

Provable Inductive Matrix Completion

Prateek Jain

Microsoft Research India, Bangalore
prajain@microsoft.com

Inderjit S. Dhillon

The University of Texas at Austin
inderjit@cs.utexas.edu

Abstract

Consider a movie recommendation system where apart from the ratings information, side information such as user’s age or movie’s genre is also available. Unlike standard matrix completion, in this setting one should be able to predict inductively on new users/movies. In this paper, we study the problem of inductive matrix completion in the exact recovery setting. That is, we assume that the ratings matrix is generated by applying feature vectors to a low-rank matrix and the goal is to recover back the underlying matrix. Furthermore, we generalize the problem to that of low-rank matrix estimation using rank-1 measurements. We study this generic problem and provide conditions that the set of measurements should satisfy so that the alternating minimization method (which otherwise is a non-convex method with no convergence guarantees) is able to recover back the *exact* underlying low-rank matrix.

In addition to inductive matrix completion, we show that two other low-rank estimation problems can be studied in our framework: a) general low-rank matrix sensing using rank-1 measurements, and b) multi-label regression with missing labels. For both the problems, we provide novel and interesting bounds on the number of measurements required by alternating minimization to provably converges to the *exact* low-rank matrix. In particular, our analysis for the general low rank matrix sensing problem significantly improves the required storage and computational cost than that required by the RIP-based matrix sensing methods [1]. Finally, we provide empirical validation of our approach and demonstrate that alternating minimization is able to recover the true matrix for the above mentioned problems using a small number of measurements.

1 Introduction

Motivated by the Netflix Challenge, recent research has addressed the problem of matrix completion where the goal is to recover the underlying low-rank “ratings” matrix by using a small number of observed entries of the matrix. However, the standard low-rank matrix completion formulation is applicable only to the transductive setting only, i.e., predictions are restricted to the existing users/movies only. However, several real-world recommendation systems have useful side-information available in the form of feature vectors for users as well as movies, and hence one should be able to make accurate predictions for new users and movies as well.

In this paper, we formulate and study the above mentioned problem which we call inductive matrix completion, where other than a small number of observations from the ratings matrix, the feature vectors for users/movies are also available. We formulate the problem as that of recovering a low-rank matrix W_* using observed entries $R_{ij} = \mathbf{x}_i^T W_* \mathbf{y}_j$ and the user/movie feature vectors $\mathbf{x}_i, \mathbf{y}_j$. By factoring $W_* = U_* V_*^T$, we see that this scheme constitutes a bi-linear prediction $(\mathbf{x}^T U_*)(V_*^T \mathbf{y})$ for a new user/movie pair (\mathbf{x}, \mathbf{y}) .

In fact, the above rank-1 measurement scheme also arises in several other important low-rank estimation problems such as: a) general low-rank matrix sensing in the signal acquisition domain, and b) multi-label regression problem with missing information.

In this paper, we generalize the above three mentioned problems to the following low-rank matrix estimation problem that we call *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:

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

where $\mathbf{x}_i, \mathbf{y}_i$ are “feature” vectors and are provided along with the measurements \mathbf{b} .

Now given measurements \mathbf{b} and the feature vectors $\{\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_m\}$, $Y = \{\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_m\}$, a canonical way to recover W_* is to find a rank- k matrix W such that $\|\mathcal{A}(W) - \mathbf{b}\|_2$ is small. While the objective function of this problem is simple least squares, the non-convex rank constraint makes it NP-hard, in general, to solve. In existing literature, there are two common approaches to handle such low-rank problems: a) Use trace-norm constraint as a proxy for the rank constraint and then solve the resulting non-smooth convex optimization problem, b) Parameterize W as $W = UV^T$ and then alternately optimize for U and V .

The first approach has been shown to be successful for a variety of problems such as matrix completion [2, 3, 4, 5], general low-rank matrix sensing [1], robust PCA [6, 7], etc. However, the resulting convex optimization methods require computation of full SVD of matrices with potentially large rank and hence do not scale to large scale problems. On the other hand, alternating minimization and its variants need to solve only least squares problems and hence are scalable in practice but might get stuck in a local minima. However, [8] recently showed that under standard set of assumptions, alternating minimization actually converges at a linear rate to the global optimum of two low-rank estimation problems: a) RIP measurements based general low-rank matrix sensing, and b) low-rank matrix completion.

Motivated by its empirical as well as theoretical success, we study a variant of alternating minimization (with appropriate initialization) for the above mentioned LRROM problem. To analyze our general LRROM problem, we present three key properties that a rank-1 measurement operator should satisfy. Assuming these properties, we show that the alternating minimization method converges to the global optima of LRROM at a linear rate. We then study the three problems individually and show that for each of the problems, the measurement operator indeed satisfies the conditions required by our general analysis and hence, for each of the problems alternating minimization converges to the global optimum at a linear rate. Below, we briefly describe the three application problems that we study and also our high-level result for each one of them:

(a) Efficient matrix sensing using Gaussian Measurements: In this problem, $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{y}_i \in \mathbb{R}^{d_2}$ are sampled from a sub-Gaussian distribution and the goal is efficient acquisition and recovery of rank- k matrix W_* . Here, we show that if the number of measurements $m = \Omega(k^4 \beta^2 (d_1 + d_2) \log(d_1 + d_2))$, where $\beta = \sigma_*^1 / \sigma_*^k$ is the condition number of W_* . Then with high probability (w.h.p.), our alternating minimization based method will recover back W_* in linear time.

Note that the problem of low-rank matrix sensing has been considered by several existing methods [1, 9, 10], however most of these methods require the measurement operator to satisfy the Restricted Isometry Property (RIP) (see Definition 2). Typically, RIP operators are constructed by sampling from distributions with bounded fourth moments and require $m = O(k(d_1 + d_2) \log(d_1 +$

d_2)) measurements to satisfy RIP for a constant $\delta > 0$. That is, the number of samples required to satisfy RIP are similar to the number of samples required by our method.

Moreover, RIP based operators are typically dense, have a large memory footprint and make the algorithm computationally intensive. For example, assuming rank and β to be constant, RIP based operators would require $O((d_1 + d_2)d_1d_2)$ storage and computational time, as opposed to $O((d_1 + d_2)^2)$ storage and computational time required by the rank-1 measurement operators. However, a drawback of such rank-1 measurements is that, unlike RIP based operators, they are not universal, i.e., a new set of $\mathbf{x}_i, \mathbf{y}_i$ needs to be sampled for any given signal W_* .

(b) Inductive Matrix Completion: As motivated earlier, consider a movie recommendation system with n_1 users and n_2 movies. Let $X \in \mathbb{R}^{n_1 \times d_1}$, $Y \in \mathbb{R}^{n_2 \times d_2}$ be feature matrices of the users and the movies, respectively. Then, the user-movie rating R_{ij} can be modeled as $R_{ij} = \mathbf{x}_i^T W \mathbf{y}_j$ and the goal is to learn W using a small number of random ratings indexed by the set of observations $\Omega \in [n_1] \times [n_2]$. Note that matrix completion is a special case of this problem when $\mathbf{x}_i = \mathbf{e}_i$ and $\mathbf{y}_j = \mathbf{e}_j$. Also, unlike standard matrix completion, accurate ratings can be predicted for users who have not rated any prior movies and vice versa.

If the feature matrices X, Y are incoherent and the number of observed entries $|\Omega| = m \geq C \cdot (k^3 \beta^2 (d_1 \cdot d_2) \log(d_1 + d_2))$, then inductive matrix completion satisfies the conditions required by our generic method and hence the global optimality result follows directly. Note that our analysis requires a quadratic number of samples, i.e., $\tilde{O}(d_1 \cdot d_2)$ samples (assuming k to be a constant) for recovery. On the other hand, applying standard matrix completion would require $\tilde{O}(n_1 + n_2)$ samples. Hence, our analysis provides significant improvement if $d_1 \cdot d_2 \ll n_1 + n_2$, i.e., when the number of features is significantly smaller than the total number of users and movies.

(c) Multi-label Regression with Missing Data: Consider a multi-variate regression problem, where the goal is to predict a set of (correlated) target variables $\mathbf{r} \in \mathbb{R}^L$ for a given $\mathbf{x} \in \mathbb{R}^{d_1}$. We model this problem as a regression problem with low-rank parameters, i.e., $\mathbf{r} = W^T \mathbf{x}$ where W is a low-rank matrix. Given training data points $X = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_{n_1}]$ and the associated target matrix R , W can be learned using a simple least squares regression. However, in most real-world applications several of the entries in R are missing and the goal is to be able to learn W “exactly”.

Now, let the set of known entries $R_{ij}, (i, j) \in \Omega$ be sampled uniformly at random from R . Then we show that, by sampling $|\Omega| = m \geq k^3 \beta^2 \cdot (d_1 \cdot L) \cdot \log(d_1 + L)$ entries, alternating minimization recovers back W_* exactly. Note that a direct approach to this problem is to first recover the label matrix R using standard matrix completion and then learn W_* from the completed label matrix. Such a method would require $\tilde{O}(n_1 + L)$ samples of R . In contrast, our more unified approach requires $\tilde{O}(d_1 \cdot L)$ samples. Hence, if the number of training points n_1 is much larger than the number of labels L , then our method provides significant improvement over first completing the matrix and then learning the true low-rank matrix.

We would like to stress that the above mentioned problems of inductive matrix completion and multi-label regression with missing labels have recently received a lot of attention from the machine learning community [11, 12]. However, to the best of our knowledge, our results are the first theoretically rigorous results that improve upon the sample complexity of first completing the target/ratings matrix and then learning the parameter matrix W_* .

Related Work: Low-rank matrix estimation problems are pervasive and have innumerable real-life applications. Popular examples of low-rank matrix estimation problems include PCA, robust PCA, non-negative matrix approximation, low-rank matrix completion, low-rank matrix

Algorithm 1 AltMin-LRROM : Alternating Minimization for LRROM

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

sensing etc. While in general low-rank matrix estimation that satisfies given (affine) observations is NP-hard, several recent results present conditions under which the optimal solution can be recovered exactly or approximately [2, 1, 3, 13, 7, 6, 8, 9].

Of these above mentioned low-rank matrix estimation problems, the most relevant problems to ours are those of matrix completion [2, 5, 8] and general matrix sensing [1, 9, 10]. The matrix completion problem is restricted to a given set of users and movies and hence does not generalize to new users/movies. On the other hand, matrix sensing methods require the measurement operator to satisfy the RIP condition, which at least for the current constructions, necessitate measurement matrices that have full rank, large number of random bits and hence high storage as well as computational time [1]. Our work on general low-rank matrix estimation (problem (a) above) alleviates this issue as our measurements are only rank-1 and hence the low-rank signal W_* can be encoded as well as decoded much more efficiently. Moreover, our result for inductive matrix completion generalizes the matrix completion work and provides, to the best of our knowledge, the first theoretical results for the problem of inductive matrix completion.

Paper Organization: We formally introduce the problem of low-rank matrix estimation with rank-one measurements in Section 2. We provide our version of the alternating minimization method and then we present a *generic* analysis for alternating minimization when applied to such rank-one measurements based problems. Our results distill out certain key problem specific properties that would imply global optimality of alternating minimization. In the subsequent sections 3, 4, and 5, we show that for each of our three problems (mentioned above) the required problem specific properties are satisfied and hence our alternating minimization method provides globally optimal solution. Finally, we provide empirical validation of our methods in Section 6.

2 Low-rank Matrix Estimation using Rank-one Measurements

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^m$ be a linear measurement operator parameterized by $\mathcal{A} = \{A_1, A_2, \dots, A_m\}$, where $A_i \in \mathbb{R}^{d_1 \times d_2}$. Then, the linear measurements of a given matrix $W \in \mathbb{R}^{d_1 \times d_2}$ are given by:

$$\mathcal{A}(W) = [\operatorname{Tr}(A_1^T W) \ \operatorname{Tr}(A_2^T W) \ \dots \ \operatorname{Tr}(A_m^T W)]^T, \quad (1)$$

where Tr denotes the trace operator.

In this paper, we mainly focus on the rank-1 measurement operators, i.e., $A_i = \mathbf{x}_i \mathbf{y}_i^T, 1 \leq i \leq m$ where $\mathbf{x}_i \in \mathbb{R}^{d_1}, \mathbf{y}_i \in \mathbb{R}^{d_2}$. Also, let $W_* \in \mathbb{R}^{d_1 \times d_2}$ be a rank- k matrix, with the singular value decomposition (SVD) $W_* = U_* \Sigma_* V_*^T$.

Then, given \mathcal{A}, \mathbf{b} , the goal of the LRROM problem is to recover back W_* efficiently. This problem can be reformulated as the following non-convex optimization problem:

$$(\text{LRROM}) : \min_{W=UV^T, U \in \mathbb{R}^{d_1 \times k}, V \in \mathbb{R}^{d_2 \times k}} \sum_{i=1}^m (b_i - \mathbf{x}_i^T W \mathbf{y}_i)^2. \quad (2)$$

Note that W to be recovered is restricted to have at most rank- k and hence W can be re-written as $W = UV^T$.

We use the standard alternating minimization algorithm with appropriate initialization to solve the above problem (2) (see Algorithm 1). Note that the above problem is non-convex in U, V and hence standard analysis would only ensure convergence to a local minima. However, [8] 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 2), even when the vectors $\mathbf{x}_i, \mathbf{y}_i$ are sampled from the normal distribution (see Claim 3). Furthermore, each measurement need not reveal exactly one entry of W_* as in the case of matrix completion. Hence, the proof of [8] does not apply directly. However, inspired by the proof of [8], 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 $\sigma_*^1 \geq \sigma_*^2 \dots \geq \sigma_*^k$. Also, let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^m$ be a linear measurement operator parameterized by m matrices, i.e., $\mathcal{A} = \{A_1, A_2, \dots, A_m\}$ where $A_i = \mathbf{x}_i \mathbf{y}_i^T$. 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:** $\|\frac{1}{m} \sum_i b_i A_i - W_*\|_2 \leq \|W_*\|_2 \cdot \delta$.
2. **Concentration of operators B_x, B_y :** Let $B_x = \frac{1}{m} \sum_{i=1}^m (\mathbf{y}_i^T \mathbf{v})^2 \mathbf{x}_i \mathbf{x}_i^T$ and $B_y = \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i^T \mathbf{u})^2 \mathbf{y}_i \mathbf{y}_i^T$, where $\mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}$ are two unit vectors that are independent of randomness in $\mathbf{x}_i, \mathbf{y}_i, \forall i$. 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 $G_x = \frac{1}{m} \sum_i (\mathbf{y}_i^T \mathbf{v})(\mathbf{y}_i \mathbf{v}_\perp) \mathbf{x}_i \mathbf{x}_i^T$, $G_y = \frac{1}{m} \sum_i (\mathbf{x}_i^T \mathbf{u})(\mathbf{u}_\perp^T \mathbf{x}_i) \mathbf{y}_i \mathbf{y}_i^T$, where $\mathbf{u}, \mathbf{u}_\perp \in \mathbb{R}^{d_1}, \mathbf{v}, \mathbf{v}_\perp \in \mathbb{R}^{d_2}$ are unit vectors, s.t., $\mathbf{u}^T \mathbf{u}_\perp = 0$ and $\mathbf{v}^T \mathbf{v}_\perp = 0$. Furthermore, let $\mathbf{u}, \mathbf{u}_\perp, \mathbf{v}, \mathbf{v}_\perp$ be independent of randomness in $\mathbf{x}_i, \mathbf{y}_i, \forall i$. 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 $H \leq 100 \log(\|W_*\|_F / \epsilon)$.

Proof. We explain the key ideas of the proof by first presenting the proof for the special case of rank-1 $W_* = \sigma_* \mathbf{u}_* \mathbf{v}_*^T$. Later in Appendix B, we extend the proof to general rank- k case.

Similar to [8], we first characterize the update for $h+1$ -th step iterates $\hat{\mathbf{v}}_{h+1}$ of Algorithm 1 and its normalized form $\mathbf{v}_{h+1} = \hat{\mathbf{v}}_{h+1} / \|\hat{\mathbf{v}}_{h+1}\|_2$.

Now, by gradient of (2) w.r.t. $\hat{\mathbf{v}}$ to be zero while keeping \mathbf{u}_h to be fixed. That is,

$$\begin{aligned}
& \sum_{i=1}^m (b_i - \mathbf{x}_i^T \mathbf{u}_h \hat{\mathbf{v}}_{h+1}^T \mathbf{y}_i) (\mathbf{x}_i^T \mathbf{u}_h) \mathbf{y}_i = 0, \\
& i.e., \sum_{i=1}^m (\mathbf{u}_h^T \mathbf{x}_i) \mathbf{y}_i (\sigma_* \mathbf{y}_i^T \mathbf{v}_* \mathbf{u}_*^T \mathbf{x}_i - \mathbf{y}_i^T \hat{\mathbf{v}}_{h+1} \mathbf{u}_h^T \mathbf{x}_i) = 0, \\
& i.e., \left(\sum_{i=1}^m (\mathbf{x}_i^T \mathbf{u}_h \mathbf{u}_h^T \mathbf{x}_i) \mathbf{y}_i \mathbf{y}_i^T \right) \hat{\mathbf{v}}_{h+1} = \sigma_* \left(\sum_{i=1}^m (\mathbf{x}_i^T \mathbf{u}_h \mathbf{u}_*^T \mathbf{x}_i) \mathbf{y}_i \mathbf{y}_i^T \right) \mathbf{v}_*, \\
& i.e., \hat{\mathbf{v}}_{h+1} = \sigma_* (\mathbf{u}_*^T \mathbf{u}_h) \mathbf{v}_* - \sigma_* B^{-1} ((\mathbf{u}_*^T \mathbf{u}_h) B - \tilde{B}) \mathbf{v}_*,
\end{aligned} \tag{3}$$

where,

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

Note that (3) shows that $\hat{\mathbf{v}}_{h+1}$ is a perturbation of \mathbf{v}_* and the goal now is to bound the spectral norm of the perturbation term:

$$\|G\|_2 = \|B^{-1}(\mathbf{u}_*^T \mathbf{u}_h B - \tilde{B}) \mathbf{v}_*\|_2 \leq \|B^{-1}\|_2 \|\mathbf{u}_*^T \mathbf{u}_h B - \tilde{B}\|_2 \|\mathbf{v}_*\|_2. \tag{4}$$

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

$$\|B - I\|_2 \leq 1/100, \quad i.e., \sigma_{\min}(B) \geq 1 - 1/100, \quad i.e., \|B^{-1}\|_2 \leq 1/(1 - 1/100). \tag{5}$$

Now,

$$\begin{aligned}
(\mathbf{u}_*^T \mathbf{u}_h) B - \tilde{B} &= \frac{1}{m} \sum_{i=1}^m \mathbf{y}_i \mathbf{y}_i^T \mathbf{x}_i^T ((\mathbf{u}_*^T \mathbf{u}_h) \mathbf{u}_h \mathbf{u}_h^T - \mathbf{u}_* \mathbf{u}_h^T) \mathbf{x}_i, \\
&= \frac{1}{m} \sum_{i=1}^m \mathbf{y}_i \mathbf{y}_i^T \mathbf{x}_i^T (\mathbf{u}_h \mathbf{u}_h^T - I) \mathbf{u}_* \mathbf{u}_h^T \mathbf{x}_i, \\
&\stackrel{\zeta_1}{\leq} \frac{1}{100} \|(\mathbf{u}_h \mathbf{u}_h^T - I) \mathbf{u}_*\|_2 \|\mathbf{u}_h^T\|_2 = \frac{1}{100} \sqrt{1 - (\mathbf{u}_h^T \mathbf{u}_*)^2},
\end{aligned} \tag{6}$$

where ζ_1 follows by observing that $(\mathbf{u}_h \mathbf{u}_h^T - I) \mathbf{u}_*$ and \mathbf{u}_h are orthogonal set of vectors and then using Property 3 given in the Theorem 1. Hence, using (5), (6), and $\|\mathbf{v}_*\|_2 = 1$ along with (4), we get:

$$\|G\|_2 \leq \frac{1}{99} \sqrt{1 - (\mathbf{u}_h^T \mathbf{u}_*)^2}. \tag{7}$$

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

Now, by left-multiplying (3) by \mathbf{v}_* and using (5) we obtain:

$$\mathbf{v}_*^T \hat{\mathbf{v}}_{h+1} = \sigma_* (\mathbf{u}_h^T \mathbf{u}_*) - \sigma_* \mathbf{v}_*^T G \geq \sigma_* (\mathbf{u}_h^T \mathbf{u}_*) - \frac{\sigma_*}{99} \sqrt{1 - (\mathbf{u}_h^T \mathbf{u}_*)^2}. \tag{8}$$

Similarly, by multiplying (3) by \mathbf{v}_*^\perp , where \mathbf{v}_*^\perp is a unit norm vector that is orthogonal to \mathbf{v}_* , we get:

$$\langle \mathbf{v}_*^\perp, \hat{\mathbf{v}}_{h+1} \rangle \leq \frac{\sigma_*}{99} \sqrt{1 - (\mathbf{u}_h^T \mathbf{u}_*)^2}. \tag{9}$$

Using (8), (9), and $\|\widehat{\mathbf{v}}_{h+1}\|_2^2 = (\mathbf{v}_*^T \widehat{\mathbf{v}}_{h+1})^2 + ((\mathbf{v}_*^\perp)^T \widehat{\mathbf{v}}_{h+1})^2$, we get:

$$\begin{aligned} 1 - (\mathbf{v}_{h+1}^T \mathbf{v}_*)^2 &= \frac{\langle \mathbf{v}_*^\perp, \widehat{\mathbf{v}}_{h+1} \rangle^2}{\langle \mathbf{v}_*, \widehat{\mathbf{v}}_{h+1} \rangle^2 + \langle \mathbf{v}_*^\perp, \widehat{\mathbf{v}}_{h+1} \rangle^2}, \\ &\leq \frac{1}{99 \cdot 99 \cdot (\mathbf{u}_h^T \mathbf{u}_* - \frac{1}{99} \sqrt{1 - (\mathbf{u}_h^T \mathbf{u}_*)^2})^2 + 1} (1 - (\mathbf{u}_h \mathbf{u}_*)^2). \end{aligned} \quad (10)$$

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 \mathbf{u}_0 on left and \mathbf{v}_0 on the right and using the fact that $(\mathbf{u}_0, \mathbf{v}_0)$ are the largest singular vectors of S , we get: $\|S\|_2 - \sigma_* \mathbf{v}_0^T \mathbf{v}_* \mathbf{u}_0^T \mathbf{u}_* \leq \sigma_*/100$. Hence, $\mathbf{u}_0^T \mathbf{u}_* \geq 9/10$.

Using the (10) along with the above given observation and by the ‘‘inductive’’ assumption $\mathbf{u}_h^T \mathbf{u}_* \geq \mathbf{u}_0^T \mathbf{u}_* \geq 9/10$ (proof of the inductive step follows directly from the below equation), we get:

$$1 - (\mathbf{v}_{h+1}^T \mathbf{v}_*)^2 \leq \frac{1}{2} (1 - (\mathbf{u}_h^T \mathbf{u}_*)^2). \quad (11)$$

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

Note that we require intermediate vectors $\mathbf{u}, \mathbf{v}, \mathbf{u}_\perp, \mathbf{v}_\perp$ to be independent of randomness in A_i ’s. Hence, we partition \mathcal{A}_{all} into $2H + 1$ partitions and at each step $(\mathcal{A}^h, \mathbf{b}^h)$ and $(\mathcal{A}^{h+1}, \mathbf{b}^{h+1})$ are supplied to the algorithm. This implies that the measurement complexity of the algorithm is given $m \cdot H = m \log(\|W_*\|_F/\epsilon)$. 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.

3 Rank-one Matrix Sensing using Gaussian Measurements

In this section, we study the problem of sensing general low-rank matrices which is an important problem in the domain of signal acquisition [1] and has several applications in a variety of areas like control theory, computer vision, etc. For this problem, the goal is to *design* the measurement matrix A_i as well as recovery algorithm, so that the true low-rank signal W_* can be recovered back from the given linear measurements.

Consider a measurement operator $\mathcal{A}_{Gauss} = \{A_1, A_2, \dots, A_m\}$ where each measurement matrix $A_i = \mathbf{x}_i \mathbf{y}_i^T$ is sampled using normal distribution, i.e., $\mathbf{x}_i \sim N(0, I)$, $\mathbf{y}_i \sim N(0, I), \forall i$. Now, for this operator \mathcal{A}_{Gauss} , we show that if $m = \Omega(k^4 \beta^2 \cdot (d_1 + d_2) \cdot \log^2(d_1 + d_2))$, 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_* .

Note that several similar recovery results for the matrix sensing problem already exist in the literature that guarantee exact recovery using $\Omega(k(d_1 + d_2) \log(d_1 + d_2))$ measurements [1, 10, 9]. However, we would like to stress that all the above mentioned existing results assume that the measurement operator \mathcal{A} satisfies the Restricted Isometry Property (RIP) defined below:

Definition 2. A linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^m$ satisfies RIP iff, $\forall W$ s.t. $\text{rank}(W) \leq k$, 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 .

Most current constructions of RIP matrices require each A_i to be sampled from a zero mean distribution with bounded fourth norm which implies that they have almost *full* rank. That is, such operators require $O(md_1d_2)$ memory just to store the operator, i.e., the storage requirement is cubic in $d_1 + d_2$. Consequently signal acquisition as well as recovery time for these algorithms is also at least cubic in $d_1 + d_2$. In contrast, our proposed rank-1 measurements require only $O(m(d_1 + d_2))$ storage and computational time. Hence, the proposed method makes the signal acquisition as well as signal recovery at least an order of magnitude faster.

Naturally, this begs the question whether we can show that our rank-1 measurement operator \mathcal{A}_{Gauss} satisfies RIP, so that the existing analysis for RIP based low-rank matrix sensing can be used [8]. We answer this question in the negative, i.e., for $m = O((d_1 + d_2) \log(d_1 + d_2))$, \mathcal{A}_{Gauss} does not satisfy RIP even for rank-1 matrices (with high probability):

Claim 3. *Let $\mathcal{A}_{Gauss} = \{A_1, A_2, \dots, A_m\}$ be a measurement operator with each $A_i = \mathbf{x}_i \mathbf{y}_i^T$, where $\mathbf{x}_i \in \mathbb{R}^{d_1} \sim \mathcal{N}(0, I)$, $\mathbf{y}_i \in \mathbb{R}^{d_2} \sim \mathcal{N}(0, I)$, $1 \leq i \leq m$. 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}_{Gauss} does not satisfy RIP for rank-1 matrices with a constant δ .*

Proof of Claim 3. The main idea behind our proof is to show that there exists two rank-1 matrices Z_U, Z_L s.t. $\|\mathcal{A}_{Gauss}(Z_U)\|_2^2$ is large while $\|\mathcal{A}_{Gauss}(Z_L)\|_2^2$ is much smaller than $\|\mathcal{A}_{Gauss}(Z_U)\|_2^2$.

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

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

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

$$\|\mathcal{A}_{Gauss}(Z_U)\|_2^2 \geq .5 d_1^2 d_2^2. \quad (12)$$

Moreover, $\|Z_U\|_F^2 \leq 2d_1d_2$ w.p. $\geq 1 - 2 \exp(-d_1 - d_2)$.

Now, consider

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

where $Z_L = \mathbf{u} \mathbf{v}^T$ and \mathbf{u}, \mathbf{v} are sampled from standard normal distribution, independent of $\mathbf{x}_i, \mathbf{y}_i, \forall i$. Since, \mathbf{u}, \mathbf{v} are independent of $\mathbf{u}^T \mathbf{x}_i \sim N(0, \|\mathbf{u}\|_2)$ and $\mathbf{v}^T \mathbf{y}_i \sim N(0, \|\mathbf{v}\|_2)$. Hence, w.p. $\geq 1 - 1/m^3$, $|\mathbf{u}^T \mathbf{x}_i| \leq \log(m) \|\mathbf{u}\|_2, |\mathbf{v}^T \mathbf{y}_i| \leq \log(m) \|\mathbf{v}\|_2, \forall i \geq 2$. Moreover, w.p. $\geq 1 - \exp(-d_1 - d_2)$, $\|\mathbf{u}\|_2 \leq 2\sqrt{d_1}$ and $\|\mathbf{v}\|_2 \leq 2\sqrt{d_2}$. That is, w.p. $1 - 1/m^3$:

$$\|\mathcal{A}_{Gauss}(Z_L)\|_2^2 \leq 4m \cdot d_1 \cdot d_2 \log^4 m. \quad (13)$$

Furthermore, $\|Z_L\|_F^2 \leq 2d_1d_2$ w.p. $\geq 1 - 2 \exp(-d_1 - d_2)$.

Using (12), (13), we get that w.p. $\geq 1 - 2/m^3 - 10 \exp(-d_1 - d_2)$:

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

Now, for RIP to be satisfied with a constant δ , the lower and upper bound on $\|\mathcal{A}_{Gauss}(Z/\|Z\|_F)\|^2$ for all rank-1 Z should be at most a constant factor apart. However, the above equation clearly shows that the upper and lower bound can match only when $m = \Omega(d_1 d_2 / \log(5d_1 d_2))$. Hence, for m that is at most linear in both d_1, d_2 , RIP cannot be satisfied with probability $\geq 1 - 1/(d_1 + d_2)^3$. \square

Now, even though \mathcal{A}_{Gauss} does not satisfy RIP, we can still show that \mathcal{A}_{Gauss} satisfies the three properties mentioned in the Theorem 1. and hence we can use Theorem 1 to obtain the exact recovery result.

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

Proof of Lemma 4. We divide the proof into three parts where each part proves a property mentioned in Theorem 1.

Proof of Property 1. Now,

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

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

As S is a sum of m random matrices, the goal is to apply matrix concentration bounds to show that S is close to $\mathbb{E}[S] = W = U_* \Sigma_* V_*^T$ for large enough m . To this end, we use Theorem 8 by [14] given below. However, Theorem 8 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:

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, & |x_{ij}| \leq C \sqrt{\log(m(d_1 + d_2))}, \\ 0, & \text{otherwise,} \end{cases} \quad (14)$$

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

$$\tilde{y}_{ij} = \begin{cases} y_{ij}, & |y_{ij}| \leq C \sqrt{\log(m(d_1 + d_2))}, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

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

$$\begin{aligned} \mathbb{E}[\tilde{x}_{ij}^2] &= \mathbb{E}[x_{ij}^2] - \frac{2}{\sqrt{2\pi}} \int_{C\sqrt{\log(m(d_1 + d_2))}}^{\infty} x^2 \exp(-x^2/2) dx, \\ &= 1 - \frac{2}{\sqrt{2\pi}} \frac{C\sqrt{\log(m(d_1 + d_2))}}{(m(d_1 + d_2))^{C^2/2}} - \frac{2}{\sqrt{2\pi}} \int_{C\sqrt{\log(m(d_1 + d_2))}}^{\infty} \exp(-x^2/2) dx, \\ &\geq 1 - \frac{2C\sqrt{\log(m(d_1 + d_2))}}{(m(d_1 + d_2))^{C^2/2}}. \end{aligned} \quad (16)$$

Similarly,

$$\mathbb{E}[\tilde{y}_{ij}^2] \geq 1 - \frac{2C\sqrt{\log(m(d_1 + d_2))}}{(m(d_1 + d_2))^{C^2/2}}. \quad (17)$$

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

$$\begin{aligned} \|\mathbb{E}[\tilde{Z}_i \tilde{Z}_i^T]\|_2 &= \|\mathbb{E}[\|\tilde{\mathbf{y}}_i\|_2^2 \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T U_* \Sigma_* V_*^T \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T V_* \Sigma_* U_*^T \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T]\|_2, \\ &\leq C^2 d_2 \log(m(d_1 + d_2)) \mathbb{E}[\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T U_* \Sigma_*^2 U_*^T \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T]\|_2, \\ &\leq C^2 d_2 \log(m(d_1 + d_2)) (\sigma_*^1)^2 \|\mathbb{E}[\|U_*^T \tilde{\mathbf{x}}_i\|_2^2 \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T]\|_2, \\ &\leq C^4 k d_2 \log^2(m(d_1 + d_2)) (\sigma_*^1)^2. \end{aligned} \quad (18)$$

Similarly,

$$\|\mathbb{E}[\tilde{Z}_i] \mathbb{E}[\tilde{Z}_i^T]\|_2 \leq (\sigma_*^{max})^2. \quad (19)$$

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

Finally, by selecting $m = \frac{C_1 k (d_1 + d_2) \log^2(d_1 + d_2)}{\delta^2}$ and applying Theorem 8 we get (w.p. $1 - \frac{1}{(d_1 + d_2)^{10}}$),

$$\left\| \frac{1}{m} \sum_{i=1}^m \tilde{Z}_i - \mathbb{E}[\tilde{Z}_i] \right\|_2 \leq \delta. \quad (20)$$

Note that $\mathbb{E}[\tilde{Z}_i] = \mathbb{E}[\tilde{x}_{i1}^2] \mathbb{E}[\tilde{y}_{i1}^2] U_* \Sigma_* V_*^T$. Hence, by using (20), (16), (17),

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

Finally, by observing that by selecting C to be large enough in the definition of $\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i$ (see (14), (15)), we get $P(\|Z_i - \tilde{Z}_i\|_2 = 0) \geq 1 - \frac{1}{(d_1 + d_2)^5}$. Hence, by assuming δ to be a constant wrt d_1, d_2 and by union bound, w.p. $1 - \frac{2\delta^{10}}{(d_1 + d_2)^5}$,

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

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

□

Global optimality of the rate of convergence of the Alternating Minimization procedure for this problem now follows directly by using Theorem 1 with the above given lemma. We would like to note that while the above result shows that the \mathcal{A}_{Gauss} 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}_{Gauss} 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 Inductive Matrix Completion

In this section, we study the problem of inductive matrix completion which is another important application of the LRROM problem. Consider a movie recommender system which contains n_1 users and n_2 movies and let $R \in \mathbb{R}^{n_1 \times n_2}$ be the corresponding “true” ratings matrix. The standard matrix completion methods only utilize the samples from the ratings matrix R and ignore the side-information that might be present in the system such as, demographic information of the user or genre of the movie. This restricts the usage of matrix completion to the transductive setting only.

Recently, [11] studied a generalization of the low-rank matrix completion problem where R_{ij} is modeled as $R_{ij} = \mathbf{x}_i^T W_* \mathbf{y}_j$; where $\mathbf{x}_i, \mathbf{y}_j$ are the feature vectors of users and movies, respectively. Using benchmark datasets, they showed empirically that their method outperforms traditional matrix completion methods. However, to the best of our knowledge, there is no existing theoretical analysis of such an inductive approach.

Now, since R is a rank- k matrix, one can still apply standard matrix completion results to recover R and hence W_* . Assuming that the observed index set Ω is sampled uniformly from $[n_1] \times [n_2]$ and that R is incoherent, a direct application of the matrix completion methods would require $|\Omega| \geq C(k(n_1 + n_2) \log(n_1 + n_2))$ samples to be known. Now, if $d_1 + d_2 \ll n_1 + n_2$ then this means that many more samples are required than the total degrees of freedom in W_* which is $O(k(d_1 + d_2))$.

Hence, a natural question here is can the above given sample complexity bound be improved? Below, we provide the answer to this question in affirmative. In particular, we show that by using the feature vectors AltMin-LRROM (see Algorithm 1) can recover the true matrix W_* using $O(kd_1d_2 \log(d_1d_2))$ random samples. Now, if $d_1d_2 \ll n_1 + n_2$, then our method requires significantly lesser number of samples than the standard matrix completion methods. Furthermore, this implies that several users/movies need not have even *one* known rating, i.e, the method can be applied to the inductive setting as well. We note that our sample size requirement is still larger than the information theoretically optimal requirement which is $O(k(d_1 + d_2) \log(d_1 + d_2))$. We leave further reduction in the sample complexity as an open problem.

Similar to the previous section, we utilize our general theorem for optimality of the LRROM problem to provide a convergence analysis of the inductive matrix completion method. In particular, we provide the following lemma which shows that assuming X, Y to be incoherent (see Definition 5), the above mentioned inductive matrix completion operator also satisfies Properties 1, 2, 3 required by Theorem 1. Hence, AltMin-LRROM (Algorithm 1) converges to the global optimum in $O(\log(\|W_*\|_F/\epsilon))$ iterations. We first provide the definition of incoherent matrices.

Definition 5. $X \in \mathbb{R}^{d \times n}$ ($d < n$) is μ -incoherent if: $\|U_X^i\|_2 \leq \frac{\mu\sqrt{d}}{\sqrt{n}}, 1 \leq i \leq d$, where $X^T = U_X \Sigma_X V_X^T$ is the SVD of X^T and $U_X^i \in \mathbb{R}^d$ is the i -th row of $U_X \in \mathbb{R}^{n \times d}$.

Lemma 6. Let both $X \in \mathbb{R}^{d_1 \times n_1}$ and $Y \in \mathbb{R}^{d_2 \times n_2}$ be μ -incoherent matrices. Let $R = X^T W_* Y$ be the “ratings” matrix and let $W_* \in \mathbb{R}^{d_1 \times d_2}$ be any fixed rank- k matrix. Let Ω be a uniformly random subset of $[n_1] \times [n_2]$, s.t., $|\Omega| = m \geq Ck^3 \cdot \beta^2 \cdot d_1d_2 \cdot \log(d_1 + d_2)$, where $\beta = \sigma_R^1/\sigma_R^k$ is the condition number of R . Then, w.p. $\geq 1 - 1/(d_1 + d_2)^{100}$, the measurement operators $A_{ij} = \sqrt{n_1n_2} \mathbf{x}_i \mathbf{y}_j^T$ satisfy¹ Properties 1,2,3 required by Theorem 1.

Proof. We first observe that both X, Y can be thought of as orthonormal matrices. The reason being, $X^T W_* Y = U_X \Sigma_X V_X^T W_* V_Y \Sigma_Y U_Y^T$, where $X^T = U_X \Sigma_X V_X^T$ and $Y^T = U_Y \Sigma_Y V_Y^T$. Hence,

¹We multiply $\mathbf{x}_i, \mathbf{y}_j$ by $\sqrt{n_1}, \sqrt{n_2}$ for normalization so that $\mathbb{E}_i[n_1 \mathbf{x}_i \mathbf{x}_i^T] = I$ and $\mathbb{E}_j[n_2 \mathbf{y}_j \mathbf{y}_j^T] = I$.

$R = X^T W_* Y = U_X (\Sigma_X V_X^T W_* V_Y \Sigma_Y) U_Y^T$. That is, 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. Hence, we prove the theorem with the assumption that X, Y are orthonormal and that β is the condition number of W_* .

We now present the proof for each of the three properties mentioned in Theorem 1.

Proof of Property 1. As mentioned above, 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 :

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

where $Z_{ij} = \mathbf{x}_i \mathbf{x}_i^T U_* \Sigma_* V_*^T \mathbf{y}_j \mathbf{y}_j^T = X \mathbf{e}_i \mathbf{e}_i^T X^T U_* \Sigma_* V_*^T Y \mathbf{e}_j \mathbf{e}_j^T Y^T$, where $\mathbf{e}_i, \mathbf{e}_j$ 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, $\mathbb{E}_i[\mathbf{e}_i \mathbf{e}_i^T] = \frac{1}{n_1} I$ and $\mathbb{E}_j[\mathbf{e}_j \mathbf{e}_j^T] = \frac{1}{n_2} I$. 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 8 to bound $\|S - W_*\|_2$. To apply the bound of Theorem 8, we first need to bound the following two quantities:

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

$$\|Z_{ij}\|_2 = \|\mathbf{x}_i \mathbf{x}_i^T U_* \Sigma_* V_*^T \mathbf{y}_j \mathbf{y}_j^T\|_2 \leq \sigma_*^1 \|\mathbf{x}_i\|_2^2 \|\mathbf{y}_j\|_2^2 \leq \frac{\sigma_*^1 \mu^4 d_1 d_2}{n_1 n_2},$$

where the last inequality follows using incoherence 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 $\|\sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T]\|_2$:

$$\begin{aligned} \left\| \sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T W_* \mathbf{y}_j \mathbf{y}_j^T \mathbf{y}_j \mathbf{y}_j^T W_*^T \mathbf{x}_i \mathbf{x}_i^T] \right\|_2, \\ &\stackrel{\zeta_1}{\leq} \frac{\mu^2 d_2}{n_2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T W_* \mathbf{y}_j \mathbf{y}_j^T W_*^T \mathbf{x}_i \mathbf{x}_i^T] \right\|_2 \stackrel{\zeta_2}{\leq} \frac{\mu^2 d_2}{n_2^2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T W_* W_*^T \mathbf{x}_i \mathbf{x}_i^T] \right\|_2, \\ &\stackrel{\zeta_3}{\leq} \frac{(\sigma_*^1)^2 \mu^4 d_1 d_2}{n_1 n_2^2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] \right\|_2 \stackrel{\zeta_4}{\leq} \frac{(\sigma_*^1)^2 \mu^4 d_1 d_2}{n_1^2 n_2^2} \cdot m, \end{aligned} \tag{21}$$

where ζ_1, ζ_3 follows by using incoherent of X, Y and $\|W_*\|_2 \leq \sigma_*^1$. ζ_2, ζ_4 follows by using $\mathbb{E}_i[\mathbf{e}_i \mathbf{e}_i^T] = \frac{1}{n_1} I$ and $\mathbb{E}_j[\mathbf{e}_j \mathbf{e}_j^T] = \frac{1}{n_2} I$.

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 8 and using the above computed bounds we get:

$$Pr(\|S - W_*\|_2 \geq \sigma_*^1 \gamma) \leq 2(d_1 + d_2) \exp\left(-\frac{m\gamma^2}{\mu^4 d_1 d_2 (1 + \gamma/3)}\right). \quad (22)$$

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

$$\|S - W_*\|_2 \leq \frac{\sigma_*^1 \mu^2 \sqrt{d_1 d_2 \log(2(d_1 + d_2)/\gamma)}}{\sqrt{m}}. \quad (23)$$

Hence, by selecting $m = \Omega(\mu^4 k^3 \cdot \beta^2 \cdot d_1 d_2 \log(2(d_1 + d_2)/\gamma))$ where $\beta = \sigma_*^1 / \sigma_*^k$, the following holds w.p. $\geq 1 - \gamma$:

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

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

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

$$\mathbb{E}[B_y] = \frac{n_1 n_2}{m} \sum_{(i,j) \in \Omega} Z_{ij} = \frac{n_1 n_2}{m} \sum_{i=1}^m \mathbb{E}_{(i,j) \in \Omega} [\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i \mathbf{y}_i \mathbf{y}_i^T] = I. \quad (24)$$

Here again, we apply Theorem 8 to bound $\|B_y - I\|_2$. To this end, we need to bound the following quantities:

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

$$\|Z_{ij}\|_2 = \|\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i \mathbf{y}_i \mathbf{y}_i^T\|_2 \leq \|\mathbf{y}_i\|_2^2 \|\mathbf{x}_i\|_2^2 \leq \frac{\mu^4 d_1 d_2}{n_1 n_2}.$$

- **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 $\|\sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T]\|_2$:

$$\begin{aligned} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[Z_{ij} Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i)^2 \|\mathbf{y}_i\|_2^2 \mathbf{y}_i \mathbf{y}_i^T] \right\|_2 \stackrel{\zeta_1}{\leq} \frac{\mu^2 d_2}{n_2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i)^2 \mathbf{y}_i \mathbf{y}_i^T] \right\|_2, \\ &\stackrel{\zeta_2}{\leq} \frac{\mu^2 d_2}{n_2^2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i)^2] \right\|_2 \stackrel{\zeta_3}{\leq} \frac{\mu^4 d_1 d_2}{n_1 n_2^2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u})^2] \right\|_2 \stackrel{\zeta_4}{\leq} \frac{\mu^4 d_1 d_2}{n_1^2 n_2^2} \cdot m. \end{aligned} \quad (25)$$

Note that the above given bounds that we obtain are exactly the same as the ones obtained in the Initialization Property's proof. Hence, by applying Theorem 8 in a similar manner, and selecting $m = \Omega(k^3\beta^2d_1 \cdot d_2 \log(1/\gamma))$ 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. \square

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 $m = \Omega(k^3\beta^2d_1 \cdot d_2 \log(1/\gamma))$ 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. \square

\square

5 Multi-label Learning

In this section, we study the problem of multi-label regression with missing values. Let $X = [\mathbf{x}_1 \dots \mathbf{x}_{n_1}] \in \mathbb{R}^{d_1 \times n_1}$ be the training matrix where \mathbf{x}_i is the feature vector of the i -th data point. Also, let $R \in \mathbb{R}^{n_1 \times L}$ be the corresponding matrix of target variables. That is, $R^i = [R_{i1} \dots R_{ij} \dots R_{iL}]$ denotes L target variables for \mathbf{x}_i . The goal is to learn a (low-rank) parameter matrix W_* s.t. $X^T W_* = R$.

The above problem is a straightforward multi-variate linear regression problem. However, in several large-scale multi-label learning problems, it is impossible to obtain all the target variables for each of the points. That is, R generally has several entries missing. The goal is to learn W_* exactly, even when only a small number of random entries of R is available.

Here again, we view the problem as a low-rank matrix estimation problem with rank-one measurements $R_{ij} = \mathbf{e}_i^T X^T W_* \mathbf{e}_j$, $(i, j) \in \Omega$, where index Ω is a uniformly sampled subset of $[n_1] \times [L]$. Note that this problem is a combination of the inductive matrix completion problem we studied in the previous section and the standard matrix completion. The left hand side measurement vector $X \mathbf{e}_i$ is similar to inductive matrix completion while the right hand measurement vector \mathbf{e}_j is a standard matrix completion type of measurement vector. That is, this problem assumes the labels to be "fixed" but is inductive w.r.t. the data points \mathbf{x} .

Similar to the previous section, we show that under a certain incoherence assumption on the feature matrix X , Properties 1, 2, 3, required by Theorem 1 are satisfied and hence alternating minimization will be able to learn the global optima W_* .

Lemma 7. *Let $X \in \mathbb{R}^{d_1 \times n_1}$ be μ -incoherent. Let $R = X^T W_* \in \mathbb{R}^{n_1 \times L}$ be the "labels" matrix. Let Ω be a uniformly random subset of $[n_1] \times [L]$, s.t., $|\Omega| = m \geq C\beta^2 \cdot d_1 n_2 \cdot \log(d_1 + n_2)$, where $\beta = \sigma_R^1 / \sigma_R^k$ is the condition number of R . Then, w.p. $\geq 1 - 1/(d_1 + L)^{100}$, the measurement operators $A_{ij} = \sqrt{n_1 n_2} \mathbf{x}_i \mathbf{e}_j^T$ satisfy² Properties 1, 2, 3 required by Theorem 1.*

²We multiply $\mathbf{x}_i, \mathbf{y}_j$ by $\sqrt{n_1}, \sqrt{n_2}$ for normalization so that $\mathbb{E}_i[n_1 \mathbf{x}_i \mathbf{x}_i^T] = I$ and $\mathbb{E}_j[n_2 \mathbf{y}_j \mathbf{y}_j^T] = I$

Assuming β, k to be constant and by ignoring log factors, the above lemma shows that using $m = d_1 \cdot L$ samples the parameter matrix W_* can be recovered exactly. In contrast, matrix completion requires $m = n_1 + L$ samples. That is, if the number of training points is significantly larger than $d_1 \cdot L$, then the above method improves upon the matrix completion approach significantly. This result can be interpreted in another way: for missing labels a standard method is to first do matrix completion and then learn W_* . Our above lemma gives an example of a setting where simultaneous learning and completion of R leads to significantly better sample complexity.

We now provide a proof of the above lemma.

Proof. Here again, we divide the proof into three parts where each part proves a property mentioned in Theorem 1.

Proof of Property 1. As mentioned in the proof of Lemma 6, 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 :

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

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

Now using the fact that (i, j) is sampled uniformly at random from $[n_1] \times [n_2]$:

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

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

As in the previous section, we first bound the following two quantities:

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

$$\|Z_{ij}\|_2 = \|\mathbf{x}_i \mathbf{x}_i^T U_* \Sigma_* V_*^T \mathbf{e}_j \mathbf{e}_j^T\|_2 \leq \sigma_*^1 \|\mathbf{x}_i\|_2^2 \leq \frac{\sigma_*^1 \mu^2 d_1}{n_1},$$

where the last inequality follows using incoherence of X and V_* .

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

$$\begin{aligned} \left\| \sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T W_* \mathbf{e}_j \mathbf{e}_j^T \mathbf{e}_j \mathbf{e}_j^T W_*^T \mathbf{x}_i \mathbf{x}_i^T] \right\|_2, \\ &\stackrel{\zeta_1}{\leq} \frac{1}{n_2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T W_* W_*^T \mathbf{x}_i \mathbf{x}_i^T] \right\|_2 \stackrel{\zeta_2}{\leq} \frac{(\sigma_*^1)^2 \mu^2 d_1}{n_1 n_2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] \right\|_2 \stackrel{\zeta_3}{\leq} \frac{(\sigma_*^1)^2 \mu^2 d_1}{n_1^2 n_2} \cdot m, \quad (26) \end{aligned}$$

where ζ_1 follows from $\mathbb{E}_j[\mathbf{e}_j \mathbf{e}_j^T] = \frac{1}{n_2} I$, ζ_2 follows from incoherence of \mathbf{x}_i , and ζ_3 follows from $\mathbb{E}_i[\mathbf{x}_i \mathbf{x}_i^T] = \frac{1}{n_1} I$.

- Bound $\|\sum_{(i,j) \in \Omega} E[Z_{ij}^T Z_{ij}]\|_2$:

$$\begin{aligned}
& \left\| \sum_{(i,j) \in \Omega} E[Z_{ij}^T Z_{ij}] \right\|_2 = \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{e}_j \mathbf{e}_j^T W_*^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{x}_i \mathbf{x}_i^T W_* \mathbf{e}_j \mathbf{e}_j^T] \right\|_2, \\
& \leq \frac{\mu^2 d_1}{n_1} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{e}_j \mathbf{e}_j^T W_*^T \mathbf{x}_i \mathbf{x}_i^T W_* \mathbf{e}_j \mathbf{e}_j^T] \right\|_2 \stackrel{\zeta_2}{=} \frac{\mu^2 d_1}{n_1^2} \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[\mathbf{e}_j \mathbf{e}_j^T W_*^T W_* \mathbf{e}_j \mathbf{e}_j^T] \right\|_2, \\
& \leq \frac{(\sigma_*^1)^2 \mu^2 d_1}{n_1^2 n_2} \cdot m, \tag{27}
\end{aligned}$$

where ζ_1 follows from incoherence of X , ζ_2, ζ_3 follows from uniform sampling of \mathbf{e}_i and \mathbf{e}_j , respectively.

Using (26), (27) we get:

$$\max \left(\left\| \sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T] \right\|_2, \left\| \sum_{(i,j) \in \Omega} E[Z_{ij}^T Z_{ij}] \right\|_2 \right) \leq \frac{(\sigma_*^1)^2 \mu^2 d_1}{n_1^2 n_2} \cdot m.$$

Using the above given bounds, and Theorem 8, we get:

$$Pr(\|S - W_*\|_2 \geq \frac{n_2 \sigma_*^1 \gamma}{\mu \sqrt{k}}) \leq 2(d_1 + n_2) \exp \left(-\frac{m \gamma^2}{\mu^4 \cdot k \cdot d_1 (1 + \gamma/3)} \right). \tag{28}$$

That is, by selecting $m = \Omega(k^3 \beta^2 \mu^2 d_1 n_2 \log(2(d_1 + n_2)/\gamma))$ with $\beta = \frac{\sigma_*^1}{\sigma_*^k}$, the following holds w.p. $\geq 1 - \gamma$:

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

where $\delta \leq \frac{1}{k^{3/2} \cdot \beta \cdot 100}$. □

Proof of Property 2. Here, we first prove the property for B_y . Now, $B_y = \frac{n_1 n_2}{m} \sum_{(i,j) \in \Omega} Z_{ij}$ where $Z_i = \mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i \mathbf{e}_j \mathbf{e}_j^T$. Note that, $\mathbb{E}[B_y] = I$.

Next, we bound the quantities required by Theorem 8:

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

$$\|Z_{ij}\|_2 = \|\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i \mathbf{e}_j \mathbf{e}_j^T\|_2 \leq \|\mathbf{x}_i\|_2^2 \leq \frac{\mu^2 d_1}{n_1},$$

where the second inequality follows from incoherence of X .

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

$$\left\| \sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T] \right\|_2 = \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i)^2 \mathbf{e}_j \mathbf{e}_j^T] \right\|_2 \stackrel{\zeta_1}{=} \frac{1}{n_2} \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{x}_i^T \mathbf{u} \mathbf{u}^T \mathbf{x}_i)^2] \stackrel{\zeta_2}{\leq} \frac{\mu^2 d_1}{n_1^2 n_2},$$

where ζ_1 follows as \mathbf{e}_j is sampled uniformly and ζ_2 follows by using incoherence of X and uniform sampling of \mathbf{e}_i .

Hence, using $m = \Omega(k^3 \cdot \beta^2 \cdot d \cdot n_2 \log(2(d_1 + n_2)/\gamma))$, then we have (w.p. $\geq 1 - \gamma$):

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

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

Now, we bound $B_x = \frac{n_1 n_2}{m} \sum_{(i,j) \in \Omega} Z_{ij}$ where $Z_i = \mathbf{e}_j^T \mathbf{v} \mathbf{v}^T \mathbf{e}_j \mathbf{x}_i \mathbf{x}_i^T$. Note that, $\mathbb{E}[B_y] = I$. Next, we bound the quantities required by Theorem 8:

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

$$\|Z_{ij}\|_2 = \|\mathbf{e}_j^T \mathbf{v} \mathbf{v}^T \mathbf{e}_j \mathbf{x}_i \mathbf{x}_i^T\|_2 \leq \|\mathbf{x}_i\|_2^2 \leq \frac{\mu^2 d_1}{n_1},$$

where the second inequality follows from incoherence of X .

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

$$\begin{aligned} \left\| \sum_{(i,j) \in \Omega} E[Z_{ij} Z_{ij}^T] \right\|_2 &= \left\| \sum_{(i,j) \in \Omega} \mathbb{E}[(\mathbf{e}_j^T \mathbf{v} \mathbf{v}^T \mathbf{e}_j)^2 \|\mathbf{x}_i\|^2 \mathbf{x}_i \mathbf{x}_i^T] \right\|_2 \\ &\stackrel{\zeta_1}{\leq} \frac{1}{n_2} \sum_{(i,j) \in \Omega} \mathbb{E}[\|\mathbf{x}_i\|^2 \mathbf{x}_i \mathbf{x}_i^T], \\ &\stackrel{\zeta_2}{\leq} \frac{\mu^2 d_1}{n_1^2 n_2}, \end{aligned} \tag{29}$$

where ζ_1 follows as \mathbf{e}_j is sampled uniformly and ζ_2 follows by using incoherence of X and uniform sampling of \mathbf{e}_i .

Hence, using $m = \Omega(k^3 \cdot \beta^2 \cdot d \cdot n_2 \log(2(d_1 + n_2)/\gamma))$, we have (w.p. $\geq 1 - \gamma$):

$$\|B_x - I\|_2 \leq \delta,$$

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

□

Proof of Property 3. We first note that $\mathbb{E}[C_x] = \mathbb{E}[C_y] = 0$. Now, here again we use Theorem 8 to say that C_x, C_y converge to their mean. The quantities we need to bound are similar to the ones proved above for Property 2. Hence, the Property 3 follows using $m = \Omega(k^3 \cdot \beta^2 \cdot d \cdot n_2 \log(2(d_1 + n_2)/\gamma))$ samples. □

□

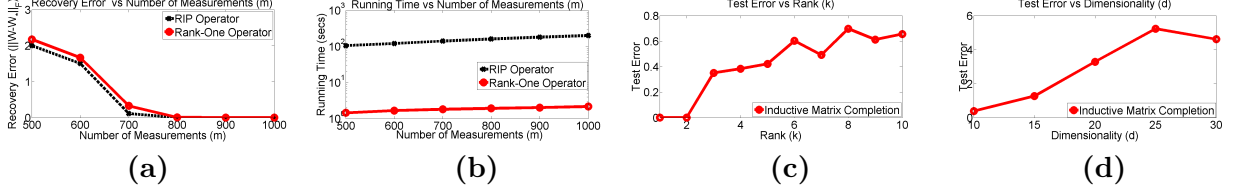


Figure 1: (a), (b): Low-rank Matrix Sensing—Comparison of RIP based and the rank-one matrices based measurement operators for low-rank matrix sensing. Clearly, our rank-one operator is significantly faster than the RIP based method while incurring similar recovery error. (c), (d): Inductive Matrix Completion—plots show the error incurred by alternating minimization on the test data with, (c): varying rank of the underlying W_* , and (d): varying dimensionality of W_* .

6 Experiments

In this section, we first demonstrate empirically that our Gaussian rank-one linear operator (\mathcal{A}_{Gauss}) is significantly more efficient for matrix sensing than the existing RIP based measurement operators. To this end, we first generated a random rank-5 signal $W_* \in \mathbb{R}^{50 \times 50}$ and then generate different number of measurements using both \mathcal{A}_{Gauss} and an RIP based operator. We run alternating minimization method for both type of measurements. Figure 1 (a) compares the Frobenius norm in recovery by both the methods. Figure 1 (b) plots (on log-scale) the running time of both the methods as m increases. Clearly, the \mathcal{A}_{Gauss} operator based measurements provide reasonably accurate recovery while the running time of our \mathcal{A}_{Gauss} based method is about two orders of magnitude better than that of RIP based measurement method.

Next, we demonstrate that by using a very small number of measurements, the multi-label regression problem can still be solved accurately. For this, we selected number of labels $L = 50$, number of points $n_1 = 100$, and varied d from 1 to 20. We then generated 100 training points $X \in \mathbb{R}^{d_1 \times 100}$ and 100 test points. We then generated $W_* \in \mathbb{R}^{d_1 \times L}$ and observed only 200 random entries of $R = X^T W_*$. Figure 1 (c), (d) plot the error incurred in prediction over the test set, as k and d vary respectively. The error is computed using $\sum_{\mathbf{x} \in TestSet} |R_{xj} - \mathbf{x}^T W_* \mathbf{e}_j|^2$. Clearly, the method is able to output fairly accurate predictions for small k, d . Moreover, the test error degrades gracefully as either k or d increases.

References

- [1] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [2] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, December 2009.
- [3] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.
- [4] David Gross. Recovering low-rank matrices from few coefficients in any basis, 2009.
- [5] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.
- [6] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.
- [7] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Sparse and low-rank matrix decompositions. In *IFAC Symposium on System Identification*, 2009.
- [8] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.
- [9] Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, pages 937–945, 2010.
- [10] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. *arXiv preprint arXiv:0903.4742*, 2009.
- [11] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.
- [12] Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013.
- [13] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. In *ICML*, pages 1129–1136, 2011.
- [14] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- [15] Ren-Cang Li. On perturbations of matrix pencils with real spectra. *Math. Comp.*, 62:231–265, 1994.

A Preliminaries

Theorem 8 (Theorem 1.6 of [14]). *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 $\gamma \geq 0$,*

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

B Proof of General Theorem for Low-rank Matrix Estimation

Here, we now generalize our above given proof to the rank- k case. In the case of rank-1 matrix recovery, we used $1 - (\mathbf{v}_{h+1}^T \mathbf{u}_*)^2$ 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 9. *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 of Theorem 1: General Rank- k Case. For simplicity of notation, we denote U_h by U , \hat{V}_{h+1} by \hat{V} , and V_{h+1} by V .

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

$$\hat{V} = W_*^T U - F, \quad (30)$$

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:

$$B_{pq} = \sum_i \mathbf{y}_i \mathbf{y}_i^T (\mathbf{x}_i^T \mathbf{u}_p) (\mathbf{x}_i^T \mathbf{u}_q), \quad (31)$$

$$C_{pq} = \sum_i \mathbf{y}_i \mathbf{y}_i^T (\mathbf{x}_i^T \mathbf{u}_p) (\mathbf{x}_i^T \mathbf{u}_{*q}), \quad (32)$$

$$D_{pq} = \mathbf{u}_p^T \mathbf{u}_{*q} I, \quad (33)$$

$$S_{pq} = \sigma_*^p I \quad \text{if } p = q, \quad \text{and} \quad 0 \quad \text{if } p \neq q. \quad (34)$$

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

Then, using the definitions given above, we get:

$$\begin{bmatrix} F_1 \\ \vdots \\ F_k \end{bmatrix} = B^{-1} (BD - C) S \cdot \text{vec}(V_*). \quad (35)$$

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

Now, using (30), $V = \hat{V}R^{-1} = (W_*^T U - F)R^{-1}$. Multiplying both the sides by V_*^\perp , where V_*^\perp is a fixed orthonormal basis of the subspace orthogonal to $\text{span}(V_*)$, we get:

$$(V_*^\perp)^T V = -(V_*^\perp)^T F R^{-1} \Rightarrow \text{dist}(V_*, V_{t+1}) = \|(V_*^\perp)^T V\|_2 \leq \|F\|_2 \|R^{-1}\|_2. \quad (36)$$

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[15], we get: $\text{dist}(U_0, U_*) \leq \frac{1}{100}$.

Theorem now follows by using Lemma 10, Lemma 11 along with the above mentioned bound on $\text{dist}(U_0, U_*)$. □

Lemma 10. *Let \mathcal{A} be a rank-one measurement operator where $A_i = \mathbf{x}_i \mathbf{y}_i^T$. Also, let \mathcal{A} satisfy Property 1, 2, 3 mentioned in Theorem 1 and let $\sigma_*^1 \geq \sigma_*^2 \geq \dots \geq \sigma_*^k$ be the singular values of W_* . Then,*

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

Lemma 11. *Let \mathcal{A} be a rank-one measurement operator where $A_i = \mathbf{x}_i \mathbf{y}_i^T$. Also, let \mathcal{A} satisfy Property 1, 2, 3 mentioned in Theorem 1. Then,*

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

Proof of Lemma 10. Recall that $\text{vec}(F) = B^{-1}(BD - C)S \cdot \text{vec}(V_*)$. Hence,

$$\|F\|_2 \leq \|F\|_F \leq \|B^{-1}\|_2 \|BD - C\|_2 \|S\|_2 \|\text{vec}(V_*)\|_2 = \sigma_*^1 \sqrt{k} \|B^{-1}\|_2 \|BD - C\|_2. \quad (37)$$

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

$$\begin{aligned} \sigma_{\min}(B) &= \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \mathbf{z}^T B \mathbf{z} = \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \sum_{1 \leq p \leq k, 1 \leq q \leq k} \mathbf{z}_p^T B_{pq} \mathbf{z}_q \\ &= \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \sum_p \mathbf{z}_p^T B_{pp} \mathbf{z}_p + \sum_{pq, p \neq q} \mathbf{z}_p^T B_{pq} \mathbf{z}_q. \end{aligned} \quad (38)$$

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

$$\sigma_{\min}(B_{pp}) \geq 1 - \delta, \quad (39)$$

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:

$$\|B_{pq}\|_2 \leq \delta. \quad (40)$$

Hence, using (38), (39), (40), we get:

$$\sigma_{\min}(B) \geq \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} (1 - \delta) \sum_p \|\mathbf{z}_p\|_2^2 - \delta \sum_{pq, p \neq q} \|\mathbf{z}_p\|_2 \|\mathbf{z}_q\|_2 = \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} 1 - \delta \sum_{pq} \|\mathbf{z}_p\|_2 \|\mathbf{z}_q\|_2 \geq 1 - k\delta. \quad (41)$$

Now, consider $BD - C$:

$$\begin{aligned} \|BD - C\|_2 &= \max_{\mathbf{z}, \|\mathbf{z}\|_2=1} |\mathbf{z}^T (BD - C) \mathbf{z}|, \\ &= \max_{\mathbf{z}, \|\mathbf{z}\|_2=1} \left| \sum_{1 \leq p \leq k, 1 \leq q \leq k} \mathbf{z}_p^T \mathbf{y}_i \mathbf{y}_i^T \mathbf{z}_q \mathbf{x}_i^T \left(\sum_{1 \leq \ell \leq k} \langle \mathbf{u}_\ell, \mathbf{u}_{*q} \rangle \mathbf{u}_p \mathbf{u}_\ell^T - \mathbf{u}_p \mathbf{u}_{*q}^T \right) \mathbf{x}_i \right|, \\ &= \max_{\mathbf{z}, \|\mathbf{z}\|_2=1} \left| \sum_{1 \leq p \leq k, 1 \leq q \leq k} \mathbf{z}_p^T \mathbf{y}_i \mathbf{y}_i^T \mathbf{z}_q \mathbf{x}_i^T \mathbf{u}_p \mathbf{u}_{*q}^T (UU^T - I) \mathbf{x}_i \right|, \\ &\stackrel{\zeta_1}{\leq} \delta \max_{\mathbf{z}, \|\mathbf{z}\|_2=1} \sum_{1 \leq p \leq k, 1 \leq q \leq k} \|(UU^T - I) \mathbf{u}_{*q}\|_2 \|\mathbf{z}_p\|_2 \|\mathbf{z}_q\|_2 \leq k \cdot \delta \cdot \text{dist}(U, U_*), \end{aligned} \quad (42)$$

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

Lemma now follows by using (42) along with (37) and (41). \square

Proof of Lemma 11. The lemma is exactly the same as Lemma 4.7 of [8]. We reproduce their proof here for completeness.

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

$$\begin{aligned} \sigma_{\min}(R) &= \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \|R\mathbf{z}\|_2 = \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \|V R \mathbf{z}\|_2 = \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \|V_* \Sigma_* U_*^T U \mathbf{z} - F \mathbf{z}\|_2, \\ &\geq \min_{\mathbf{z}, \|\mathbf{z}\|_2=1} \|V_* \Sigma_* U_*^T U \mathbf{z}\|_2 - \|F \mathbf{z}\|_2 \geq \sigma_*^k \sigma_{\min}(U^T U_*) - \|F\|_2, \\ &\geq \sigma_*^k \sqrt{1 - \|U^T U_*^\perp\|_2^2} - \|F\|_2 = \sigma_*^k \sqrt{1 - \text{dist}(U_*, U)^2} - \|F\|_2. \end{aligned} \quad (43)$$

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