}^m \boldsymbol{x}_l \boldsymbol{x}_l^T U_* \Sigma_* V_*^T \boldsymbol{y}_l \boldsymbol{y}_l^T = \frac{1}{m}\sum_{l=1}^m Z_l \ ,$$

where $Z_l = \boldsymbol{x}_l \boldsymbol{x}_l^T U_* \Sigma_* V_*^T \boldsymbol{y}_l \boldsymbol{y}_l^T$. Note that $\mathbb{E}[Z_l] = U_* \Sigma_* V_*^T$. Hence, to prove the initialization result, we need a tail bound for sums of random matrices. To this end, we use Theorem 6 in [22]. However, Theorem 6 in [22] requires a bounded

random variable while Z_l is an unbounded variable. We handle this issue by clipping Z_l to ensure that its spectral norm is always bounded. Furthermore, by using properties of normal distribution, we can ensure that w.p. $\geq 1 - 1/m^3$, Z_l's do not require clipping and the new "clipped" variables converge to nearly the same quantity as the original "non-clipped" Z_l's. See Appendix B for more details.

Concentration of B_x, B_y, G_x, G_y: Consider $G_x = \frac{1}{m}\sum_{l=1}^{m} \boldsymbol{x}_l \boldsymbol{x}_l^T \boldsymbol{y}_l^T \boldsymbol{v}\boldsymbol{v}_\perp{}^T \boldsymbol{y}_l$. As, $\boldsymbol{v}, \boldsymbol{v}_\perp$ are unit-norm vectors, $\boldsymbol{y}_l^T \boldsymbol{v} \sim \mathcal{N}(0,1)$ and $\boldsymbol{v}_\perp^T \boldsymbol{x}_l \sim \mathcal{N}(0,1)$. Also, since $\boldsymbol{v}$ and $\boldsymbol{v}_\perp$ are orthogonal, $\boldsymbol{y}_l^T \boldsymbol{v}$ and $\boldsymbol{v}_\perp^T \boldsymbol{y}_l$ are independent variables. Hence, $G_x = \frac{1}{m}\sum_{l=1}^{m} Z_l$ where $\mathbb{E}[Z_l] = 0$. Here again, we apply Theorem 6 in [22] after using a clipping argument. We can obtain the required bounds for B_x, B_y, G_y also in a similar manner.

Note that the clipping procedure ensures that Z_l's don't need to be clipped with probability $\geq 1 - 1/m^3$ only. That is, we cannot apply the *union* bound to ensure that the concentration result holds *for all* $\boldsymbol{v}, \boldsymbol{v}_\perp$. Hence, we need a fresh set of measurements after each iteration to ensure concentration.

Global optimality of the rate of convergence of the Alternating Minimization procedure for this problem now follows directly by using Theorem 1. We would like to note that while the above result shows that the $\mathcal{A}_{GI}$ operator is almost as powerful as the RIP based operators for matrix sensing, there is one critical drawback: while RIP based operators are universal that is they can be used to recover any rank-k W_*, $\mathcal{A}_{GI}$ needs to be resampled for each W_*. We believe that the two operators are at two extreme ends of randomness vs universality trade-off and intermediate operators with higher success probability but using larger number of random bits should be possible.

4.3 Dependent Gaussian Measurements

For the dependent Gaussian measurements, the alternating minimization formulation is given by:

$$\min_{U \in \mathbb{R}^{d_1 \times k}, V \in \mathbb{R}^{d_2 \times k}} \sum_{(i,j) \in \Omega} (\boldsymbol{x}_i^T U V^T \boldsymbol{y}_j - R_{ij})^2 \,. \tag{5}$$

Here again, we can solve the problem by alternatively optimizing for U and V. Later in Section 4.4, we show that using such dependent measurements leads to a faster recovery algorithm when compared to the recovery algorithm for independent measurements.

Note that both the measurement matrices X, Y can be thought of as orthonormal matrices. The reason being, $XW_*Y^T = U_X \Sigma_X V_X^T W_* V_Y \Sigma_Y U_Y^T$, where $X = U_X \Sigma_X V_X^T$ and $Y = U_Y \Sigma_Y V_Y^T$ is the SVD of X, Y respectively. Hence, $R = XW_*Y^T = U_X(\Sigma_X V_X^T W_* V_Y \Sigma_Y) U_Y^T$. Now U_X, U_Y can be treated as the true "X", "Y" matrices and $W_* \leftarrow (\Sigma_X V_X^T W_* V_Y \Sigma_Y)$ can be thought of as W_*. Then the "true" W_* can be recovered using the obtained W_H as: $W_H \leftarrow V_X \Sigma_X^{-1} W_H \Sigma_Y^{-1} V_Y^T$. We also note that such a transformation implies that the

condition number of R and that of $W_* \leftarrow (\Sigma_X V_X^T W_* V_Y \Sigma_Y)$ are exactly the same.

Similar to the previous section, we utilize our general theorem for optimality of the LRROM problem to provide a convergence analysis of rank-one Gaussian dependent operators $\mathcal{A}_{GD}$. We prove if X and Y are random orthogonal matrices, defined in [3], the above mentioned dependent measurement operator $\mathcal{A}_{GD}$ generated from X, Y also satisfies Properties 1, 2, 3 in Theorem 1. Hence, AltMin-LRROM (Algorithm 1) converges to the global optimum in $O(\log(\|W_*\|_F/\epsilon))$ iterations.

Theorem 3 (Rank-One Dependent Gaussian Measurements using ALS). *Let $X_0 \in \mathbb{R}^{n_1 \times d_1}$ and $Y_0 \in \mathbb{R}^{n_2 \times d_2}$ be Gaussian matrices, i.e. every entry is sampled i.i.d from $\mathcal{N}(0,1)$. Let $X_0 = X\Sigma_X V_X^T$ and $Y_0 = Y\Sigma_Y V_Y^T$ be the thin SVD of X_0 and Y_0 respectively. Then the rank-one dependent operator $\mathcal{A}_{GD}$ formed by X, Y with $m \geq O(k^4\beta^2(d_1+d_2)\log(d_1+d_2))$ satisfy Property 1,2,3 required by Theorem 1 with high probability.*

See Appendix C for a detailed proof. Interestingly, our proof does not require X, Y to be Gaussian. It instead utilizes only two key properties about X, Y which are given by:

1. **Incoherence:** For some constant μ, c,

$$\max_{i\in[n]} \|\boldsymbol{x}_i\|_2^2 \leq \frac{\mu\bar{d}}{n}, \tag{6}$$

   where $\bar{d} = \max(d, \log n)$
2. **Averaging Property:** For H different orthogonal matrices $U_h \in \mathbb{R}^{d\times k}$, $h = 1, 2, \ldots, H$, the following hold for these U_h's,

$$\max_{i\in[n]} \|U_h^T \boldsymbol{x}_i\|_2^2 \leq \frac{\mu_0\bar{k}}{n}, \tag{7}$$

   where μ_0, c are some constants and $\bar{k} = \max(k, \log n)$.

Hence, the above theorem can be easily generalized to solve the inductive matrix completion problem (IMC), i.e., solve (5) for arbitrary X, Y. Moreover, the sample complexity of the analysis would be nearly in $(d_1 + d_2)$, instead of $(n_1 + n_2)$ samples required by the standard matrix completion methods.

The following lemma shows that the above two properties hold w.h.p. for random orthogonal matrices .

Lemma 1. *If $X \in \mathbb{R}^{n\times d}$ is a random orthogonal matrix, then both Incoherence and Averaging properties are satisfied with probability $\geq 1 - (c/n^3)\log n$, where c is a constant.*

The proof of Lemma 1 can be found in Appendix C.

4.4 Computational Complexity for Alternating Minimization

In this subsection, we briefly discuss the computational complexity for Algorithm 1. For simplicity, we set $d = d_1 + d_2$ and $n = n_1 + n_2$, and in practical implementation, we don't divide the measurements and use the whole measurement operator $\mathcal{A}$ for every iteration. The most time-consuming part of Algorithm 1 is the step for solving the least square problem. Given $U = U_h$, V can be obtained by solving the following linear system,

$$\sum_{l=1}^{m} \langle V, A_l^T U_h \rangle A_l U_h = \sum_{l=1}^{m} b_l A_l^T U_h \ . \tag{8}$$

The dimension of this linear system is kd, which could be large, thus we use conjugate gradient (CG) method to solve it. In each CG iteration, different measurement operators have different computational complexity. For RIP-based full-rank operators, the computational complexity for each CG step is $O(d^2 m)$ while it is $O(kdm)$ for rank-one independent operators. However, for rank-one dependent operators, using techniques introduced in [25], we can reduce the per iteration complexity to be $O(kdn + md)$. Furthermore, if $n = d$, the computational complexity of dependent operators is only $O(kd^2 + md)$, which is better than the complexity of rank-one independent operators in an order of k.

5 Rank-one Matrix Sensing via Nuclear norm Minimization

In this section, we consider solving LRROM by nuclear norm relaxation. We first note that using nuclear norm relaxation, [15] provided the analysis for independent rank-one measurement operators when the underlying matrix is Hermitian. It can be shown that non-Hermitian matrices problem can be transformed to Hermitian cases. Their proof uses the bowling scheme and only requires $O(k(d_1+d_2))$ measurements for Gaussian case and $O(k(d_1+d_2)\log(d_1+d_2)\log(n_1+n_2))$ measurements for 4-designs case. In this paper, we consider dependent measurement operators which have a similar sample complexity as the independent operators, but less computational complexity and memory footprint than those of independent ones.

The nuclear norm minimization using rank-one dependent Gaussian operator is of form,

$$\begin{aligned} &\min \ \|W\|_* \\ &\text{s.t. } \boldsymbol{x}_i^T W \boldsymbol{y}_j = R_{ij},\ (i,j) \in \Omega \ . \end{aligned} \tag{9}$$

(9) can be solved exactly by semi-definite programming or approximated by

$$\min_W \sum_{(i,j)\in\Omega} (\boldsymbol{x}_i^T W \boldsymbol{y}_j - R_{ij})^2 + \lambda \|W\|_* \ , \tag{10}$$

where λ is a constant and can be viewed as a Lagrange multiplier.

5.1 Recovery Guarantee for Nuclear-norm Minimization

In this subsection, we show that using rank-one dependent Gaussian operators, the nuclear-norm minimization can recover *any* low-rank matrix exactly with $O(k(d_1+d_2)\log(d_1+d_2)\log(n_1+n_2))$ measurements. We also generalize the theorem to the IMC problem where the feature matrices X and Y can be arbitrary instead of being Gaussian. We show that as long as X, W_*, Y satisfy certain incoherence style properties, the nuclear norm minimization can guarantee exact recovery using only $O(k(d_1+d_2)\log(d_1+d_2))$ samples.

We first provide recovery guarantees for our rank-one dependent operator, i.e., when X, Y are sampled from the Gaussian distribution.

Theorem 4 (Rank-one Dependent Gaussian Measurements using Nuclear-norm Minimization). *Let $W_* = U_*\Sigma_*V_*^T \in \mathbb{R}^{d_1\times d_2}$ be a rank-k matrix. Let $X \in \mathbb{R}^{n_1\times d_1}$ and $Y \in \mathbb{R}^{n_2\times d_2}$ be random orthogonal matrices. Assume each $(i,j) \in \Omega$ is sampled from $[n_1]\times[n_2]$ i.i.d.. Then if $m = |\Omega| \geq O(k(d_1+d_2)\log(d_1+d_2)\log(n_1+n_2))$, the minimizer to the problem (9) is unique and equal to W_* with probability at least $1 - c_1(d_1+d_2)^{-c_2}$, where c_1 and c_2 are universal constants.*

The above theorem is a directly corollary of Theorem 5 combined with Lemma 1. Lemma 1 shows that random orthonormal matrices X, Y (can be generated using Gaussian matrices as stated in Theorem 2) satisfy the requirements of Theorem 5.

Nuclear-norm minimization approach for inductive matrix completion (9) has also been studied by [23]. However, their recovery guarantee holds under a much more restrictive set of assumptions on X, W_*, Y and in fact requires that the number of samples is not only lower bounded by certain quantity but also upper bounded by some other quantity. Our general analysis below doesn't rely on this upper bound. Moreover, their proof also requires a "joint" incoherent condition, i.e., an upper bound on $\max_{i,j}|\boldsymbol{x}_i^T U_* V_*^T \boldsymbol{y}_j|$ which is not required by our method; to this end, we use a technique introduced by [5] to bound an $\ell_{\infty,2}$-norm.

Theorem 5 (Inductive Matrix Completion using Nuclear-norm Minimization). *Let $W_* = U_*\Sigma_*V_*^T \in \mathbb{R}^{d_1\times d_2}$ be a rank-k matrix. Assume X, Y are orthogonal matrices, and satisfy the following conditions with respect to W_* for some constant μ and μ_0,*

$$\textbf{C1.}\ \max_{i\in[n_1]}\|\boldsymbol{x}_i\|_2^2 \leq \frac{\mu d_1}{n_1},\ \max_{j\in[n_2]}\|\boldsymbol{y}_j\|_2^2 \leq \frac{\mu d_2}{n_2}\ ,$$

$$\textbf{C2.}\ \max_{i\in[n_1]}\|U_*^T\boldsymbol{x}_i\|_2^2 \leq \frac{\mu_0 k}{n_1},\ \max_{j\in[n_2]}\|V_*^T\boldsymbol{y}_j\|_2^2 \leq \frac{\mu_0 k}{n_2}\ .$$

Then if each observed entry $(i,j) \in \Omega$ is sampled from $[n_1]\times[n_2]$ i.i.d. with probability p,

$$p \geq \max\left\{\frac{c_0\mu_0\mu k d\log(d)\log(n)}{n_1 n_2}, \frac{1}{\min\{n_1,n_2\}^{10}}\right\}, \tag{11}$$

the minimizer to the problem (9) *is unique and equal to* W_* *with probability at least* $1 - c_1 d^{-c_2}$*, where* c_0*,* c_1 *and* c_2 *are universal constants,* $d = d_1 + d_2$ *and* $n = n_1 + n_2$*.*

Note that the first condition **C1** is actually the incoherence condition on X, Y, while the second one **C2** is the incoherence of XU_*, YV_*. Additionally, **C2** is weaker than the *Averaging* property in Lemma 1, as it only asks for one U_* rather than H different U_h's to satisfy the property.

Proof. We follow the popular proof ideas used by [3][20], that is, finding a dual feasible solution for (9) to certify the uniqueness of the minimizer of (9). Unlike the analysis in [23], we build our dual certificate in the $\mathbb{R}^{d_1 \times d_2}$ matrix space rather than the $\mathbb{R}^{n_1 \times n_2}$ space. This choice makes it easy to follow the analysis in standard matrix completion problem. In Proposition 1 in Appendix D, we give certain conditions the dual certificate should satisfy for the uniqueness. Then we apply golfing scheme [6] to find such a certificate. When building the dual certificate, we use an $\ell_{\infty,2}$-norm adapted from [5]. This enables us to discard the assumption of "joint" incoherence. The details can be found in Appendix D.

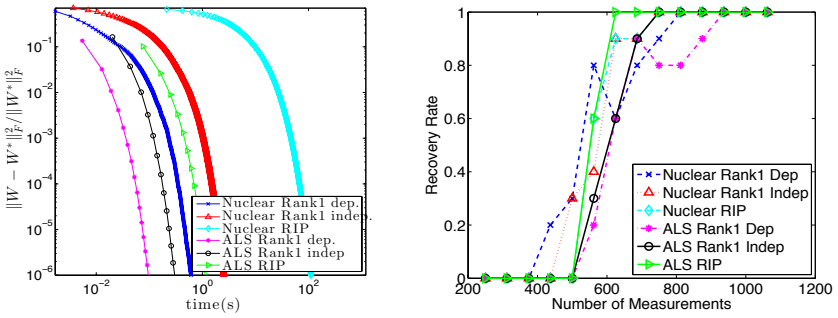
5.2 Computational Complexity for Nuclear-norm minimization

The optimization for nuclear-norm formulation is much more complex. Recently [10] proposed an active subspace method to solve Problem (10). The computational bottleneck is the approxSVD step and the inner problem step, both of which involve calculating a similar equation as shown on the left hand side of Eq (8). However, the rank of U or V is not fixed in each iteration as that of ALS, and in the worst case, it can be as large as $\min\{d_1, d_2\}$. The computational complexity for this basic operation is shown in Table 1.

6 Experiments

In this section, we demonstrate empirically that our Gaussian rank-one linear operators are significantly more efficient for matrix sensing than the existing RIP based measurement operators. In particular, we apply the two recovery methods namely alternating minimization (ALS) and nuclear norm minimization (Nuclear) to the measurements obtained using three different operators: rank-one independent (Rank1 Indep), rank-one dependent (Rank1 Dep), and a RIP based operator generated using random Gaussian matrices (RIP).

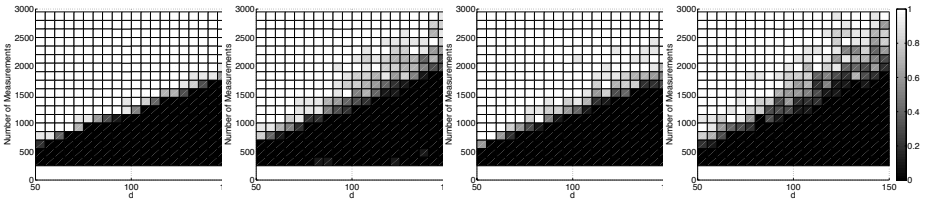
The experiments are conducted on Matlab and the nuclear-norm solver is adapted from the code by [10]. We first generated a random rank-5 signal $W_* \in \mathbb{R}^{50 \times 50}$, and compute $m = 1000$ measurements using different measurement operators. Here, we fix a small $\lambda = 10^{-6}$ for solving Eq (10) in order to exactly recover the matrix. And we set the maximum possible rank $\hat{k} = k$ as the input of the nuclear-norm solver. Figure 1a plots the relative error in recovery, $err = \|W - W^*\|_F^2 / \|W^*\|_F^2$, against computational time required by each method. Clearly, recovery using rank-one measurements requires significantly



(a) Relative error in recovery v.s. computation time

(b) Recovery rate v.s. number of measurements

Fig. 1. Comparison of computational complexity and measurement complexity for different approaches and different operators



Indep. ALS Dep. ALS Indep. Nuclear Dep. Nuclear

Fig. 2. Recovery rate for different matrix dimension d (x-axis) and different number of measurements m (y-axis). The color reflects the recovery rate scaled from 0 to 1. The white color indicates perfect recovery, while the black color denotes failure in all the experiments.

less time compared to the RIP based operator. Moreover, ALS in general seems to be significantly faster than Nuclear methods.

Next, we compare the measurement complexity (m) of each method. Here again, we first generate a random rank-5 signal $W_* \in \mathbb{R}^{50\times 50}$ and its measurements using different operators. We then measure error in recovery by each of the method and consider success if the relative error $err \leq 0.05$. We repeat the experiment 10 times to obtain the recovery rate (number of success/10) for each value of m (number of measurements). Figure 1b plots the recovery rate of different approaches for different m. Clearly, the rank-one based measurements have similar recovery rate and measurement complexity as the RIP based operators. However, our rank-one operator based methods are much faster than the corresponding methods for the RIP-based measurement scheme.

Finally, in Figure 2, we validate our theoretical analysis on the measurement complexity by showing the recovery rate for different d and m. We fix the rank $k = 5$, set $d = d_1 = d_2$ and $n_1 = d_1$, $n_2 = d_2$ for dependent operators. Figure 3 plots the recovery rate for various d and m. As shown in Figure 2, both indepen-

dent and dependent operators using alternating minimization or nuclear-norm minimization require a number of measurements proportional to the dimension of d. We also see that dependent operators require a slight larger number of measurements than that of independent ones. Another interesting observation is that although our theoretical analysis requires a higher measurement complexity of ALS than that of Nuclear methods, the empirical results show their measurement complexities are almost identical for the same measurement operator.

References

1. Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., Tandon, R.: Learning sparsely used overcomplete dictionaries via alternating minimization. COLT (2014)
2. Cai, T.T., Zhang, A., et al.: Rop: Matrix recovery via rank-one projections. The Annals of Statistics 43(1), 102–138 (2015)
3. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9(6), 717–772 (December 2009)
4. Candès, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inform. Theory 56(5), 2053–2080 (2009)
5. Chen, Y.: Incoherence-optimal matrix completion. arXiv preprint arXiv:1310.0154 (2013)
6. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. Information Theory, IEEE Transactions on 57(3), 1548–1566 (2011)
7. Hardt, M.: Understanding alternating minimization for matrix completion. In: Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on. pp. 651–660. IEEE (2014)
8. Hardt, M., Wootters, M.: Fast matrix completion without the condition number. In: Proceedings of The 27th Conference on Learning Theory. pp. 638–678 (2014)
9. Hsieh, C.J., Dhillon, I.S., Ravikumar, P.K., Becker, S., Olsen, P.A.: Quic & dirty: A quadratic approximation approach for dirty statistical models. In: Advances in Neural Information Processing Systems. pp. 2006–2014 (2014)
10. Hsieh, C.J., Olsen, P.: Nuclear norm minimization via active subspace selection. In: Proceedings of The 31st International Conference on Machine Learning. pp. 575–583 (2014)
11. Jain, P., Dhillon, I.S.: Provable inductive matrix completion. CoRR abs/1306.0626 (2013), `http://arxiv.org/abs/1306.0626`
12. Jain, P., Meka, R., Dhillon, I.S.: Guaranteed rank minimization via singular value projection. In: NIPS. pp. 937–945 (2010)
13. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: STOC (2013)
14. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Transactions on Information Theory 56(6), 2980–2998 (2010)
15. Kueng, R., Rauhut, H., Terstiege, U.: Low rank matrix recovery from rank one measurements. arXiv preprint arXiv:1410.6913 (2014)
16. Lee, K., Bresler, Y.: Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. arXiv preprint arXiv:0903.4742 (2009)
17. Li, R.C.: On perturbations of matrix pencils with real spectra. Math. Comp. 62, 231–265 (1994)

18. Liu, Y.K.: Universal low-rank matrix recovery from pauli measurements. In: Advances in Neural Information Processing Systems. pp. 1638–1646 (2011)
19. Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., Jain, P.: Non-convex robust PCA. In: Advances in Neural Information Processing Systems. pp. 1107–1115 (2014)
20. Recht, B.: A simpler approach to matrix completion. The Journal of Machine Learning Research 12, 3413–3430 (2011)
21. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review 52(3), 471–501 (2010)
22. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics 12(4), 389–434 (2012)
23. Xu, M., Jin, R., Zhou, Z.H.: Speedup matrix completion with side information: Application to multi-label learning. In: Advances in Neural Information Processing Systems. pp. 2301–2309 (2013)
24. Yu, H.F., Hsieh, C.J., Si, S., Dhillon, I.S.: Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In: ICDM. pp. 765–774 (2012)
25. Yu, H.F., Jain, P., Kar, P., Dhillon, I.S.: Large-scale multi-label learning with missing labels. In: Proceedings of The 31st International Conference on Machine Learning. pp. 593–601 (2014)
26. Zuk, O., Wagner, A.: Low-rank matrix recovery from row-and-column affine measurements. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. pp. 2012–2020 (2015)

Appendix

A Proof of Theorem 1

Proof. We explain the key ideas of the proof by first presenting the proof for the special case of rank-1 $W_* = \sigma_* \boldsymbol{u}_* \boldsymbol{v}_*^T$. We later extend the proof to general rank-k case.

Similar to [13], we first characterize the update for $h+1$-th step iterates $\widehat{\boldsymbol{v}}_{h+1}$ of Algorithm 1 and its normalized form $\boldsymbol{v}_{h+1} = \widehat{\boldsymbol{v}}_{h+1}/\|\widehat{\boldsymbol{v}}_{h+1}\|_2$.

Now, by gradient of (4) w.r.t. $\widehat{\boldsymbol{v}}$ to be zero while keeping $\boldsymbol{u}_h$ to be fixed. That is,

$$\begin{aligned}
&\sum_{i=1}^m (b_i - \boldsymbol{x}_i^T \boldsymbol{u}_h \widehat{\boldsymbol{v}}_{h+1}^T \boldsymbol{y}_i)(\boldsymbol{x}_i^T \boldsymbol{u}_h)\boldsymbol{y}_i = 0, \\
i.e.,\ &\sum_{i=1}^m (\boldsymbol{u}_h^T \boldsymbol{x}_i)\boldsymbol{y}_i(\sigma_* \boldsymbol{y}_i^T \boldsymbol{v}_* \boldsymbol{u}_*^T \boldsymbol{x}_i - \boldsymbol{y}_i^T \widehat{\boldsymbol{v}}_{h+1} \boldsymbol{u}_h^T \boldsymbol{x}_i) = 0, \\
i.e.,\ &\left(\sum_{i=1}^m (\boldsymbol{x}_i^T \boldsymbol{u}_h \boldsymbol{u}_h^T \boldsymbol{x}_i)\boldsymbol{y}_i \boldsymbol{y}_i^T\right) \widehat{\boldsymbol{v}}_{h+1} = \sigma_* \left(\sum_{i=1}^m (\boldsymbol{x}_i^T \boldsymbol{u}_h \boldsymbol{u}_*^T \boldsymbol{x}_i)\boldsymbol{y}_i \boldsymbol{y}_i^T\right) \boldsymbol{v}_*, \\
i.e.,\ &\widehat{\boldsymbol{v}}_{h+1} = \sigma_* (\boldsymbol{u}_*^T \boldsymbol{u}_h)\boldsymbol{v}_* - \sigma_* B^{-1}((\boldsymbol{u}_*^T \boldsymbol{u}_h)B - \widetilde{B})\boldsymbol{v}_*, \qquad (12)
\end{aligned}$$

where,

$$B = \frac{1}{m}\sum_{i=1}^m (\boldsymbol{x}_i^T \boldsymbol{u}_h \boldsymbol{u}_h^T \boldsymbol{x}_i)\boldsymbol{y}_i \boldsymbol{y}_i^T, \quad \widetilde{B} = \frac{1}{m}\sum_{i=1}^m (\boldsymbol{x}_i^T \boldsymbol{u}_h \boldsymbol{u}_*^T \boldsymbol{x}_i)\boldsymbol{y}_i \boldsymbol{y}_i^T.$$

Note that (12) shows that $\widehat{\boldsymbol{v}}_{h+1}$ is a perturbation of $\boldsymbol{v}_*$ and the goal now is to bound the spectral norm of the perturbation term:

$$\|G\|_2 = \|B^{-1}(\boldsymbol{u}_*^T \boldsymbol{u}_h B - \widetilde{B})\boldsymbol{v}_*\|_2 \le \|B^{-1}\|_2 \|\boldsymbol{u}_*^T \boldsymbol{u}_h B - \widetilde{B}\|_2 \|\boldsymbol{v}_*\|_2. \qquad (13)$$

Now,, using Property 2 mentioned in the theorem, we get:

$$\|B - I\|_2 \le 1/100, \quad i.e.,\ \sigma_{min}(B) \ge 1 - 1/100, \quad i.e.,\ \|B^{-1}\|_2 \le 1/(1 - 1/100). \qquad (14)$$

Now,

$$\begin{aligned}
(\boldsymbol{u}_*^T \boldsymbol{u}_h)B - \widetilde{B} &= \frac{1}{m}\sum_{i=1}^m \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{x}_i^T ((\boldsymbol{u}_*^T \boldsymbol{u}_h)\boldsymbol{u}_h \boldsymbol{u}_h^T - \boldsymbol{u}_* \boldsymbol{u}_h^T)\boldsymbol{x}_i, \\
&= \frac{1}{m}\sum_{i=1}^m \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{x}_i^T (\boldsymbol{u}_h \boldsymbol{u}_h^T - I)\boldsymbol{u}_* \boldsymbol{u}_h^T \boldsymbol{x}_i, \\
&\overset{\zeta_1}{\le} \frac{1}{100}\|(\boldsymbol{u}_h \boldsymbol{u}_h^T - I)\boldsymbol{u}_*\|_2 \|\boldsymbol{u}_h^T\|_2 = \frac{1}{100}\sqrt{1 - (\boldsymbol{u}_h^T \boldsymbol{u}_*)^2}, \qquad (15)
\end{aligned}$$

where ζ_1 follows by observing that $(\boldsymbol{u}_h\boldsymbol{u}_h^T - I)\boldsymbol{u}_*$ and $\boldsymbol{u}_h$ are orthogonal set of vectors and then using Property 3 given in the Theorem 1. Hence, using (14), (15), and $\|\boldsymbol{v}_*\|_2 = 1$ along with (13), we get:

$$\|G\|_2 \leq \frac{1}{99}\sqrt{1-(\boldsymbol{u}_h^T\boldsymbol{u}_*)^2}. \tag{16}$$

We are now ready to lower bound the component of $\widehat{\boldsymbol{v}}_h$ along the correct direction $\boldsymbol{v}_*$ and the component of $\widehat{\boldsymbol{v}}_h$ that is perpendicular to the optimal direction $\boldsymbol{v}_*$.

Now, by left-multiplying (12) by $\boldsymbol{v}_*$ and using (14) we obtain:

$$\boldsymbol{v}_*^T\widehat{\boldsymbol{v}}_{h+1} = \sigma_*(\boldsymbol{u}_h^T\boldsymbol{u}_*) - \sigma_*\boldsymbol{v}_*^T G \geq \sigma_*(\boldsymbol{u}_h^T\boldsymbol{u}_*) - \frac{\sigma_*}{99}\sqrt{1-(\boldsymbol{u}_h^T\boldsymbol{u}_*)^2}. \tag{17}$$

Similarly, by multiplying (12) by $\boldsymbol{v}_*^\perp$, where $\boldsymbol{v}_*^\perp$ is a unit norm vector that is orthogonal to $\boldsymbol{v}_*$, we get:

$$\langle \boldsymbol{v}_*^\perp, \widehat{\boldsymbol{v}}_{h+1}\rangle \leq \frac{\sigma_*}{99}\sqrt{1-(\boldsymbol{u}_h^T\boldsymbol{u}_*)^2}. \tag{18}$$

Using (17), (18), and $\|\widehat{\boldsymbol{v}}_{h+1}\|_2^2 = (\boldsymbol{v}_*^T\widehat{\boldsymbol{v}}_{h+1})^2 + ((\boldsymbol{v}_*^\perp)^T\widehat{\boldsymbol{v}}_{h+1})^2$, we get:

$$\begin{aligned} 1-(\boldsymbol{v}_{h+1}^T\boldsymbol{v}_*)^2 &= \frac{\langle \boldsymbol{v}_*^\perp, \widehat{\boldsymbol{v}}_{h+1}\rangle^2}{\langle \boldsymbol{v}_*, \widehat{\boldsymbol{v}}_{h+1}\rangle^2 + \langle \boldsymbol{v}_*^\perp, \widehat{\boldsymbol{v}}_{h+1}\rangle^2}, \\ &\leq \frac{1}{99\cdot 99\cdot(\boldsymbol{u}_h^T\boldsymbol{u}_* - \frac{1}{99}\sqrt{1-(\boldsymbol{u}_h^T\boldsymbol{u}_*)^2})^2 + 1}(1-(\boldsymbol{u}_h\boldsymbol{u}_*)^2). \end{aligned} \tag{19}$$

Also, using Property 1 of Theorem 1, for $S = \frac{1}{m}\sum_{i=1}^m b_i A_i$, we get: $\|S\|_2 \geq \frac{99\sigma_*}{100}$. Moreover, by multiplying $S - W_*$ by $\boldsymbol{u}_0$ on left and $\boldsymbol{v}_0$ on the right and using the fact that $(\boldsymbol{u}_0, \boldsymbol{v}_0)$ are the largest singular vectors of S, we get: $\|S\|_2 - \sigma_*\boldsymbol{v}_0^T\boldsymbol{v}_*\boldsymbol{u}_0^T\boldsymbol{u}_* \leq \sigma_*/100$. Hence, $\boldsymbol{u}_0^T\boldsymbol{u}_* \geq 9/10$.

Using the (19) along with the above given observation and by the "inductive" assumption $\boldsymbol{u}_h^T\boldsymbol{u}_* \geq \boldsymbol{u}_0^T\boldsymbol{u}_* \geq 9/10$ (proof of the inductive step follows directly from the below equation) , we get:

$$1-(\boldsymbol{v}_{h+1}^T\boldsymbol{v}_*)^2 \leq \frac{1}{2}(1-(\boldsymbol{u}_h^T\boldsymbol{u}_*)^2). \tag{20}$$

Similarly, we can show that $1-(\boldsymbol{u}_{h+1}^T\boldsymbol{u}_*)^2 \leq \frac{1}{2}(1-(\boldsymbol{v}_{h+1}^T\boldsymbol{v}_*)^2)$. Hence, after $H = O(\log(\sigma_*/\epsilon))$ iterations, we obtain $W_H = \boldsymbol{u}_H\widehat{\boldsymbol{v}}_H^T$, s.t., $\|W_H - W_*\|_2 \leq \epsilon$.

We now generalize our above given proof to the rank-k case. In the case of rank-1 matrix recovery, we used $1-(\boldsymbol{v}_{h+1}^T\boldsymbol{u}_*)^2$ as the error or distance function and show at each step that the error decreases by at least a constant factor. For general rank-k case, we need to generalize the distance function to be a distance over subspaces of dimension-k. To this end, we use the standard principle angle based subspace distance. That is,

Definition 2. *Let $U_1, U_2 \in \mathbb{R}^{d\times k}$ be k-dimensional subspaces. Then the principle angle based distance* $\text{dist}(U_1, U_2)$ *between U_1, U_2 is given by:*

$$\text{dist}(U_1, U_2) = \|U_\perp^T U_2\|_2,$$

where $U_\perp$ is the subspace orthogonal to U_1.

Proof (Proof of Theorem 1: General Rank-k Case). For simplicity of notation, we denote U_h by U, $\widehat{V}_{h+1}$ by $\widehat{V}$, and V_{h+1} by V.

Similar to the above given proof, we first present the update equation for $\widehat{V}_{(t+1)}$. Recall that $\widehat{V}_{(t+1)} = \text{argmin}_{V\in\mathbb{R}^{d_2\times k}} \sum_i (\boldsymbol{x}_i^T W_* \boldsymbol{y}_i - \boldsymbol{x}_i^T U_t \widehat{V}^T \boldsymbol{y}_i)^2$. Hence, by setting gradient of this objective function to 0, using the above given notation and by simplifications, we get:

$$\widehat{V} = {W_*}^T U - F, \tag{21}$$

where $F = [F_1 F_2 \dots F_k]$ is the "error" matrix.

Before specifying F, we first introduce *block matrices* $B, C, D, S \in \mathbb{R}^{kd_2\times kd_2}$ with (p, q)-th block $B_{pq}, C_{pq}, S_{pq}, D_{pq}$ given by:

$$B_{pq} = \sum_i \boldsymbol{y}_i \boldsymbol{y}_i^T (\boldsymbol{x}_i^T \boldsymbol{u}_p)(\boldsymbol{x}_i^T \boldsymbol{u}_q), \tag{22}$$

$$C_{pq} = \sum_i \boldsymbol{y}_i \boldsymbol{y}_i^T (\boldsymbol{x}_i^T \boldsymbol{u}_p)(\boldsymbol{x}_i^T \boldsymbol{u}_{*q}), \tag{23}$$

$$D_{pq} = \boldsymbol{u}_p^T \boldsymbol{u}_{*q} I, \tag{24}$$

$$S_{pq} = \sigma_*^p I \quad \text{if } p = q, \quad \text{and} \quad 0 \quad \text{if } p \neq q. \tag{25}$$

where $\sigma_*^p = \Sigma_*(p, p)$, i.e., the p-th singular value of W_* and $\boldsymbol{u}_{*q}$ is the q-th column of U_*.

Then, using the definitions given above, we get:

$$\begin{bmatrix} F_1 \\ \vdots \\ F_k \end{bmatrix} = B^{-1}(BD - C)S \cdot \text{vec}(V_*). \tag{26}$$

Now, recall that in the $t+1$-th iteration of Algorithm 1, V_{t+1} is obtained by QR decomposition of $\widehat{V}_{t+1}$. Using notation mentioned above, $\widehat{V} = VR$ where R denotes the lower triangular matrix R_{t+1} obtained by the QR decomposition of V_{t+1}.

Now, using (21), $V = \widehat{V}R^{-1} = (W_*^T U - F)R^{-1}$. Multiplying both the sides by $V_*^\perp$, where $V_*^\perp$ is a fixed orthonormal basis of the subspace orthogonal to $span(V_*)$, we get:

$$(V_*^\perp)^T V = -(V_*^\perp)^T F R^{-1} \Rightarrow dist(V_*, V_{t+1}) = \|(V_*^\perp)^T V\|_2 \leq \|F\|_2 \|R^{-1}\|_2. \tag{27}$$

Also, note that using the initialization property (1) mentioned in Theorem 1, we get $\|S - W_*\|_2 \leq \frac{\sigma_*^k}{100}$. Now, using the standard sin theta theorem for singular vector perturbation[17], we get: $\text{dist}(U_0, U_*) \leq \frac{1}{100}$.

Theorem 1 now follows by using Lemma 2, Lemma 3 along with the above mentioned bound on $\text{dist}(U_0, U_*)$.

Lemma 2. *Let $\mathcal{A}$ be a rank-one measurement operator where $A_i = \boldsymbol{x}_i \boldsymbol{y}_i^T$. Also, let $\mathcal{A}$ satisfy Property 1, 2, 3 mentioned in Theorem 1 and let $\sigma_*^1 \geq \sigma_*^2 \geq \cdots \geq \sigma_*^k$ be the singular values of W_*. Then,*

$$\|F\|_2 \leq \frac{\sigma_*^k}{100} \text{dist}(U_t, U_*).$$

Lemma 3. *Let $\mathcal{A}$ be a rank-one measurement operator where $A_i = \boldsymbol{x}_i \boldsymbol{y}_i^T$. Also, let $\mathcal{A}$ satisfy Property 1, 2, 3 mentioned in Theorem 1. Then,*

$$\|R^{-1}\|_2 \leq \frac{1}{\sigma_*^k \cdot \sqrt{1 - dist^2(U_t, U_*)} - \|F\|_2}.$$

Proof (Proof of Lemma 2). Recall that $\text{vec}(F) = B^{-1}(BD - C)S \cdot \text{vec}(V_*)$. Hence,

$$\|F\|_2 \leq \|F\|_F \leq \|B^{-1}\|_2 \|BD - C\|_2 \|S\|_2 \|\text{vec}(V_*)\|_2 = \sigma_*^1 \sqrt{k} \|B^{-1}\|_2 \|BD - C\|_2. \tag{28}$$

Now, we first bound $\|B^{-1}\|_2 = 1/(\sigma_{\min}(B))$. Also, let $Z = [\boldsymbol{z}_1 \boldsymbol{z}_2 \ldots \boldsymbol{z}_k]$ and let $\boldsymbol{z} = \text{vec}(Z)$. Then,

$$\begin{aligned} \sigma_{\min}(B) &= \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2 = 1} \boldsymbol{z}^T B \boldsymbol{z} = \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2 = 1} \sum_{1 \leq p \leq k, 1 \leq q \leq k} \boldsymbol{z}_p^T B_{pq} \boldsymbol{z}_q \\ &= \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2 = 1} \sum_p \boldsymbol{z}_p^T B_{pp} \boldsymbol{z}_p + \sum_{pq, p \neq q} \boldsymbol{z}_p^T B_{pq} \boldsymbol{z}_q. \end{aligned} \tag{29}$$

Recall that, $B_{pp} = \frac{1}{m} \sum_{i=1}^m \boldsymbol{y}_i \boldsymbol{y}_i^T (\boldsymbol{x}_i^T \boldsymbol{u}_p)^2$ and $\boldsymbol{u}_p$ is independent of $\xi, \boldsymbol{y}_i, \forall i$. Hence, using Property 2 given in Theorem 1, we get:

$$\sigma_{\min}(B_{pp}) \geq 1 - \delta, \tag{30}$$

where,

$$\delta = \frac{1}{k^{3/2} \cdot \beta \cdot 100},$$

and $\beta = \sigma_*^1 / \sigma_*^k$ is the condition number of W_*.

Similarly, using Property (3), we get:

$$\|B_{pq}\|_2 \leq \delta. \tag{31}$$

Hence, using (29), (30), (31), we get:

$$\sigma_{\min}(B) \geq \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2 = 1} (1 - \delta) \sum_p \|\boldsymbol{z}_p\|_2^2 - \delta \sum_{pq, p \neq q} \|\boldsymbol{z}_p\|_2 \|\boldsymbol{z}_q\|_2 = \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2 = 1} 1 - \delta \sum_{pq} \|\boldsymbol{z}_p\|_2 \|\boldsymbol{z}_q\|_2 \geq 1 - k\delta. \tag{32}$$

Now, consider $BD - C$:

$$\begin{aligned}
\|BD - C\|_2 &= \max_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} |\boldsymbol{z}^T (BD - C)\boldsymbol{z}|, \\
&= \max_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \left| \sum_{1 \le p \le k, 1 \le q \le k} \boldsymbol{z}_p^T \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{z}_q \boldsymbol{x}_i^T \left(\sum_{1 \le \ell \le k} \langle \boldsymbol{u}_\ell, \boldsymbol{u}_{*q} \rangle \boldsymbol{u}_p \boldsymbol{u}_\ell^T - \boldsymbol{u}_p \boldsymbol{u}_{*q}^T \right) \boldsymbol{x}_i \right|, \\
&= \max_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \left| \sum_{1 \le p \le k, 1 \le q \le k} \boldsymbol{z}_p^T \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{z}_q \boldsymbol{x}_i^T \boldsymbol{u}_p \boldsymbol{u}_{*q}^T (UU^T - I) \boldsymbol{x}_i \right|, \\
&\overset{\zeta_1}{\le} \delta \max_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \sum_{1 \le p \le k, 1 \le q \le k} \|(UU^T - I)\boldsymbol{u}_{*q}\|_2 \|\boldsymbol{z}_p\|_2 \|\boldsymbol{z}_q\|_2 \le k \cdot \delta \cdot \mathrm{dist}(U, U_*),
\end{aligned} \tag{33}$$

where ζ_1 follows by observing that $\boldsymbol{u}_{*q}^T (UU^T - I)\boldsymbol{u}_p = 0$ and then by applying Property (3) mentioned in Theorem 1.

Lemma now follows by using (33) along with (28) and (32).

Proof (Proof of Lemma 3). The lemma is exactly the same as Lemma 4.7 of [13]. We reproduce their proof here for completeness.

Let $\sigma_{\min}(R)$ be the smallest singular value of R, then:

$$\begin{aligned}
\sigma_{\min}(R) &= \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \|R\boldsymbol{z}\|_2 = \min_{z, \|z\|_2=1} \|VR\boldsymbol{z}\|_2 = \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \|V_* \Sigma_* U_*^T U \boldsymbol{z} - F\boldsymbol{z}\|_2, \\
&\ge \min_{\boldsymbol{z}, \|\boldsymbol{z}\|_2=1} \|V_* \Sigma_* U_*^T U \boldsymbol{z}\|_2 - \|F\boldsymbol{z}\|_2 \ge \sigma_*^k \sigma_{\min}(U^T U_*) - \|F\|_2, \\
&\ge \sigma_*^k \sqrt{1 - \|U^T U_*^\perp\|_2^2} - \|F\|_2 = \sigma_*^k \sqrt{1 - \mathrm{dist}(U_*, U)^2} - \|F\|_2.
\end{aligned} \tag{34}$$

Lemma now follows by using the above inequality along with the fact that $\|R^{-1}\|_2 \le 1/\sigma_{\min}(R)$.

B Proofs for Matrix Sensing using Rank-one Independent Gaussian Measurements

B.1 Proof of Claim 4.2

Proof. The main idea behind our proof is to show that there exists two rank-1 matrices Z_U, Z_L s.t. $\|\mathcal{A}_{GI}(Z_U)\|_2^2$ is large while $\|\mathcal{A}_{GI}(Z_L)\|_2^2$ is much smaller than $\|\mathcal{A}_{GI}(Z_U)\|_2^2$.

In particular, let $Z_U = \boldsymbol{x}_1 \boldsymbol{y}_1^T$ and let $Z_L = \boldsymbol{u}\boldsymbol{v}^T$ where $\boldsymbol{u}, \boldsymbol{v}$ are sampled from normal distribution independent of X, Y. Now,

$$\|\mathcal{A}_{GI}(Z_U)\|_2^2 = \sum_{i=1}^m \|\boldsymbol{x}_1\|_2^4 \|\boldsymbol{y}_1\|_2^4 + \sum_{i=2}^m (\boldsymbol{x}_1^T \boldsymbol{x}_i)^2 (\boldsymbol{y}_1^T \boldsymbol{y}_i)^2.$$

Now, as $\boldsymbol{x}_i, \boldsymbol{y}_i, \forall i$ are multi-variate normal random variables, $\|\boldsymbol{x}_1\|_2^4 \|\boldsymbol{y}_1\|_2^4 \geq 0.5 d_1^2 d_2^2$ w.p. $\geq 1 - 2\exp(-d_1 - d_2)$.

$$\|\mathcal{A}_{GI}(Z_U)\|_2^2 \geq .5 d_1^2 d_2^2. \tag{35}$$

Moreover, $\|Z_U\|_F^2 \leq 2 d_1 d_2$ w.p. $\geq 1 - 2\exp(-d_1 - d_2)$.

Now, consider

$$\|\mathcal{A}_{GI}(Z_L)\|_2^2 = \sum_{i=2}^{m} (\boldsymbol{u}^T \boldsymbol{x}_i)^2 (\boldsymbol{v}^T \boldsymbol{y}_i)^2,$$

where $Z_L = \boldsymbol{u}\boldsymbol{v}^T$ and $\boldsymbol{u}, \boldsymbol{v}$ are sampled from standard normal distribution, independent of $\boldsymbol{x}_i, \boldsymbol{y}_i, \forall i$. Since, $\boldsymbol{u}, \boldsymbol{v}$ are independent of $\boldsymbol{x}_1^T \boldsymbol{x}_i \sim N(0, \|\boldsymbol{x}_1\|_2)$ and $\boldsymbol{y}_1^T \boldsymbol{y}_i \sim N(0, \|\boldsymbol{y}_1\|_2)$. Hence, w.p. $\geq 1 - 1/m^3$, $|\boldsymbol{u}^T \boldsymbol{x}_i| \leq \log(m)\|u\|_2, |\boldsymbol{v}^T \boldsymbol{y}_i| \leq \log(m)\|v\|_2, \forall i \geq 2$. That is, w.p. $1 - 1/m^3$:

$$\|\mathcal{A}_{GI}(Z_L)\|_2^2 \leq 4m \log^4 m d_1 d_2. \tag{36}$$

Furthermore, $\|Z_L\|_F^2 \leq 2 d_1 d_2$ w.p. $\geq 1 - 2\exp(-d_1 - d_2)$.

Using (35), (36), we get that w.p. $\geq 1 - 2/m^3 - 10\exp(-d_1 - d_2)$:

$$40m \log^4 m \leq \|\mathcal{A}_{GI}(Z/\|Z\|_F)\|^2 \leq .05 d_1 d_2.$$

Now, for RIP to be satisfied with a constant δ, the lower and upper bound should be at most a constant factor apart. However, the above equation clearly shows that the upper and lower bound can match only when $m = \Omega(d_1 d_2 / \log(5 d_1 d_2))$. Hence, for m that at most linear in both d_1, d_2 cannot be satisfied with probability $\geq 1 - 1/(d_1 + d_2)^3$.

B.2 Proof of Thoerem 2

Proof. We divide the proof into three parts where each part proves a property mentioned in Theorem 1.

Proof (Proof of Property 1). Now,

$$S = \frac{1}{m} \sum_{i=1}^{m} b_i \boldsymbol{x}_i \boldsymbol{y}_i^T = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T U_* \Sigma_* V_*^T \boldsymbol{y}_i \boldsymbol{y}_i^T = \frac{1}{m} \sum_{i=1}^{m} Z_i,$$

where $Z_i = \boldsymbol{x}_i \boldsymbol{x}_i^T U_* \Sigma_* V_*^T \boldsymbol{y}_i \boldsymbol{y}_i^T$. Note that $\mathbb{E}[Z_i] = U_* \Sigma_* V_*^T$. Also, both $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are spherical Gaussian variables and hence are rotationally invariant. Therefore, wlog, we can assume that $U_* = [\boldsymbol{e}_1 \boldsymbol{e}_2 \dots \boldsymbol{e}_k]$ and $V_* = [\boldsymbol{e}_1 \boldsymbol{e}_2 \dots \boldsymbol{e}_k]$ where e_i is the i-th canonical basis vector.

As S is a sum of m random matrices, the goal is to apply Theorem 6 by [22] to show that S is close to $\mathbb{E}[S] = W = U_* \Sigma_* V_*^T$ for large enough m. However, Theorem 6 requires bounded random variable while Z_i is an unbounded variable.

We handle this issue by clipping Z_i to ensure that its spectral norm is always bounded. In particular, consider the following random variable:

$$\widetilde{x}_{ij} = \begin{cases} x_{ij}, & |x_{ij}| \leq C\sqrt{\log(m(d_1+d_2))}, \\ 0, & \text{otherwise}, \end{cases} \tag{37}$$

where x_{ij} is the j-th co-ordinate of $\boldsymbol{x}_i$. Similarly, define:

$$\widetilde{y}_{ij} = \begin{cases} y_{ij}, & |y_{ij}| \leq C\sqrt{\log(m(d_1+d_2))}, \\ 0, & \text{otherwise}. \end{cases} \tag{38}$$

Note that, $\mathbb{P}(x_{ij} = \widetilde{x}_{ij}) \geq 1 - \frac{1}{(m(d_1+d_2))^C}$ and $\mathbb{P}(y_{ij} = \widetilde{y}_{ij}) \geq 1 - \frac{1}{(m(d_1+d_2))^C}$. Also, $\widetilde{x}_{ij}, \widetilde{y}_{ij}$ are still symmetric and independent random variables, i.e., $\mathbb{E}[\widetilde{x}_{ij}] = \mathbb{E}[\widetilde{y}_{ij}] = 0,\ \forall i,j$. Hence, $\mathbb{E}[\widetilde{x}_{ij}\widetilde{x}_{i\ell}] = 0, \forall j \neq \ell$. Furthermore, $\forall j$,

$$\begin{aligned} \mathbb{E}[\widetilde{x}_{ij}^2] &= \mathbb{E}[x_{ij}^2] - \frac{2}{\sqrt{2\pi}} \int_{C\sqrt{\log(m(d_1+d_2))}}^{\infty} x^2 \exp(-x^2/2) dx, \\ &= 1 - \frac{2}{\sqrt{2\pi}} \frac{C\sqrt{\log(m(d_1+d_2))}}{(m(d_1+d_2))^{C^2/2}} - \frac{2}{\sqrt{2\pi}} \int_{C\sqrt{\log(m(d_1+d_2))}}^{\infty} \exp(-x^2/2) dx, \\ &\geq 1 - \frac{2C\sqrt{\log(m(d_1+d_2))}}{(m(d_1+d_2))^{C^2/2}}. \end{aligned} \tag{39}$$

Similarly,

$$\mathbb{E}[\widetilde{y}_{ij}^2] \geq 1 - \frac{2C\sqrt{\log(m(d_1+d_2))}}{(m(d_1+d_2))^{C^2/2}}. \tag{40}$$

Now, consider RV, $\widetilde{Z}_i = \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T U_* \Sigma_* V_*^T \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i^T$. Note that, $\|\widetilde{Z}_i\|_2 \leq C^4\sqrt{d_1 d_2} k \log^2(m(d_1+d_2))\sigma_*^1$ and $\|\mathbb{E}[\widetilde{Z}_i]\|_2 \leq \sigma_*^1$. Also,

$$\begin{aligned} \|\mathbb{E}[\widetilde{Z}_i\widetilde{Z}_i^T]\|_2 &= \|\mathbb{E}[\|\widetilde{\boldsymbol{y}}_i\|_2^2 \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T U_*\Sigma_* V_*^T \widetilde{\boldsymbol{y}}_i\widetilde{\boldsymbol{y}}_i^T V_*\Sigma_* U_*^T \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T]\|_2, \\ &\leq C^2 d_2 \log(m(d_1+d_2))\mathbb{E}[\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T U_*\Sigma_*^2 U_*^T \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T]\|_2, \\ &\leq C^2 d_2 \log(m(d_1+d_2))(\sigma_*^1)^2 \|\mathbb{E}[\|U_*^T\widetilde{\boldsymbol{x}}_i\|_2^2 \widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T]\|_2, \\ &\leq C^4 k d_2 \log^2(m(d_1+d_2))(\sigma_*^1)^2. \end{aligned} \tag{41}$$

Similarly,

$$\|\mathbb{E}[\widetilde{Z}_i]\mathbb{E}[\widetilde{Z}_i^T]\|_2 \leq (\sigma_*^{max})^2. \tag{42}$$

Similarly, we can obtain bounds for $\|\mathbb{E}[\widetilde{Z}_i^T\widetilde{Z}_i]\|_2$, $\|\mathbb{E}[\widetilde{Z}_i]^T\mathbb{E}[\widetilde{Z}_i]\|_2$.

Finally, by selecting $m = \frac{C_1 k(d_1+d_2)\log^2(d_1+d_2)}{\delta^2}$ and applying Theorem 6 we get (w.p. $1 - \frac{1}{(d_1+d_2)^{10}}$),

$$\|\frac{1}{m}\sum_{i=1}^{m} \widetilde{Z}_i - \mathbb{E}[\widetilde{Z}_i]\|_2 \leq \delta. \tag{43}$$

Note that $\mathbb{E}[\widetilde{Z}_i] = \mathbb{E}[\widetilde{x}_{i1}^2]\mathbb{E}[\widetilde{y}_{i1}^2]U_*\Sigma_*V_*^T$. Hence, by using (43), (39), (40),

$$\|\frac{1}{m}\sum_{i=1}^{m}\widetilde{Z}_i - U_*\Sigma_*V_*^T\|_2 \leq \delta + \frac{\sigma_*^1}{(d_1+d_2)^{100}}.$$

Finally, by observing that by selecting C to be large enough in the definition of $\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{y}}_i$ (see (37), (38)), we get $P(\|Z_i - \widetilde{Z}_i\|_2 = 0) \geq 1 - \frac{1}{(d_1+d_2)^5}$. Hence, by assuming δ to be a constant wrt d_1, d_2 and by union bound, w.p. $1 - \frac{2\delta^{10}}{(d_1+d_2)^5}$,

$$\|\frac{1}{m}\sum_{i=1}^{m}Z_i - W_*\|_2 \leq 5\delta\|W_*\|_2.$$

Now, the theorem follows directly by setting $\delta = \frac{1}{100k^{3/2}\beta}$.

Proof (Proof of Property 2). Here again, the goal is to show that the random matrix B_x concentrates around its mean which is given by I. Now, as $\boldsymbol{x}_i, \boldsymbol{y}_i$ are rotationally invariant random variables, wlog, we can assume $\boldsymbol{u}_h = \boldsymbol{e}_1$. That is, $(\boldsymbol{x}_i^T\boldsymbol{u}_h\boldsymbol{u}_h^T\boldsymbol{x}_i) = x_{i1}^2$ where x_{i1} is the first coordinate of $\boldsymbol{x}_i$. Furthermore, similar to (37), (38), we define clipped random variables $\widetilde{\boldsymbol{x}}_{i1}$ and $\widetilde{\boldsymbol{y}}_i$ below:

$$\widetilde{x}_{i1} = \begin{cases} x_{i1}, & |x_{i1}| \leq C\sqrt{\log(m)}, \\ 0, & \text{otherwise.} \end{cases} \tag{44}$$

$$\widetilde{\boldsymbol{y}}_i = \begin{cases} \boldsymbol{y}_i, & \|\boldsymbol{y}_i\|_2^2 \leq 2(d_1+d_2), \\ 0, & \text{otherwise.} \end{cases} \tag{45}$$

Now, consider $\widetilde{B} = \frac{1}{m}\sum_{i=1}^{m}\widetilde{Z}_i$, where $\widetilde{Z}_i = \widetilde{x}_{i1}^2\widetilde{\boldsymbol{y}}_i\widetilde{\boldsymbol{y}}_i^T$. Note that, $\|\widetilde{Z}_i\|_2 \leq 2C^2(d_1+d_2)\log(m)$. Similarly, $\|\mathbb{E}[\sum_i \widetilde{Z}_i\widetilde{Z}_i^T]\|_2 \leq 2mC^4(d_1+d_2)\log^2(m)$. Hence, using Theorem 6, we get:

$$\mathbb{P}\left(\|\frac{1}{m}\sum_i \widetilde{Z}_i - \mathbb{E}[\widetilde{Z}_i]\|_2 \geq \gamma\right) \leq \exp\left(-\frac{m\gamma^2}{2C^4(d_1+d_2)\log^2(m)(1+\gamma/3)}\right). \tag{46}$$

Now, using argument similar to (39), we get $\|\mathbb{E}[\widetilde{Z}_i] - I\|_2 \leq \frac{2C\log(m)}{m^{C^2/2}}$. Furthermore, $\mathbb{P}(\widetilde{Z}_i = Z_i) \geq 1 - \frac{1}{m^3}$. Hence for $m = \Omega(k(d_1+d_2)\log(d_1+d_2)/\delta^2)$, w.p. $\geq 1 - \frac{2}{m^3}$,

$$\|B_x - I\|_2 \leq \delta. \tag{47}$$

Similarly, we can prove the bound for B_y using exactly same set of arguments.

Proof (Proof of Property 3). Let, $C = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{y}_i\boldsymbol{y}_i^T\boldsymbol{x}_i^T\boldsymbol{u}\boldsymbol{u}_\perp^T\boldsymbol{x}_i$ where $\boldsymbol{u}$, $\boldsymbol{u}_\perp$ are *fixed* orthogonal unit vectors. Now $\boldsymbol{x}_i^T\boldsymbol{u}_h \sim \mathcal{N}(0,1)$ and $\boldsymbol{u}_\perp^T\boldsymbol{x}_i \sim \mathcal{N}(0,1)$ are both normal variables. Also, note that $\boldsymbol{u}$ and $\boldsymbol{u}_\perp$ are orthogonal, hence $\boldsymbol{x}_i^T\boldsymbol{u}_h$ and $\boldsymbol{u}_\perp^T\boldsymbol{x}_i$ are independent variables.

Hence, $\mathbb{E}[\boldsymbol{x}_i^T\boldsymbol{u}\boldsymbol{u}_\perp^T\boldsymbol{x}_i] = 0$, i.e., $\mathbb{E}[C] = 0$. Now, let $m = \Omega(k(d_1+d_2)\log(d_1+d_2)/\delta^2)$. Then, using the clipping argument (used in the previous proof) with Theorem 6, Property 3 is satisfied w.p. $\geq 1 - \frac{2}{m^3}$. That is, $\|C_y\|_2 \leq \delta$. Moreover, $\|C_x\|_2 \leq \delta$ also can be proved using similar proof to the one given above.

C Proof of Matrix Sensing using Rank-one Dependent Gaussian Measurements

C.1 Proof of Lemma 1

Proof. **Incoherence:** The Incoherence property directly follows Lemma 2.2 in [3].

Averaging Property: Given any orthogonal matrix $U \in \mathbb{R}^{d\times k}(d \geq k)$, let $Q = [U, U_\perp]$, where $U_\perp$ is a complementary matrix of U. Define $S = XQ = (XQX^T)X$. The matrix XQX^T can be viewed as a rotation matrix constrained in the column space of X. Thus, S is a constrained rotation of X, which implies S is also a random orthogonal matrix and so is the first k columns of S. We use $T \in \mathbb{R}^{n\times k}$ to denote the first k columns of S.

$$\max_i \|U^T \boldsymbol{z}_i\| = \max_i \|U^T Q \boldsymbol{s}_i\| = \max_i \|\boldsymbol{t}_i\|$$

where $\boldsymbol{t}_i$ is the transpose of the i-th row of T. Now this property follows from Lemma 2.2 in [3].

C.2 Proof of Theorem 3

Proof. Similar to the proof of Theorem 2, we divide the proof into three parts where each part proves a property mentioned in Theorem 1. And in this proof, we set $d = d_1 + d_2$ and $n = n_1 + n_2$.

Proof (Proof of Property 1). As mentioned in the proof sketch, wlog, we can assume that both X, Y are orthonormal matrices and that the condition number of R is same as condition number of W_*.

We first recall the definition of S:

$$S = \frac{n_1 n_2}{m} \sum_{(i,j)\in\Omega}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T U_* \Sigma_* V_*^T \boldsymbol{y}_j \boldsymbol{y}_j^T = \frac{n_1 n_2}{m} \sum_{(i,j)\in\Omega}^{m} Z_{ij},$$

where $Z_{ij} = \boldsymbol{x}_i \boldsymbol{x}_i^T U_* \Sigma_* V_*^T \boldsymbol{y}_j \boldsymbol{y}_j^T = X \boldsymbol{e}_i \boldsymbol{e}_i^T X^T U_* \Sigma_* V_*^T Y \boldsymbol{e}_j \boldsymbol{e}_j^T Y^T$, where $\boldsymbol{e}_i, \boldsymbol{e}_j$ denotes the i-th, j-th canonical basis vectors, respectively.

Also, since (i, j) is sampled uniformly at random from $[n_1] \times [n_2]$. Hence, $\mathbb{E}_i[\boldsymbol{e}_i \boldsymbol{e}_i^T] = \frac{1}{n_1} I$ and $\mathbb{E}_j[\boldsymbol{e}_j \boldsymbol{e}_j^T] = \frac{1}{n_2} I$. That is,

$$\mathbb{E}_{ij}[Z_{ij}] = \frac{1}{n_1 n_2} X X^T U_* \Sigma_* V_*^T Y Y^T = U_* \Sigma_* V_*^T = W_*/(n_1 \cdot n_2),$$

where $XX^T = I$, $YY^T = I$ follows by orthonormality of both X and Y.

We now use the matrix concentration bound of Theorem 6 to bound $\|S - W_*\|_2$. To apply the bound of Theorem 6, we first need to bound the following two quantities:

– **Bound** $\max_{ij} \|Z_{ij}\|_2$: Now,

$$\|Z_{ij}\|_2 = \|\boldsymbol{x}_i\boldsymbol{x}_i^T U_* \Sigma_* V_*^T \boldsymbol{y}_j \boldsymbol{y}_j^T\|_2 \leq \sigma_*^1 \|U_*^T \boldsymbol{x}_i\|_2 \|V_*^T \boldsymbol{y}_j\|_2 \|\boldsymbol{x}_i\|_2 \|\boldsymbol{y}_j\|_2 \leq \frac{\sigma_*^1 \mu\mu_0 \sqrt{d_1 d_2} k}{n_1 n_2},$$

where the last inequality follows using the two properties of random orthogonal matrices of X, Y.

– **Bound** $\|\sum_{(i,j)\in\Omega} E[Z_{ij}Z_{ij}^T]\|_2$ **and** $\|\sum_{(i,j)\in\Omega} E[Z_{ij}^T Z_{ij}]\|_2$: We first consider $\|\sum_{(i,j)\in\Omega} E[Z_{ij}Z_{ij}^T]\|_2$:

$$\begin{aligned}
\left\| \sum_{(i,j)\in\Omega} E[Z_{ij}Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^T W_* \boldsymbol{y}_j\boldsymbol{y}_j^T \boldsymbol{y}_j\boldsymbol{y}_j^T W_*^T \boldsymbol{x}_i\boldsymbol{x}_i^T] \right\|_2, \\
&\overset{\zeta_1}{\leq} \frac{\mu d_2}{n_2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^T W_* \boldsymbol{y}_j\boldsymbol{y}_j^T W_*^T \boldsymbol{x}_i\boldsymbol{x}_i^T] \right\|_2, \\
&\overset{\zeta_2}{=} \frac{\mu^2 d_2}{n_2^2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^T W_* W_*^T \boldsymbol{x}_i\boldsymbol{x}_i^T] \right\|_2, \\
&\overset{\zeta_3}{\leq} \frac{(\sigma_*^1)^2 \mu\mu_0 k d_2}{n_1 n_2^2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^T] \right\|_2, \\
&\overset{\zeta_4}{=} \frac{(\sigma_*^1)^2 \mu\mu_0 k d_2}{n_1^2 n_2^2} \cdot m, \qquad (48)
\end{aligned}$$

where ζ_1, ζ_3 follows by using the two properties of X, Y and $\|W_*\|_2 \leq \sigma_*^1$. ζ_2, ζ_4 follows by using $\mathbb{E}_i[\boldsymbol{e}_i\boldsymbol{e}_i^T] = \frac{1}{n_1} I$ and $\mathbb{E}_j[\boldsymbol{e}_j\boldsymbol{e}_j^T] = \frac{1}{n_2} I$.
Now, bound for $\|\sum_{(i,j)\in\Omega} E[Z_{ij}^T Z_{ij}]\|_2$ also turns out to be exactly the same and can be easily computed using exactly same arguments as above.

Now, by applying Theorem 6 and using the above computed bounds we get:

$$Pr(\|S - W_*\|_2 \geq \sigma_*^1 \gamma) \leq d \exp\left(-\frac{m\gamma^2}{\mu\mu_0 k d(1+\gamma/3)} \right). \qquad (49)$$

That is, w.p. $\geq 1 - \gamma$:

$$\|S - W_*\|_2 \leq \frac{\sigma_*^1 \sqrt{2\mu\mu_0 k d \log(d/\gamma)}}{\sqrt{m}}. \qquad (50)$$

Because the properties for random orthogonal matrices fail with probability $cn^{-3}\log n$, we assume *gamma* at least have the same magnitude of such a failure probability to simplify the result, i.e., $\gamma \geq cn^{-3}\log n$. Hence, by selecting $m = O(\mu\mu_0 k^4 \cdot \beta^2 \cdot d\log(d/\gamma))$ where $\beta = \sigma_*^1/\sigma_*^k$, the following holds w.p. $\geq 1 - \gamma$:

$$\|S - W_*\|_2 \leq \|W_*\|_2 \cdot \delta,$$

where $\delta = 1/(k^{3/2} \cdot \beta \cdot 100)$.

Proof (Proof of Property 2). We prove the property for B_y; proof for B_y follows analogously. Now, let $B_y = \frac{n_1 n_2}{m} \sum_{(i,j)\in\Omega} Z_{ij}$ where $Z_i = \boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i \boldsymbol{y}_i \boldsymbol{y}_i^T$. Then,

$$\mathbb{E}[B_y] = \frac{n_1 n_2}{m} \sum_{(i,j)\in\Omega} Z_{ij} = \frac{n_1 n_2}{m} \sum_{i=1}^{m} \mathbb{E}_{(i,j)\in\Omega}[\boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i \boldsymbol{y}_i \boldsymbol{y}_i^T] = I. \tag{51}$$

Here again, we apply Theorem 6 to bound $\|B_y - I\|_2$. To this end, we need to bound the following quantities:

- **Bound** $\max_{ij} \|Z_{ij}\|_2$: Now,

$$\|Z_{ij}\|_2 = \|\boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i \boldsymbol{y}_i \boldsymbol{y}_i^T\|_2 \leq \|\boldsymbol{y}_i\|_2^2 \|\boldsymbol{u}^T \boldsymbol{x}_i\|_2^2 \leq \frac{\mu\mu_0 d_2 \log n_2}{n_1 n_2}.$$

  The log factor comes from the second property of random orthogonal matrices.
- **Bound** $\|\sum_{(i,j)\in\Omega} E[Z_{ij} Z_{ij}^T]\|_2$ **and** $\|\sum_{(i,j)\in\Omega} E[Z_{ij}^T Z_{ij}]\|_2$: We first consider $\|\sum_{(i,j)\in\Omega} E[Z_{ij} Z_{ij}^T]\|_2$:

$$\begin{aligned}
\left\| \sum_{(i,j)\in\Omega} \mathbb{E}[Z_{ij} Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[(\boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i)^2 \|\boldsymbol{y}_i\|_2^2 \boldsymbol{y}_i \boldsymbol{y}_i^T] \right\|_2, \\
&\overset{\zeta_1}{\leq} \frac{\mu d_2}{n_2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[(\boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i)^2 \boldsymbol{y}_i \boldsymbol{y}_i^T] \right\|_2, \\
&\overset{\zeta_2}{=} \frac{\mu d_2}{n_2^2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[(\boldsymbol{x}_i^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{x}_i)^2 \right\|_2, \\
&\overset{\zeta_3}{\leq} \frac{\mu\mu_0 d_2 \log n_2}{n_1 n_2^2} \left\| \sum_{(i,j)\in\Omega} \mathbb{E}[(\boldsymbol{x}_i^T \boldsymbol{u})^2] \right\|_2, \\
&\overset{\zeta_4}{=} \frac{\mu\mu_0 d_2 \log n_2}{n_1^2 n_2^2} \cdot m
\end{aligned} \tag{52}$$

Note that if we assume $k \geq \log n$, the above given bounds are less than the ones obtained in the Initialization Property's proof respectively. Hence, by applying Theorem 6 in a similar manner, and selecting $m = O(\mu\mu_0 k^4 \cdot \beta^2 \cdot d \log(d/\gamma))$ and $\delta = 1/(k^{3/2} \cdot \beta \cdot 100)$, we get w.p. $\geq 1 - \gamma$:

$$\|B_y - I\|_2 \leq \delta.$$

Hence Proved. $\|B_x - I\|_2 \leq \delta$ can be proved similarly.

Proof (Proof of Property 3). Note that $\mathbb{E}[C_y] = \mathbb{E}[\sum_{(i,j)\in\Omega} Z_{ij}] = 0$. Furthermore, both $\|Z_{ij}\|_2$ and $\|\mathbb{E}[\sum_{(i,j)\in\Omega} Z_{ij} Z_{ij}^T]\|_2$ have exactly the same bounds as those given in the Property 2's proof above. Hence, we obtain similar bounds.

That is, if $m = O(\mu\mu_0 k^4 \cdot \beta^2 \cdot d\log(d/\gamma))$ and $\delta = 1/(k^{3/2} \cdot \beta \cdot 100)$, we get w.p. $\geq 1-\gamma$:

$$\|C_y\|_2 \leq \delta.$$

Hence Proved. $\|C_x\|_2$ can also be bounded analogously.

D Proofs for Rank-one Matrix Sensing using Nuclear-norm Minimization

D.1 Preliminaries

We first define some notations. The orthogonal projection $\mathcal{P}_T : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ is defined as

$$\mathcal{P}_T(Z) = U_*U_*^T Z + ZV_*V_*^T - U_*U_*^T ZV_*V_*^T.$$

The corresponding orthogonal projection onto $T^\perp$ is given by

$$\mathcal{P}_{T^\perp}(Z) = (I_{d_1} - U_*U_*^T)Z(I_{d_2} - V_*V_*^T)$$

where I_d denotes the $d \times d$ identity matrix. Define the sampling operator $\mathcal{R}_\Omega : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ as

$$\mathcal{R}_\Omega(\hat{Z}) := \sum_{i,j} \frac{\delta_{ij}}{p} \hat{Z}_{ij}\boldsymbol{e}_i\boldsymbol{e}_j^T$$

where $\delta_{ij} = \mathbb{I}[(i,j) \in \Omega]$, and $\mathbb{P}[\delta_{ij} = 1] = p$. When $\hat{Z} = XZY^T$ for any $Z \in \mathbb{R}^{d_1\times d_2}$,

$$X^T\mathcal{R}_\Omega(XZY^T)Y = \sum_{i,j} \frac{\delta_{ij}}{p} \boldsymbol{x}_i^T Z\boldsymbol{y}_j \boldsymbol{x}_i\boldsymbol{y}_j^T$$

Now we introduce some important norms used in the proofs. The modified ℓ_∞ norm of matrix Z with respect to X and Y is defined by $\|Z\|_\infty := \max_{i,j} \frac{\sqrt{n_1n_2}}{\mu k}|\boldsymbol{x}_i^T Z\boldsymbol{y}_j|$. This norm generalizes the standard ℓ_∞-norm in the matrix completion problem to our setting. Another norm, the $\ell_{\infty,2}$ norm of Z with respect to X and Y, which enables us to discard the assumption of "joint" incoherence, is defined as

$$\begin{aligned}\|Z\|_{\infty,2} := \max\Big\{ &\max_i \sqrt{\frac{n_1}{\mu_0 k}}\|Z^T\boldsymbol{x}_i\|_2, \\ &\max_j \sqrt{\frac{n_2}{\mu_0 k}}\|Z\boldsymbol{y}_j\|_2\Big\}\end{aligned} \tag{53}$$

The following fact is frequently used in the proofs.

$$\begin{aligned}&\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\|_F^2 \\ =&\langle \mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T), \boldsymbol{x}_i\boldsymbol{y}_j^T\rangle \\ =&\|U^T\boldsymbol{x}_i\|_2^2\|\boldsymbol{y}_j\|_2^2 + \|V^T\boldsymbol{y}_j\|_2^2\|\boldsymbol{x}_i\|_2^2 - \|U^T\boldsymbol{x}_i\|_2^2\|V^T\boldsymbol{y}_j\|_2^2 \\ \leq&\frac{p}{c\log(d_1+d_2)}\end{aligned} \tag{54}$$

The following inequality and its corollary is very important for the proof of this section.

Theorem 6 (Matrix Bernstein Inequality, Theorem 1.6 of [22]). *Consider a finite sequence Z_i of independent, random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies $\mathbb{E}[Z_i] = 0$ and $\|Z_i\|_2 \leq R$ almost surely. Define, $\sigma^2 := \max\{\|\sum_i \mathbb{E}[Z_i Z_i^T]\|_2, \|\sum_i \mathbb{E}[Z_i^T Z_i]\|_2\}$. Then, for all $\gamma \geq 0$,*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m} Z_i\right\|_2 \geq \gamma\right) \leq (d_1+d_2)\exp\left(\frac{-m^2\gamma^2}{\sigma^2 + Rm\gamma/3}\right).$$

We will frequently use the following corollary from Theorem 6.

Corollary 1. *Let $Z_1, Z_2, ..., Z_N \in \mathbb{R}^{d_1\times d_2}$ be independent zero mean random matrices. Suppose*

$$\max\{\|\sum_{k=1}^{N}\mathbb{E}(Z_k Z_k^T)\|, \|\sum_{k=1}^{N}\mathbb{E}(Z_k^T Z_k)\|\} \leq \sigma^2$$

and $\|Z_k\| \leq R$ almost surely for all k. Then for any $c > 0$, we have,

$$\|\sum_{k=1}^{N} Z_k\| \leq 2\sqrt{c\sigma^2\log(d_1+d_2)} + cR\log(d_1+d_2).$$

with probability at least $1-(d_1+d_2)^{-c+1}$.

D.2 Lemmas

To prove Theorem 5, we need the following lemmas.

Lemma 4. *Assuming X and Y satisfy the two properties in Lemma 1, and p satisfies the following inequality with a large enough c,*

$$p \geq c\mu_0\mu\frac{k(d_1+d_2)}{n_1 n_2}\log(d_1+d_2), \tag{55}$$

then for any $Z \in \mathbb{R}^{d_1\times d_2}$, w.h.p.

$$\|\mathcal{P}_T(X^T\mathcal{R}_\Omega(X\mathcal{P}_T(Z)Y^T)Y) - \mathcal{P}_T(Z)\|_F \leq \frac{1}{2}\|Z\|_F$$

Proof.

$$\mathcal{P}_T(X^T\mathcal{R}_\Omega(X\mathcal{P}_T(Z)Y^T)Y) - \mathcal{P}_T(Z) = \sum_{i,j}\left(\frac{\delta_{ij}}{p}-1\right)\langle Z, \mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\rangle\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T) =: \sum_{i,j}\mathcal{S}_{ij}(Z)$$

To use Matrix Berstein Inequality, we can view the linear operator $\mathcal{S}_{ij}$ as a $d_1d_2 \times d_1d_2$ square matrix S_{ij}, which acts on the vectorized Z. So $\|S_{ij}\| = \|\mathcal{S}_{ij}\|$, where the first norm is spectral norm of a matrix and the second is the operator

norm of an operator. It can be shown that S_{ij} is symmetric. As $\mathbb{P}(\delta_{ij}=1)=p$, $\mathbb{E}(\mathcal{S}_{ij}(Z))=0$. Besides, we need to bound $\|\mathcal{S}_{ij}\|$ and $\|\sum_{i,j}\mathbb{E}(\mathcal{S}_{ij}^2)\|$.

$$\|\mathcal{S}_{ij}(Z)\|_F \le \frac{1}{p}|\langle Z, \mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\rangle|\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\|_F \le \frac{1}{p}\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\|_F^2\|Z\|_F \le \frac{2}{c\log(2d_1d_2)}\|Z\|_F$$

$$\begin{aligned}
&\left\|\sum_{i,j}\mathbb{E}(\mathcal{S}_{ij}^2)(Z)\right\|_F \\
=&\left\|\sum_{i,j}\frac{1-p}{p}\langle Z,\mathcal{P}_T(\boldsymbol{x}_i,\boldsymbol{y}_j^T)\rangle\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\|_F^2\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\right\|_F \\
\le&\frac{2}{c\log(2d_1d_2)}\left\|\mathcal{P}_T\left(\sum_{i,j}\langle\mathcal{P}_T(Z),\boldsymbol{x}_i\boldsymbol{y}_j^T\rangle\boldsymbol{x}_i\boldsymbol{y}_j^T\right)\right\|_F \\
=&\frac{2}{c\log(2d_1d_2)}\left\|\mathcal{P}_T\left(X^TX\mathcal{P}_T(Z)Y^TY\right)\right\|_F \\
=&\frac{2}{c\log(2d_1d_2)}\|Z\|_F
\end{aligned}$$

Thus applying the Matrix Berstein Inequality, Lemma 4 holds with a sufficiently large c.

Lemma 5. *Assuming X and Y satisfy the two properties in Lemma 1, and p satisfies inequality* (55) *with a large enough c, then for any $Z\in\mathbb{R}^{d_1\times d_2}$, w.h.p.*

$$\|X^T\mathcal{R}_\Omega(XZY^T)Y-Z\| \le \frac{1}{12}(\|Z\|_\infty+\|Z\|_{\infty,2})$$

Proof.

$$X^T\mathcal{R}_\Omega(XZY^T)Y-Z=\sum_{i,j}\left(\frac{\delta_{ij}}{p}-1\right)\langle Z,\boldsymbol{x}_i\boldsymbol{y}_j^T\rangle\boldsymbol{x}_i\boldsymbol{y}_j^T =: \sum_{i,j}S_{ij}$$

Here $S_{ij}\in\mathbb{R}^{d_1\times d_2}$. Using $p\ge 2c\mu_0\mu_1\frac{r\sqrt{d_1d_2}}{n_1n_2}\log(d_1+d_2)$,

$$\|S_{ij}\| \le \frac{1}{p}\left|\boldsymbol{x}_i^TZ\boldsymbol{y}_j\right|\|\boldsymbol{x}_i\|_2\|\boldsymbol{y}_j\|_2 \le \frac{1}{c\log(d_1+d_2)}\|Z\|_\infty$$

Then using $p\ge c\mu_0\mu_1\frac{rd_1}{n_1n_2}\log(d_1+d_2)$, we have

$$\left\|\sum_{i,j}\mathbb{E}(S_{ij}^TS_{ij})\right\| \le \left\|\sum_{i,j}\frac{1}{p}\left|\boldsymbol{x}_i^TZ\boldsymbol{y}_j\right|^2\|\boldsymbol{x}_i\|_2^2\boldsymbol{y}_j\boldsymbol{y}_j^T\right\| \le \frac{1}{c\log(d_1+d_2)}\left\|\sum_j\|Z\|_{\infty,2}^2\boldsymbol{y}_j\boldsymbol{y}_j^T\right\| = \frac{\|Z\|_{\infty,2}^2}{c\log(d_1+d_2)}$$

We can also prove the same bound for $\left\|\sum_{i,j}\mathbb{E}(S_{ij}S_{ij}^T)\right\|$. Now Lemma 5 follows by applying Matrix Berstein Inequality.

Lemma 6. *Assuming X and Y satisfy the two properties in Lemma 1, and p satisfies inequality* (55) *with a large enough c, then for any $Z \in \mathbb{R}^{d_1 \times d_2}$, w.h.p.*

$$\|\mathcal{P}_T(X^T\mathcal{R}_\Omega(XZY^T)Y) - \mathcal{P}_T(Z)\|_{\infty,2} \le \frac{1}{2}(\|Z\|_\infty + \|Z\|_{\infty,2})$$

Proof.

$$\sqrt{\frac{n_2}{\mu_0 r}}\left(\mathcal{P}_T(X^T\mathcal{R}_\Omega(XZY^T)Y)\boldsymbol{y}_b - \mathcal{P}_T(Z)\boldsymbol{y}_b\right) = \sum_{i,j}\left(\frac{\delta_{ij}}{p} - 1\right)\sqrt{\frac{n_2}{\mu_0 r}}\langle Z, \boldsymbol{x}_i\boldsymbol{y}_j^T\rangle\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\boldsymbol{y}_b =: \sum_{i,j} S_{ij}$$

Here $S_{ij} \in \mathbb{R}^{d_1 \times 1}$.

$$\begin{aligned}
\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\boldsymbol{y}_b\|_2 &= \|UU^T\boldsymbol{x}_i\boldsymbol{y}_j^T\boldsymbol{y}_b + (I_{d_1} - UU^T)\boldsymbol{x}_i\boldsymbol{y}_j^TVV^T\boldsymbol{y}_b\|_2 \\
&\le \|U^T\boldsymbol{x}_i\|_2\|\boldsymbol{y}_j\|_2\|\boldsymbol{y}_b\|_2 + \|\boldsymbol{x}_i\|_2\|V^T\boldsymbol{y}_j\|_2\|V^T\boldsymbol{y}_b\|_2 \\
&\le \sqrt{\frac{\mu_0 r}{n_1}}\frac{\mu_1 d_2}{n_2} + \sqrt{\frac{\mu_1 d_1}{n_1}}\frac{\mu_0 r}{n_2}
\end{aligned}$$

Using the fact that $\frac{\mu_0 r}{n_1} \le \frac{\mu_1 d_1}{n_1}$ because $\|U^T\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_2$, and the inequality condition (55) of p, we have,

$$\|S_{ij}\|_2 \le \frac{\|Z\|_\infty}{c\log(d_1+d_2)} \le \frac{\|Z\|_\infty}{c\log(d_1+1)}$$

$$\begin{aligned}
\left\|\sum_{i,j}\mathbb{E}(S_{ij}^TS_{ij})\right\| &\le \left|\sum_{i,j}\frac{1}{p}\frac{n_2}{\mu_0 r}(\boldsymbol{x}_i^TZ\boldsymbol{y}_j)^2\left[\|U^T\boldsymbol{x}_i\|_2^2(\boldsymbol{y}_j^T\boldsymbol{y}_b)^2 + \|\boldsymbol{x}_i\|_2^2(\boldsymbol{y}_j^TVV^T\boldsymbol{y}_b)^2\right]\right| \\
&\le \left|\sum_{i,j}\frac{1}{p}\frac{n_2}{\mu_0 r}(\boldsymbol{x}_i^TZ\boldsymbol{y}_j)^2\left[\frac{\mu_0 r}{n_1}(\boldsymbol{y}_j^T\boldsymbol{y}_b)^2 + \frac{\mu_1 d_1}{n_1}(\boldsymbol{y}_j^TVV^T\boldsymbol{y}_b)^2\right]\right| \\
&\le \left|\|Z\|_{\infty,2}^2\frac{1}{p}\left[\frac{\mu_0 r}{n_1}\sum_j(\boldsymbol{y}_j^T\boldsymbol{y}_b)^2 + \frac{\mu_1 d_1}{n_1}\sum_j(\boldsymbol{y}_j^TVV^T\boldsymbol{y}_b)^2\right]\right| \\
&\le \frac{\|Z\|_\infty^2}{c\log(d_1+1)}
\end{aligned}$$

As S_{ij} is a vector, $\left\|\sum_{i,j}\mathbb{E}(S_{ij}S_{ij}^T)\right\| = \left\|\sum_{i,j}\mathbb{E}(S_{ij}^TS_{ij})\right\|$.
Now Lemma 6 follows by applying Matrix Berstein Inequality

Lemma 7. *Assuming X and Y satisfy the two properties in Lemma 1, and p satisfies inequality* (55) *with a large enough c, then for any $Z \in \mathbb{R}^{d_1 \times d_2}$, w.h.p.*

$$\|\mathcal{P}_T(X^T\mathcal{R}_\Omega(XZY^T)Y) - \mathcal{P}_T(Z)\|_\infty \le \frac{1}{2}\|Z\|_\infty$$

Proof.

$$\frac{\sqrt{n_1 n_2}}{\mu_0 r}\boldsymbol{x}_a^T\left(\mathcal{P}_T(X^T\mathcal{R}_\Omega(XZY^T)Y)-\mathcal{P}_T(Z)\right)\boldsymbol{y}_b=\sum_{i,j}\frac{\sqrt{n_1 n_2}}{\mu_0 r}\left(\frac{\delta_{ij}}{p}-1\right)\boldsymbol{x}_i^T Z\boldsymbol{y}_j\langle\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T),\boldsymbol{x}_a\boldsymbol{y}_b^T\rangle=:\sum_{i,j}s_{ij}$$

$$|s_{ij}|\le\frac{\sqrt{n_1 n_2}}{\mu_0 r}\frac{1}{p}\left|\boldsymbol{x}_i^T Z\boldsymbol{y}_j\right|\|\mathcal{P}_T(\boldsymbol{x}_i\boldsymbol{y}_j^T)\|_F\|\mathcal{P}_T(\boldsymbol{x}_a\boldsymbol{y}_b^T)\|_F\le\frac{1}{c\log(2)}\frac{\sqrt{n_1 n_2}}{\mu_0 r}\left|\boldsymbol{x}_i^T Z\boldsymbol{y}_j\right|\|\le\frac{1}{c\log(2)}\|Z\|_\infty$$

$$\begin{aligned}\left\|\sum_{i,j}\mathbb{E}(s_{ij}^2)\right\|&\le\left|\sum_{i,j}\frac{1}{p}\frac{n_1 n_2}{(\mu_0 r)^2}(\boldsymbol{x}_i^T Z\boldsymbol{y}_j)^2\langle\boldsymbol{x}_i\boldsymbol{y}_j^T,\mathcal{P}_T(\boldsymbol{x}_a\boldsymbol{y}_b^T)\rangle^2\right|\\&\le\|Z\|_\infty^2\left|\sum_{i,j}\frac{1}{p}\langle\boldsymbol{x}_i\boldsymbol{y}_j^T,\mathcal{P}_T(\boldsymbol{x}_a\boldsymbol{y}_b^T)\rangle^2\right|\\&=\|Z\|_\infty^2\frac{1}{p}\|\mathcal{P}_T(\boldsymbol{x}_a\boldsymbol{y}_b^T)\|_F^2\\&\le\frac{1}{c\log(2)}\|Z\|_\infty^2\end{aligned}$$

Now Lemma 7 follows by applying Matrix Berstein Inequality

Lemma 8. *Assuming X and Y satisfy the two properties in Lemma 1, and p satisfies the following inequality with a large enough c*

$$p\ge\max\left\{c_0\mu_0\mu\frac{k(d_1+d_2)}{n_1 n_2}\log(d_1+d_2),\frac{1}{n^{10}}\right\}$$

where $n=\min\{n_1^{10},n_2^{10}\}$, then for any matrix, $\Delta\in\mathbb{R}^{d_1\times d_2}$, such that $\mathcal{R}_\Omega(X\Delta Y^T)=0$, we have, w.h.p.

$$\|\mathcal{P}_T(\Delta)\|_F\le\sqrt{2}n^5\|\mathcal{P}_{T^\perp}(\Delta)\|_*$$

Proof. Define a sampling operator,

$$\mathcal{R}_\Omega^{1/2}(\hat{Z}):=\sum_{i,j}\frac{\delta_{ij}}{\sqrt{p}}\hat{Z}_{ij}\boldsymbol{e}_i\boldsymbol{e}_j^T$$

As $\mathcal{R}_\Omega(X\Delta Y^T)$=0, $\mathcal{R}_\Omega^{1/2}(X\Delta Y^T)=0$. Thus,

$$\|\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_T(\Delta)Y^T)\|_F=\|\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_{T^\perp}(\Delta)Y^T)\|_F$$

$$\begin{aligned}\|\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_T(\Delta)Y^T)\|_F^2&=\langle\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_T(\Delta)Y^T),\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_T(\Delta)Y^T)\rangle\\&=\langle X\mathcal{P}_T(\Delta)Y^T,\mathcal{R}_\Omega(X\mathcal{P}_T(\Delta)Y^T)\rangle\\&=\langle\mathcal{P}_T(\Delta),X\mathcal{R}_\Omega(X\mathcal{P}_T(\Delta)Y^T)Y^T\rangle\\&=\langle\mathcal{P}_T(\Delta),\mathcal{P}_T(\Delta)\rangle-\langle\mathcal{P}_T(\Delta),\mathcal{P}_T(\Delta)-X^T\mathcal{R}_\Omega(X\mathcal{P}_T(\Delta)Y^T)Y\rangle\\&\ge\frac{1}{2}\|\mathcal{P}_T(\Delta)\|_F^2\end{aligned}$$

The last inequality comes from Lemma 4.

$$\|\mathcal{R}_\Omega^{1/2}(X\mathcal{P}_{T^\perp}(\Delta)Y^T)\|_F \leq \frac{1}{\sqrt{p}}\|X\mathcal{P}_{T^\perp}(\Delta)Y^T\|_F \leq n^5\|\mathcal{P}_{T^\perp}(\Delta)\|_F$$

Applying the fact that $\|Z\|_* \geq \|Z\|_F$, we have w.h.p.

$$\|\mathcal{P}_T(\Delta)\|_F \leq \sqrt{2}n^5\|\mathcal{P}_{T^\perp}(\Delta)\|_*$$

D.3 Dual Certificate

Proposition 1. *For any $\Delta \in \mathbb{R}^{d_1\times d_2}$, with $\mathcal{R}_\Omega(X\Delta Y^T) = 0$, if there exists a dual certificate $M \in \mathbb{R}^{d_1\times d_2}$ satisfying the following conditions,*

$$\begin{aligned} &\boldsymbol{A}1. \quad \langle M, \Delta\rangle = 0 \\ &\boldsymbol{A}2. \quad \|UV^T - \mathcal{P}_T(M)\|_F \leq \frac{1}{4n^5} \\ &\boldsymbol{A}3. \quad \|\mathcal{P}_{T^\perp}(M)\| \leq \frac{1}{2} \end{aligned} \tag{56}$$

then $\|W_* + \Delta\|_* > \|W_*\|_*$, i.e. the solution of problem (9) is unique and equal to W_*.

Proof. We assume the SVD of $\mathcal{P}_{T^\perp}(\Delta)$ is $U_\perp\Sigma_\perp V_\perp^T$. So $\langle U_\perp V_\perp^T, \mathcal{P}_{T^\perp}(\Delta)\rangle = \|\mathcal{P}_{T^\perp}(\Delta)\|_*$. Then we have

$$\begin{aligned} \|W_* + \Delta\|_* &\geq \langle UV^T + U_\perp V_\perp^T, W_* + \Delta\rangle \\ &= \langle UV^T, W_* + \Delta\rangle + \langle U_\perp V_\perp^T, W_* + \Delta\rangle \\ &= \|W_*\|_* + \langle UV^T, \Delta\rangle + \langle U_\perp V_\perp^T, \Delta\rangle \\ &= \|W_*\|_* + \langle UV^T - M, \Delta\rangle + \langle U_\perp V_\perp^T, \Delta\rangle \\ &= \|W_*\|_* + \langle UV^T - \mathcal{P}_T(M) - \mathcal{P}_{T^\perp}(M), \mathcal{P}_T(\Delta) + \mathcal{P}_{T^\perp}(\Delta)\rangle + \langle U_\perp V_\perp^T, \mathcal{P}_T(\Delta) + \mathcal{P}_{T^\perp}(\Delta)\rangle \\ &= \|W_*\|_* + \langle UV^T - \mathcal{P}_T(M), \mathcal{P}_T(\Delta)\rangle + \langle U_\perp V_\perp^T - \mathcal{P}_{T^\perp}(M), \mathcal{P}_{T^\perp}(\Delta)\rangle \\ &\geq \|W_*\|_* - \|UV^T - \mathcal{P}_T(M)\|_F\|\mathcal{P}_T(\Delta)\|_F + \|\mathcal{P}_{T^\perp}(\Delta)\|_* - \|\mathcal{P}_{T^\perp}(M)\|\|\mathcal{P}_{T^\perp}(\Delta)\|_* \\ &> \|W_*\|_* - \frac{1}{2}\|\mathcal{P}_{T^\perp}(\Delta)\|_* + \|\mathcal{P}_{T^\perp}(\Delta)\|_* - \frac{1}{2}\|\mathcal{P}_{T^\perp}(\Delta)\|_* \\ &= \|W_*\|_* \end{aligned}$$

The first inequality follows from the fact that the dual norm of spectral norm is the nuclear norm. The details can be found in Lemma 3.2 in [3]. We also use this fact to prove $\langle \mathcal{P}_{T^\perp}(M), \mathcal{P}_{T^\perp}(\Delta)\rangle \leq \|\mathcal{P}_{T^\perp}(M)\|\|\mathcal{P}_{T^\perp}(\Delta)\|_*$ in the second inequality. The third equality comes from Condition $\boldsymbol{A}1$. The last inequality comes from Condition $\boldsymbol{A}2$, Lemma 8 and Condition $\boldsymbol{A}3$.

D.4 Build the Dual Certificate

The remainder of the proof shows that there exists such a M with high probability. We use golfing scheme [6] to construct it. Let $\Omega = \cup_{k=1}^{k_0} \Omega_j$, where $k_0 = 20\log(n_1 + n_2)$ and Ω_k is sampled independently of $\Omega_{k'}$, $k' \neq k$. The sampling rule of Ω_k is also based on Bernoulli model with an identical probability $q = 1 - (1-p)^{1/k_0}$. As $p = 1 - (1-q)^{k_0} < k_0 q$, if p satisfies the inequality (11), then q will satisfies the inequality (55).
Similarly, we define the sampling operator for every Ω_k, such that.

$$X^T \mathcal{R}_{\Omega_k}(XZY^T)Y = \sum_{i,j} \frac{\delta_{ij}^{(k)}}{q} \boldsymbol{x}_i^T Z \boldsymbol{y}_j \boldsymbol{x}_i \boldsymbol{y}_j^T$$

where $\delta_{ij}^{(k)} = \mathbb{I}[(i,j) \in \Omega_k]$.
So the lemmas from Lemma 4 to Lemma 7 hold for every Ω_k. Now we can construct a sequence,

$$\begin{aligned} S_0 &= UV^T \\ M_k &= \sum_{l=1}^{k} X^T \mathcal{R}_{\Omega_l}(XS_{l-1}Y^T)Y, \quad k = 1, 2...k_0 \\ S_k &= UV^T - \mathcal{P}_T(M_k), \quad k = 1, 2...k_0 \end{aligned}$$

We can prove that $M = M_{k_0}$ will satisfy the three conditions. For the first condition $\boldsymbol{A}1$,

$$\langle M, \Delta \rangle = \langle \sum_{l=1}^{k_0} \mathcal{R}_{\Omega_l}(XS_{l-1}Y^T), X\Delta Y^T \rangle = 0$$

According to the construction of the sequence,

$$S_k = \mathcal{P}_T(S_{k-1}) - \mathcal{P}_T(X^T \mathcal{R}_{\Omega_k}(XS_{k-1}Y^T)Y)$$

$$\|S_k\|_F = \|\mathcal{P}_T(S_{k-1}) - \mathcal{P}_T(X^T \mathcal{R}_{\Omega_k}(XS_{k-1}Y^T)Y)\|_F \leq \frac{1}{2}\|S_{k-1}\|_F$$

Condition $\boldsymbol{A}2$ follows: $\|UV^T - \mathcal{P}_T(M_{k_0})\|_F \leq \frac{1}{2^{k_0}}\|UV^T\|_F \leq \frac{\sqrt{k}}{2^{20\log n}} < \frac{1}{4n^5}$.
To see Condition $\boldsymbol{A}3$,

$$\begin{aligned}
\|\mathcal{P}_{T^\perp}(M_{k_0})\| &= \|\mathcal{P}_{T^\perp}(\sum_{l=1}^{k_0} X^T\mathcal{R}_{\Omega_l}(XS_{l-1}Y^T)Y)\| \\
&\leq \sum_{l=1}^{k_0}\|\mathcal{P}_{T^\perp}(X^T\mathcal{R}_{\Omega_l}(XS_{l-1}Y^T)Y)\| \\
&= \sum_{l=1}^{k_0}\|\mathcal{P}_{T^\perp}(X^T\mathcal{R}_{\Omega_l}(XS_{l-1}Y^T)Y - S_{l-1})\| \\
&\leq \sum_{l=1}^{k_0}\|X^T\mathcal{R}_{\Omega_l}(XS_{l-1}Y^T)Y - S_{l-1}\|
\end{aligned}$$

The second inequality uses the fact that $\|\mathcal{P}_{T^\perp}(Z)\| = \|(I_{d_1} - P_U)Z(I_{d_2} - P_V)\| \leq \|Z\|$.
According to Lemma 5, we have

$$\|X^T\mathcal{R}_{\Omega_k}(XS_{k-1}Y^T)Y - S_{k-1}\| \leq \frac{1}{12}(\|S_{k-1}\|_\infty + \|S_{k-1}\|_{\infty,2})$$

According to Lemma 6 and Lemma 7, we have

$$\begin{aligned}
\|S_k\|_{\infty,2} &\leq \frac{1}{2}(\|S_{k-1}\|_\infty + \|S_{k-1}\|_{\infty,2}) \\
\|S_k\|_\infty &\leq \frac{1}{2}\|S_{k-1}\|_\infty
\end{aligned}$$

So $\|S_k\|_\infty \leq \frac{1}{2^k}\|UV^T\|_\infty \leq \frac{1}{2^k}$. For the $\ell_{\infty,2}$ norm,

$$\begin{aligned}
\|S_k\|_{\infty,2} &\leq \frac{1}{2}(\|S_{k-1}\|_\infty + \|S_{k-1}\|_{\infty,2}) \\
&\leq \frac{1}{2}\|S_{k-1}\|_{\infty,2} + \frac{1}{2^k} \\
&\leq \frac{1}{2^2}\|S_{k-2}\|_{\infty,2} + 2\frac{1}{2^k} \\
&\leq .... \\
&\leq \frac{1}{2^k}\|S_0\|_{\infty,2} + k\frac{1}{2^k} \\
&\leq (k+1)\frac{1}{2^k}
\end{aligned}$$

Thus, $\|X^T\mathcal{R}_{\Omega_k}(XS_{k-1}Y^T)Y - S_{k-1}\| \leq \frac{k+1}{3\times 2^{k+1}}$.

$$\|\mathcal{P}_{T^\perp}(M_{k_0})\| \leq \sum_{k=1}^{k_0}\frac{k+1}{3\times 2^{k+1}} < \frac{1}{2}$$