Noisy Matrix Completion Using Alternating Minimization

Suriya Gunasekar, Ayan Acharya, Neeraj Gaur, and Joydeep Ghosh

Department of ECE, University of Texas at Austin, USA
{suriya,aacharya,neeraj.gaur}@utexas.edu, ghosh@ece.utexas.edu

Abstract. The task of matrix completion involves estimating the entries of a matrix, $M \in \mathbb{R}^{m \times n}$, when a subset, $\Omega \subset \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$ of the entries are observed. A particular set of low rank models for this task approximate the matrix as a product of two low rank matrices, $\widehat{M} = UV^T$, where $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$ and $k \ll \min\{m,n\}$. A popular algorithm of choice in practice for recovering M from the partially observed matrix using the low rank assumption is alternating least square (ALS) minimization, which involves optimizing over U and V in an alternating manner to minimize the squared error over observed entries while keeping the other factor fixed. Despite being widely experimented in practice, only recently were theoretical guarantees established bounding the error of the matrix estimated from ALS to that of the original matrix M. In this work we extend the results for a noiseless setting and provide *the first guarantees for recovery under noise for alternating minimization*. We specifically show that for well conditioned matrices corrupted by random noise of bounded Frobenius norm, if the number of observed entries is $\mathcal{O}\left(k^7 n \log n\right)$, then the ALS algorithm recovers the original matrix within an error bound that depends on the norm of the noise matrix. The sample complexity is the same as derived in [7] for the noise–free matrix completion using ALS.

1 Introduction

The problem of matrix completion has found application in a number of research areas such as in recommender systems [10], multi-task learning [15], remote sensing[12] and image inpainting [1]. In a typical setting for matrix completion, a matrix $M \in \mathbb{R}^{m \times n}$ is observed on a subset of entries $\Omega \subset \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$, while a large number of entries are missing. The task is then to fill in the missing entries of the matrix yielding an estimate $\widehat{M}$ of the complete matrix that is consistent with the original matrix M.

Among the many models that try and tackle the matrix completion problem, low rank models have enjoyed a great deal of success in practice and have proven to be very popular and effective for the matrix completion task on real life datasets [3,8,10,13,11]. Low rank models with numerous variations have been heavily used in practice for matrix completion specially towards the application of collaborative filtering [10,13]. Though it is one of the most widely used techniques to model incomplete matrix data, there are only a few algorithms for which theoretical guarantees have been established, most notably the nuclear norm minimization [3,4] and OptSpace [8]. However, these algorithms are computationally expensive and hence not scalable.

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part II, LNAI 8189, pp. 194–209, 2013.

A popular algorithm that is heavily used in practice for recovering M from the entries observed on Ω under the low rank assumption is the alternating least squares minimization (ALS) [16,10]. The algorithm makes the assumption that the matrix M is of a fixed low rank that has a latent factor representation $M = UV^T$, where $U \in \mathbb{R}^{m \times k}$, $V \in \mathbb{R}^{n \times k}$ and $k \ll n, m$. Hence, one is interested in solving the following:

$$\min_{U,V} \|P_\Omega(M) - P_\Omega(UV^T)\|_F^2$$

Where Ω is the set of observed entries and $P_\Omega(M)$, also denoted by M^Ω, is the projection of the matrix M onto the observed set Ω, given by, $M_{ij}^\Omega = \begin{cases} M_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$

The above problem as described is jointly non–convex in U and V. Alternating minimization proceeds by alternatively fixing one of the latent factors and optimizing the other. Once one of the factors (say U) is fixed, solving for the other (V) is a convex problem. In fact, it is a simple least squares problem. This simplicity of the alternating minimization has made it a popular approach for low rank matrix factorization in practice. Recent results [7,6,14] give recovery guarantees for ALS in a noiseless setting. However theoretical guarantees for ALS when the observed entries are corrupted by noise are still lacking. On the other hand, in real life applications, the matrix entries are often corrupted by various means including the noise in the matrix generation process, outliers and inaccurate measurements. In this work we present *the first guarantees for recovery under noise for alternating least squares minimization*. We rely heavily on the analysis of [7,6] and also borrow results from [9].

The paper is organized as follows. After explaining the notations and defining a few quantities in Section 1.1, we briefly review relevant work in Section 2. In Section 3, we describe the algorithm and state the main result of the paper and compare the results with the existing results. Our primary contribution in this paper is the proof of the result stated in Section 3. We build the proof in Section 4. As the proof is fairly involved, the proof of various lemmata in this section are deferred to the Appendix. We conclude with an analysis of the results and possible future directions in Section 5.

1.1 Notations and Preliminaries

Unless stated otherwise, we use the following notation in the rest of the paper. Matrices are represented by uppercase letters. For a matrix M, M_i represents the i^{th} column, $M^{(i)}$ represents the vector corresponding to the i^{th} row, (all the vectors are column vectors *i.e* they are or dimension $d \times 1$, where d is the length of the vector) and M_{ij} is the $(i,j)^{\text{th}}$ entry. The spectral norm and Frobenius norm of a matrix M are denoted by $\|M\|_2$ and $\|M\|_F$, respectively. The max norm of M, denoted by M_{max}, is the maximum of the absolute values of the entries of M. The transpose of a matrix M is denoted by $M^\dagger$. Vectors are denoted by lowercase letters. For a vector u, u_i is the ith component of u. The p-norm of a vector is given by $\|u\|_p = \left(\sum_i |u_i|^p\right)^{1/p}$, $p \geq 1$. Finally, set of integers from 1 to m is denoted by $[m] = \{1, 2, \ldots, m\}$.

Definition 1 (SVD (or truncated SVD)). *The singular value decomposition (SVD) of a matrix $M \in \mathbb{R}^{m \times n}$ of rank k is given by $M = U\Sigma V^\dagger$, where $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ have orthonormal columns, i.e. $U^\dagger U = V^\dagger V = I$ and $\Sigma \in \mathbb{R}^{k\times k}$ is a diagonal matrix whose entries are $(\sigma_1, \sigma_2, \ldots, \sigma_k)$. Here, the columns of U and V are called the* ***left and right singular vectors*** *of M respectively and $\sigma_1 \geq \sigma_2, \ldots, \sigma_k > 0$ are the* ***singular values****.*

Definition 2 (Condition number). *Consider a matrix M of rank k, with singular values, $\sigma_1 \geq \sigma_2, \ldots, \sigma_k > 0$. The condition number of the matrix M, denoted by κ_M is defined as $\kappa_M = \frac{\sigma_1}{\sigma_k}$*

Definition 3 (Reduced–QR factorization (or simply QR factorization)). *The Reduced–QR factorization, which is often overloaded as QR factorization, of a matrix $X \in \mathbb{R}^{m\times k}$, $m \geq k$, is given by $X = QR$, where $Q \in \mathbb{R}^{m\times k}$ has orthonormal columns and $R \in \mathbb{R}^{k\times k}$ is an upper triangular matrix. The columns of the matrix Q is an orthonormal basis for the subspace spanned by the columns of X.*

Definition 4 (Distance between two matrices [5]). *Given two matrices $\widehat{U}, \widehat{W} \in \mathbb{R}^{m\times k}$, the distance between the subspaces spanned by the columns of $\widehat{U}$ and $\widehat{W}$ is given by $dist(\widehat{U}, \widehat{W}) = \|U_\perp^\dagger W\|_2 = \|UW_\perp^\dagger\|_2$ where U and W are orthonormal bases of the spaces $span(\widehat{U})$ and $span(\widehat{W})$, respectively. Similarly, $U_\perp$ and $W_\perp$ are orthonormal bases of the spaces $span(\widehat{U}_\perp)$ and $span(\widehat{W}_\perp)$.*

Definition 5 (Incoherence of a matrix). *A matrix $M \in \mathbb{R}^{m\times n}$ is incoherent with parameter μ if $\|U^{(i)}\|_2 \leq \mu\frac{\sqrt{k}}{\sqrt{m}}$ $\forall i \in [m]$ and $\|V^{(j)}\|_2 \leq \mu\frac{\sqrt{k}}{\sqrt{n}}$ $\forall j \in [n]$ where $M = U\Sigma V^\dagger$ is the SVD of M. We remind that $X^{(i)}$ is the i^{th} row of matrix X.*

Definition 6 (Vector to matrix conversion). *The operator vec2mat() converts a vector to matrix in column–order, i.e. $\forall$ $x \in \mathbb{R}^{nk}$, $vec2mat(x) =$*

$$\begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ x_{1:n} & x_{n+1:2n} & \cdots & x_{(k-1)n+1:kn} \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix}$$

2 Related Work

Candès and Recht [3] first demonstrated that under the assumptions of random sampling and incoherence conditions $O(kn^{1.2}\log n)$ samples allow for exact recovery of the true underlying matrix via convex nuclear–norm based minimization. The sample complexity result was further improved to $O(kn \log n)$ by Candès and Tao [4]. Later on, Candès and Plan [2] analyzed the recovery guarantees for nuclear–norm based optimization algorithm under bounded noise added to the true underlying matrix. However, one should note that nuclear–norm based minimization approach is computationally expensive and infeasible in practice for large scale matrices.

In the OptSpace algorithm [8], Keshavan *et al.* adopted a different approach for the matrix completion problem where they first took the SVD of the matrix M^Ω. Their analysis showed that such a SVD provides a reasonably good initial estimate for the spanning subspace, which can further be refined by gradient descent on a Grassmanian manifold. They show asymptotic recovery guarantees of original matrix if the number of samples is $O(nk\,(\sigma_1^*/\sigma_k^*)^2 \log n)$. In a later paper, Keshavan *et al.* [9] also examined the reconstruction guarantee of OptSpace under two noise models. The analysis (for both noiseless and noisy recovery) of the algorithm only guarantees asymptotic convergence and the convergence might take exponential time in the problem size in the worst case.

In practice, however, alternating minimization based approach produces good optimal solution. Though the underlying optimization problem is non–convex, each step is convex, computationally cheaper and solutions close to global optimal are often reported in experiments [11]. The algorithm and its variations have been practically deployed in many real life collaborative filtering datasets and have shown good performance [10,13]. Wang and Xu [14] first showed that given a factorization algorithm attains a global optimum, the space of the factors, U and V, and the estimated matrix $\widehat{M}$ are robust against corruption of the observed entries by bounded noise. Jain *et al.* [7,6], however, were the first to formulate the conditions for recovery of the underlying matrix using alternating minimization. They showed that the true underlying matrix M can be recovered within an error of ϵ in $O(\log(\|M\|_F/\epsilon))$ steps and this requires $O((\sigma_1^*/\sigma_k^*)^4 k^7 n \log n \log \|M\|_F/\epsilon)$ number of samples. We build on the results of Jain *et al.* [7] and provide recovery guarantees of noisy matrix completion problem with alternating minimization.

3 Main Result

In the rest of the paper, the underlying true rank–k matrix to be completed is denoted by $M \in \mathbb{R}^{m\times n}$. With a slight abuse of notation, the truncated SVD of M is given by $M = U^*\Sigma^*V^{*\dagger}$ with $U^* \in \mathbb{R}^{m\times k}$, $V^* \in \mathbb{R}^{n\times k}$ and $\Sigma^* = \text{diag}(\sigma_1^*, \sigma_2^*, \ldots, \sigma_k^*)$. Without loss of generality, it is assumed that $m \leq n$ and $\alpha = n/m \geq 1$ is a constant (independent of n). The noisy matrix which is partially observed is given by $\widetilde{M} = M + N$, where $N \in \mathbb{R}^{m\times n}$ is the noise matrix. Further, let $N = U_N \Sigma_N V_N^\dagger$ be the SVD of the noise matrix with $U_N \in \mathbb{R}^{m\times m}$, $V_N \in \mathbb{R}^{n\times m}$ and $\Sigma_N = \text{diag}(\sigma_1^N, \sigma_2^N, \ldots, \sigma_m^N)$. Each entry of the matrix $\widetilde{M}$ is independently observed with probability p. Let Ω be the set of indices where the matrix $\widetilde{M}$ is observed. The task is to estimate M given $\widetilde{M}^\Omega$ and Ω.

3.1 Noise Model

We consider a fairly general, *worst case model* for the noise matrix N, also used in [9]. In this model N is distributed arbitrarily but bounded as $|N_{ij}| \leq N_{\text{max}}$. This is a generic setting, and any noise distribution with sub Gaussian tails can be approximated by this model with high probability. However, tighter bounds can be obtained for individual cases. Our bounds primarily depend on N_{max} and the fractional operator norm of N^Ω, $\|N^\Omega\|_2/p$. We use the following result from [9]:

Theorem 1 ([9]). *If N is a matrix from the worst case model, then for any realization of N, $\|N^\Omega\|_2 \leq \frac{2|\Omega|}{m\sqrt{\alpha}} N_{max}$.*

Using, $|\Omega| \approx pmn$ in Theorem 1, we have the following bound:

$$\frac{\|N^{\Omega}\|_2}{p} \leq 2\sqrt{mn}N_{max}, \tag{1}$$

3.2 Algorithm

The algorithm analyzed in this paper is presented below [7]:

Algorithm 1. ALSM

```
1: Input: observed set $\Omega$, values $P_{\Omega}(\widetilde{M})$
2: Create $(2T+1)$ subsets from $\Omega$ — $\Omega_1, \Omega_2 \cdots, \Omega_{2T}$, each of size $|\Omega|$, with the elements of $\Omega$ belonging to one of the $\Omega_t$'s with equal probability and sampled independently
3: Set $\widehat{U}^0 = \mathrm{SVD}(P_{\Omega_0}(\widetilde{M})/p, k)$ i.e., top-k left singular vectors of $P_{\Omega_0}(\widetilde{M})/p$
4: Clipping step: Set all elements of $\widehat{U}^0$ that have magnitude greater than $\frac{2\mu\sqrt{k}}{\sqrt{n}}$ to zero and orthonormalize the columns of $\widehat{U}^0$ (using QR decomposition)
5: for $t = 0, \cdots, (T-1)$ do
       $\widehat{V}^{(t+1)} \leftarrow \underset{V \in \mathbb{R}^{n \times k}}{\text{argmin}} \|P_{\Omega^{(t+1)}}(\widehat{U}^t V^{\dagger} - \widetilde{M})\|_F$    (2)
       $\widehat{U}^{(t+1)} \leftarrow \underset{U \in \mathbb{R}^{m \times k}}{\text{argmin}} \|P_{\Omega^{(T+t+1)}}\left(U(\widehat{V}^{(t+1)})^{\dagger} - \widetilde{M}\right)\|_F$    (3)
   end
6: Output: $X = \widehat{U}^T(\widehat{V}^T)^{\dagger}$
```

For ease of analysis, we have modified the standard ALS algorithm. In Step 2 of the algorithm, independently sampled subsets of Ω are generated that are further used in the rest of the algorithm. This modification was introduced purely for the ease of theoretical analysis and is not required in practice. In the above algorithm, in each iteration, t, the observed set $\Omega^{(t)}$ is independent of the other iterations and hence, each iteration could be analyzed independently. In the proof of our main result, while analyzing iteration, t, we overload Ω to represent $\Omega^{(t)}$ to avoid cluttering of symbols. Thus, the final sample complexity for recovery would be $2T$ times the sample complexity requirements in each iteration, where T is the total number of iterations required for convergence.

3.3 Result

Theorem 2. *Let $M = U^*\Sigma^*(V^*)^{\dagger} \in \mathbb{R}^{m \times n}$ be a rank–k, incoherent matrix with both U^* and V^* being μ incoherent. Further, it is assumed that, $N_{max} \leq C_3 \frac{\sigma_k^*}{n\sqrt{k}}$ and $\frac{\|N^{\Omega}\|_2}{p} \leq C_2 \frac{\sigma_k^*}{\kappa_M k}$. Additionally, let each entry of $\widetilde{M} = M + N$ be observed uniformly and independently with probability*

$$p > C \frac{\kappa_M^4 \mu^4 k^7 \log n \log \frac{\|M\|_F}{\epsilon}}{m\delta_{2k}^2} \tag{4}$$

where, $\kappa_M = \frac{\sigma_1^*}{\sigma_k^*}$ *is the condition number of the* M*,* $\delta_{2k} \leq \frac{\sigma_k^*}{64\sigma_1^*}$ *and* $C > 0$ *is a global constant. Then with high probability, for* $T \geq C' \log \frac{\|M\|_F}{\epsilon}$*, the outputs* $\widehat{U}^T$ *and* $\widehat{V}^T$ *of Algorithm 1 with input* $(\Omega, P_\Omega(\widetilde{M}))$ *satisfy*

$$\frac{1}{\sqrt{mn}}\|M - \widehat{U}^T(\widehat{V}^T)^\dagger\|_F \leq \epsilon + 20\mu\kappa_M^2 k^{1.5}\left(\frac{\|N^\Omega\|_2}{|\Omega|}\right) \leq \epsilon + 40\mu\kappa_M^2 k^{1.5} N_{max} \quad (5)$$

Worst Case Noise Model Requirements. The theorem requires that $N_{max} \leq C_3 \frac{\sigma_k^*}{n\sqrt{k}}$ and $\frac{\|N^\Omega\|_2}{p} \leq C_2 \frac{\sigma_k^*}{\kappa_M k}$. For the *worst case noise model*, if $N_{max} \leq C_2 \frac{\sigma_k^*}{2\kappa_M nk} \implies \frac{\|N^\Omega\|_2}{p} \leq 2\sqrt{mn} N_{max} \leq C_2 \frac{\sigma_k^*}{\kappa_M k}$ Further, $N_{max} \leq C_3 \frac{\sigma_k^*}{\kappa_M nk} \implies N_{max} \leq C_3 \frac{\sigma_k^*}{nk} \leq C_3 \frac{\sigma_k^*}{n\sqrt{k}}$. Thus, choosing $C = \min\{C_2/2, C_3\}$, and $N_{max} \leq C \frac{\sigma_k^*}{\kappa_M nk}$, both the conditions on noise matrix for Theorem 2 are satisfied.

For a well conditioned matrix M of condition number close to 1, the above requirement is approximately equivalent to $N_{max} \leq C' k^{-1.5} \frac{\|M\|_F}{\sqrt{mn}}$, which is $k^{-1.5}$ fraction of root mean square value of the entries of matrix M. This is a fairly reasonable assumption on the noise matrix for recovery guarantees.

3.4 Comparison with Similar Results

The most relevant work for our analysis is the analysis of low rank matrix completion under alternating minimization approach proposed by Jain et. al. [7]. They have the following result for ALS under noiseless setting, $N = 0$:

Theorem 3 ([7]). *Let* $M = U^* \Sigma^* (V^*)^\dagger \in \mathbb{R}^{m \times n}$ *be a rank–k, incoherent matrix with both* U^* *and* V^* *being* μ *incoherent. Let each entry of* M *be observed uniformly and independently with probability,* $p > C \frac{\kappa_M^4 \mu^4 k^7 \log n \log \frac{\sqrt{k}\|M\|_2}{\epsilon}}{m\delta_{2k}^2}$ *where,* $\delta_{2k} \leq \frac{\sigma_k^*}{64\sigma_1^*}$ *and* $C > 0$ *is a global constant. Then with high probability, for* $T \geq C' \log \frac{\|M\|_F}{\epsilon}$*, the outputs* $\widehat{U}^T$ *and* $\widehat{V}^T$ *of Algorithm 1 with input* $(\Omega, P_\Omega(\widetilde{M}))$ *satisfy* $\|M - \widehat{U}^T(\widehat{V}^T)^\dagger\|_F \leq \epsilon$

Even for a very general noise model, the sample complexity required for our analysis is the same as that required by the noise–free analysis.

Next, we compare our bounds with the bounds obtained for noisy matrix completion by Keshavan et. al [9]. The algorithm suggested by Keshavan et. al., OptSpace, involves optimizing the initial estimate from SVD of $P_\Omega(\widetilde{M})$ over a Grassmann manifold. The main result in their paper is stated below:

Theorem 4 ([9]). *Let* $\widetilde{M} = M + N$*, where* M *is a rank–k,* μ *incoherent matrix. A subset,* $\Omega \subset [m] \times [n]$*, of entries of* $\widetilde{M}$ *are revealed. Let* $\widehat{M}$ *be the output of OptSpace on the input of* $(\widetilde{M}, \Omega)$*. Then, there exists numerical constants,* C *and* C' *such that, if* $|\Omega| \geq Cn\sqrt{\alpha}\kappa_M^2 \max\{\mu k\sqrt{\alpha}\log n; \mu^2 k^2 \alpha \kappa_M^4\}$ *and* $\frac{\|N^\Omega\|_2}{p} \leq C' \frac{\sigma_k^*}{\kappa_M^2 \sqrt{k}}$ *then, with probability atleast* $1 - 1/n^3$*,* $\frac{1}{\sqrt{mn}}\|M - \widehat{M}\|_F \leq C' \kappa_M^2 k^{0.5} \left(\frac{\|N^\Omega\|_2}{|\Omega|}\right)$

The requirements on the noise matrix for recovery guarantees by OptSpace is close to that derived in our results for Alternating minimization. Also, the error in the recovered matrix in our analysis is off by a small factor of k as compared to the analysis in [9]. However, the sample complexity required by ALS as evaluated by our analysis is much higher than that of Keshavan et. al.

4 Proof of Theorem 2

In this section we present the proof of Theorem 2. The outline of the proof is as follows. In Section 4.1, Theorem 5 states that the initialization step of the Algorithm described in 1 provides a good starting point. In Section 4.2, we first propose a modification to the ALSM algorithm and prove that the modification in practice is equivalent to the original ALSM algorithm, while the modified algorithm is easier to analyze. Theorem 6 is then stated without proof. This theorem establishes that the space spanned by ALSM estimates of $\widehat{U}$ and $\widehat{V}$ converge towards U^* and V^* respectively. Finally, we combine the results on initialization and above mentioned theorem to prove the main result. The proof of Theorem 6 is deferred to Section 4.3. In each subsection, the relevant lemmata are first presented and then the main theorems are proved. The proofs of the lemmata are provided in the Appendix.

4.1 Initialization

Lemma 1 (Theorem 1.1 **of [8]).** *Let* $\widetilde{M} = M + N$ *be such that* M *is rank–*k *and* μ*–incoherent and* $|\Omega| \geq Cnk\max\{\log n, k\}$. *Further, from the SVD of* $\frac{\widetilde{M}^{\Omega}}{p}$, *we get a rank–*$k$ *approximation as,* $\widetilde{M}_k^{\Omega} = \widetilde{U}^0\widetilde{\Sigma}^0\widetilde{V}^{0\dagger}$, *where* $\widetilde{U}^0 \in \mathbb{R}^{m\times k}$ *and* $\widetilde{V}^0 \in \mathbb{R}^{n\times k}$. *Let* $\alpha = n/m \geq 1$. *Then the following is true with probability greater than* $(1 - 1/n^3)$,

$$\frac{1}{\sqrt{mn}}\|M - \widetilde{M}_k^{\Omega}\|_2 \leq CM_{max}\left(\frac{m\alpha^{3/2}}{|\Omega|}\right)^{1/2} + \frac{2m\sqrt{\alpha}}{|\Omega|}\|N^{\Omega}\|_2. \qquad (6)$$

Lemma 2. *Let* $\widetilde{U}^0$ *be defined as in Lemma 1. Further, under the conditions of Theorem 2, the following is true with probability greater than* $(1 - 1/n^3)$,

$$dist(\widetilde{U}^0, U^*) \leq \frac{1}{64k}.$$

The proof of Lemma 2 is presented in Appendix B.1.

Theorem 5 (ALSM has a good initial point). *Let* U^c *be obtained from* $\widetilde{U}^0$ *defined above, by setting all the entries greater than* $\frac{2\mu\sqrt{k}}{\sqrt{m}}$ *to zero. Let* U^0 *be the orthonormal basis of* U^c. *Then under the conditions of Lemma 2, w.h.p. we have*

- $dist(U^0, U^*) \leq 1/2$.
- U^0 *is incoherent with parameter* $\mu_1 = \frac{32\sigma_1^*\mu\sqrt{k}}{\sigma_k^*}$.

The proof follows directly from Lemma C.2 in [6] and Lemma 2. □

Note that the U^0 defined above is the same as the the initial estimate, $\widehat{U}^0$ from the initialization step of the Algorithm 1.

4.2 Convergence of ALS Minimization

Consider the following modification to Equation 2 and 3 of Algorithm 1:

$$\begin{aligned}
\widehat{V}^{(t+1)} &\leftarrow \underset{\widehat{V} \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \|P_{\Omega^{(t+1)}}(U^t \widehat{V}^\dagger - \widetilde{M})\|_F \\
V^{(t+1)} R_V^{(t+1)} &= \widehat{V}^{(t+1)} \quad \text{(QR decomposition)} \\
\widehat{U}^{(t+1)} &\leftarrow \underset{\widehat{U} \in \mathbb{R}^{m \times k}}{\operatorname{argmin}} \|P_{\Omega^{(T+t+1)}} \left(\widehat{U} V^{(t+1)^\dagger} - \widetilde{M} \right) \|_F \\
U^{(t+1)} R_U^{(t+1)} &= \widehat{U}^{(t+1)} \quad \text{(QR decomposition)}
\end{aligned} \tag{7}$$

Lemma 3 (Lemma 4.4 of [7]). *Let $\widehat{U}^{(t)}$ be the t-th step iterate of ALSM Algorithm 1, and $\widetilde{U}^{(t)} = U^{(t)} R_U^{(t)}$ be that of the modified algorithm presented above. Suppose that both $\widetilde{U}^{(t)}$ and $\widehat{U}^{(t)}$ are full rank and span the same space, then the same will be true for subsequent iterates. i.e $span(\widehat{V}^{(t+1)}) = span(\widetilde{V}^{(t+1)})$ and $span(\widehat{U}^{(t+1)}) = span(\widetilde{U}^{(t+1)})$ and all the matrices at iterate $t+1$ are full rank.*

Proof. As both $\widetilde{U}^{(t)}, \widehat{U}^{(t)} \in \mathbb{R}^{m \times k}$ have full rank and span same subspace, there exists a $k \times k$ full rank matrix R such that $\widehat{U}^{(t)} = \widetilde{U}^{(t)} R = U^{(t)} R_U^{(t)} R$. Thus,

$$\begin{aligned}
\min_{V \in \mathbb{R}^{n \times k}} \|P_{\Omega^{(t+1)}} \left(\widehat{U}^{(t)} V^\dagger - \widetilde{M} \right) \|_2 \quad &= \|P_{\Omega^{(t+1)}} \left(\widehat{U}^{(t)} \widehat{V}^{(t+1)\dagger} - \widetilde{M} \right) \|_2 \\
= \|P_{\Omega^{(t+1)}} \left(U^{(t)} (\widehat{V}^{(t+1)} (R_U^{(t)} R)^\dagger)^\dagger - \widetilde{M} \right) \|_2 \quad &\geq \min_{V \in \mathbb{R}^{n \times k}} \|P_{\Omega^{(t+1)}} \left(U^{(t)} V^\dagger - \widetilde{M} \right) \|_2 \\
&= \|P_{\Omega^{(t+1)}} \left(U^{(t)} \widetilde{V}^{(t+1)\dagger} - \widetilde{M} \right) \|_2
\end{aligned}$$

The above equation holds with equality for $\widehat{V}^{(t+1)} = \widetilde{V}^{(t+1)} \left((R_U^t R)^\dagger \right)^{-1}$. Further Theorem 6 shows that $\widetilde{V}^{(t+1)}$ is full rank (as dist$(\widetilde{V}^{(t+1)}, V^*) < 1$) and hence, $\widehat{V}^{(t+1)} = \widetilde{V}^{(t+1)} \left((R_U^t R)^\dagger \right)^{-1}$ is full rank and their columns span the same subspace. Similar arguments can be used to show that $\widehat{U}^{(t+1)}$ and $\widetilde{U}^{(t+1)}$ are both full rank and span the same subspace. □

Further, as the initial estimate, $\widehat{U}^0$ satisfies the conditions of the above lemma, in the rest of the proof it is assumed that the distances dist$(\widehat{U}^t, U^*)$ and dist$(\widehat{V}^t, V^*)$ are the same for the updates from both ALSM and its modified version presented above.

Theorem 6 (Each step of ALSM is good). *Under the assumptions of Theorem 2, the $(t+1)^{th}$ iterates, $\widehat{U}^{t+1}$ and $\widehat{V}^{t+1}$ satisfy the following w.h.p:*

$$\begin{aligned}
dist\left(\widehat{V}^{t+1}, V^*\right) &\leq \frac{1}{4} dist\left(\widehat{U}^t, U^*\right) + 10 \frac{\mu \kappa_M \|N^\Omega\|_2 k}{\sigma_k^* p} \\
dist\left(\widehat{U}^{t+1}, U^*\right) &\leq \frac{1}{4} dist\left(\widehat{V}^{t+1}, U^*\right) + 10 \frac{\mu \kappa_M \|N^\Omega\|_2 k}{\sigma_k^* p}
\end{aligned} \tag{8}$$

where, $\kappa_M = \sigma_1^* / \sigma_k^*$ is the condition number of the matrix M.

The proof of Theorem 6, involves few other lemmata and is deferred to Section 4.3. The main theorem is now proved using the results from Theorem 5 and 6. □

Proof of Main Result, Theorem 2. From Theorem 6, after $T = \mathcal{O}(\log \frac{\sqrt{k}\|M\|_2}{\epsilon})$ steps, we have:

$$\text{dist}\left(\widehat{U}^T, U^*\right) \leq \frac{\epsilon}{2\sqrt{k}\|M\|_2} + 10\frac{\mu\kappa_M\|N^\Omega\|_2 k}{\sigma_k^* p} \tag{9}$$

Using Lemma 4, we have that:

$$\|M - \widehat{U}^T\widehat{V}^{(T+1)\dagger}\|_F \leq \|(I - \widehat{U}^T\widehat{U}^{T\dagger})U^*\Sigma^*\|_F + \|F\|_F + \|N_{res}\|_F \tag{10}$$

Note that the bounds on $\|F\|_2$ and $\|N_{res}\|_2$ from Lemma 5 and Equation 16, also hold for both $\|F\|_F$ and $\|N_{res}\|_F$ respectively. This can be seen from the proofs of Lemmata 5 and 6. Using these bounds we have the following:

$$\begin{aligned}\|M - \widehat{U}^T\widehat{V}^{(T+1)\dagger}\|_F \leq & \sqrt{k}\sigma_1^*\text{dist}(\widehat{U}^T, U^*) + \frac{\delta_{2k}\sigma_1^*}{1-\delta_{2k}}\text{dist}(\widehat{U}^T, U^*) + C\sigma_k^*\frac{10\mu\kappa_M\|N^\Omega\|_2 k}{\sigma_k^* p} \\ \leq & \ \epsilon + \frac{20\mu\kappa_M^2\|N^\Omega\|_2 k^{1.5}}{p}\end{aligned} \tag{11}$$

Further, in order that each of the $2T+1$ sub-sampled indices Ω_t has $\mathcal{O}\left(\frac{\mu^4\kappa_M^4 k^7\log n}{m\delta_{2k}^2}\right)$ samples, the total sample complexity required is $\mathcal{O}\left(\frac{\mu^4\kappa_M^4 k^7 \log n \log\frac{\sqrt{k}\|M\|_2}{\epsilon}}{m\delta_{2k}^2}\right)$. □

4.3 Proof of Theorem 6

To avoid cluttering of notations we define a few quantities first. In the following definitions, we recall that U^* and U_N are the left singular vectors of M and N respectively, and U^t is the t^{th} step iterate of the modified algorithm ($U^t = \widehat{U}^t(R_U^t)^{-1}$). Further $U^{(i)}$ and U_i represent the i^{th} row and column vectors of U respectively and U_{ij} is the $(i,j)^{\text{th}}$ entry of U. For $1 \leq p \leq k$ and $1 \leq q \leq k$ we define diagonal matrices B_{pq}, C_{pq}, $D_{pq} \in \mathbb{R}^{n\times n}$, where, $D_{pq} = \langle U_p^t, U_q^*\rangle \mathbb{I}_{n\times n}$ and the, j^{th} diagonal entries of B_{pq} and C_{pq} are given by:

$$(B_{pq})_{jj} = \left[\frac{1}{p}\sum_{i:(i,j)\in\Omega} U_{ip}^t U_{iq}^t\right], (C_{pq})_{jj} = \left[\frac{1}{p}\sum_{i:(i,j)\in\Omega} U_{ip}^t U_{iq}^*\right].$$

Using the above matrices, we define the following matrices of dimension $nk \times nk$:

$$B \triangleq \begin{bmatrix} B_{11} & \cdots & B_{1k} \\ \vdots & \ddots & \vdots \\ B_{k1} & \cdots & B_{kk}\end{bmatrix}, C \triangleq \begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \cdots & C_{kk}\end{bmatrix}, D \triangleq \begin{bmatrix} D_{11} & \cdots & D_{1k} \\ \vdots & \ddots & \vdots \\ D_{k1} & \cdots & D_{kk}\end{bmatrix}, S \triangleq \begin{bmatrix} \sigma_1^*\mathbb{I}_n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k^*\mathbb{I}_n\end{bmatrix}.$$

Analogously, we define matrices, $C^N \in \mathbb{R}^{nk\times nm}$ and $S^N \in \mathbb{R}^{nm\times nm}$ as follows:

$$C^N \triangleq \begin{bmatrix} C_{11}^N & \cdots & C_{1m}^N \\ \vdots & \ddots & \vdots \\ C_{k1}^N & \cdots & C_{km}^N\end{bmatrix}, S^N \triangleq \begin{bmatrix} \sigma_1^N\mathbb{I}_n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^N\mathbb{I}_n\end{bmatrix} \tag{12}$$

where, $\forall\ 1 \leq p \leq k$ and $1 \leq q \leq m$, diagonal matrices $C^N_{pq} \in \mathbb{R}^{n\times n}$ are defined as $(C^N_{pq})_{jj} = \left[\frac{1}{p}\sum_{i:(i,j)\in\Omega} U^t_{ip}U^N_{iq}\right]$. Additionally, we define the following vectors:

$$v^* = [V_1^{\dagger}, V_2^{\dagger}, \cdots, V_k^{\dagger}]^{\dagger} \in \mathbb{R}^{nk}, \quad v^N = [V_1^{N\dagger}, V_2^{N\dagger}, \cdots, V_m^{N\dagger}]^{\dagger} \in \mathbb{R}^{nm}.$$

Finally, we define the matrices, $F = \text{vec2mat}\left(B^{-1}(BD - C)Sv^*\right) \in \mathbb{R}^{n\times k}$ and $N_{res} = \text{vec2mat}\left(B^{-1}C^N S^N v^N\right) \in \mathbb{R}^{n\times k}$.

Lemma 4. *Let $\widehat{U}^t$ be the t^{th} step iterate of the above algorithm and let U^t, $\widehat{V}^{t+1}$ and V^{t+1} be obtained by Updates in 7. Then, using the matrices defined above, we have:*

$$\widehat{V}^{t+1} = V^*\Sigma^*U^{*\dagger}U^t - F + N_{res} \tag{13}$$

The proof of the above lemma is provided in Appendix B.2.

Lemma 5. *Let F be the error matrix defined above and let U^t be a μ_1 incoherent orthonormal matrix obtained from the t^{th} update. Under the conditions of Theorem 2, with probability at least $1 - 1/n^3$, $\|F\|_2 \leq \frac{\delta_{2k}\sigma_1^*}{1-\delta_{2k}}$ dist(U^t, U^*).*

This is the same as Lemma 5.6 of [7] and the proof follows exactly for the noisy case.

Lemma 6. *Let N_{res} be the matrix defined above. Under the conditions of the Theorem 2 with probability at least $1 - 1/n^3$*

$$\|N_{res}\|_2 \leq \frac{\mu_1\sqrt{k}}{(1-\delta_{2k})}\left(\frac{\|N^{\Omega}\|_2}{p}\right) \tag{14}$$

Lemma 7. *Let $R_V^{(t+1)}$ be the upper triangular matrix obtained by QR decomposition of $\widehat{V}^{t+1}$ an. Let F, N_{res} and U^t be defined as above. Then,*

$$\left\|\left(R_V^{(t+1)}\right)^{-1}\right\|_2 \leq \frac{1}{\left[\sigma_k^*\sqrt{1 - dist^2(U^t, U^*)} - \|F\|_2 - \|N_{res}\|_2\right]} \tag{15}$$

The proof of Lemma 6 and 7 are provided in Appendix B.3 and B.4 respectively. We now use the above lemmata to prove Theorem 6.

If $\delta_{2k} \leq \frac{\sigma_k^*}{C\sigma_1^*}$ for appropriate $C > 1$, then $\frac{1}{1-\delta_{2k}} \leq C/(C-1) = C_1$. Further as dist$(U^{(t)}, U^*) \leq$ dist$(U^{(0)}, U^*) \leq 1/2$, we have $\sqrt{1 - dist^2(U^{(t)}, U^*)} \geq \sqrt{3}/2$. Finally, from Lemma 8, we have $\mu_1 = \frac{32\sigma_1^*\mu\sqrt{k}}{\sigma_k^*}$. This implies that $\|N_{res}\|_2 \leq \frac{32\mu\kappa_M k}{1-\delta_{2k}}\left(\frac{\|N^{\Omega}\|_2}{p}\right)$. If further we have that $\frac{\|N^{\Omega}\|_2}{p} \leq C_2\frac{\sigma_k^*}{\kappa_M k}$, then for small enough C_2, we have

$$\|N_{res}\|_2 \leq C_4\mu\kappa_M k\frac{\|N^{\Omega}\|_2}{p} \leq C'\sigma_k^* \tag{16}$$

Using Lemma 4, we have:

$$\begin{aligned}\text{dist}\left(V^*, V^{(t+1)}\right) &= \left\| \left[V_{\perp}^{*\dagger} V^* \Sigma^* U^{*\dagger} - V_{\perp}^{*\dagger} F + V_{\perp}^{*\dagger} N_{res} \right] (R_V^{(t+1)})^{-1} \right\|_2 \\ &\leq \left(\|F\|_2 + \|N_{res}\|_2 \right) \left\| \left(R_V^{(t+1)} \right)^{-1} \right\|_2 \qquad (17)\end{aligned}$$

For appropriate choice of $C > 1$ and small enough $C' < 1$, we have From Lemma 5, 6 and 7 and Equation 17, we have the following:

$$\text{dist}(V^*, V^{(t+1)}) \leq \frac{1}{4} dist(U^{(t)}, U^*) + 10 \frac{\mu \kappa_M \|N^{\Omega}\|_2 k}{\sigma_k^* p}$$

Incoherence of Solutions in Each Iteration

Lemma 8. *Under the conditions of Theorem 2, let U^t be the t^{th} step iterate obtained by Eq. 3. If U^t is $\mu_1 = \frac{32\sigma_1^* \mu \sqrt{k}}{\sigma_k^*}$ incoherent then with probability at least $(1 - 1/n^3)$, the solution $V^{(t+1)}$ obtained from Eq. 7 is also μ_1 incoherent.*

The proof of the above lemma can be found in Appendix B.5. As for $t = 0$, U^0 is μ_1 incoherent, the theorem can be used for inductively proving that U^t and V^t are μ_1 incoherent for all t. □

5 Conclusion

We have established the first theoretical guaranties for recovery of a low rank matrix perturbed by bounded noise, using alternating least squares minimization algorithm. The algorithm is computationally more scalable than the algorithms that have previously established error bounds under noisy observations. We use the *worst case noise model* and it is observed that for well conditioned matrices, the main result requires a reasonable bound on the maximum noise entry. The results establish that under the conditions of incoherence of the underlying matrix M and bounded noise, with sufficient samples, the Frobenius norm of the deviation of the recovered matrix, $\widehat{M}$, from the original matrix M, $\frac{\|M - \widehat{M}\|_F}{\sqrt{mn}}$ can be made arbitrarily close to $Ck^{1.5}N_{max}$. Finally, for well conditioned matrices, the sample complexity is $\mathcal{O}(k^7 n \log n)$. This is the same complexity as that required by the current proof of recovery guaranties of ALSM under noiseless setting. However, this is looser compared to the established bounds of other algorithms like nuclear norm minimization and OptSpace and tightening the sample complexity will be considered in the future work. Another direction for future work would include bounding the ALSM algorithm with cost function modified to include regularization on the factors U and V.

Acknowledgments. This research was supported by ONR ATL Grant N00014-11-1-0105, NSF Grants (IIS-0713142 and IIS-1016614). We also thank Praneeth Netrapalli for his valuable comments.

References

1. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. IEEE Transactions on Image Processing (2003)
2. Candès, E.J., Plan, Y.: Matrix completion with noise. CoRR (2009)
3. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics (2009)
4. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory (2010)
5. Golub, G.H., van Van Loan, C.F.: Matrix Computations (Johns Hopkins Studies in Mathematical Sciences), 3rd edn. The Johns Hopkins University Press (1996)
6. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. ArXiv e-prints (December 2012)
7. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: STOC (2013)
8. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Transactions on Information Theory (2010)
9. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. JMLR (2010)
10. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer (2009)
11. Mitra, K., Sheorey, S., Chellappa, R.: Large-scale matrix factorization with missing data under additional constraints. In: NIPS (2010)
12. So, A.M.C., Ye, Y.: Theory of semidefinite programming for sensor network localization. In: ACM-SIAM Symposium on Discrete Algorithms (2005)
13. Takács, G., Pilászy, I., Németh, B., Tikk, D.: Investigation of various matrix factorization methods for large recommender systems. In: KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition (2008)
14. Wang, Y., Xu, H.: Stability of matrix factorization for collaborative filtering. In: ICML (2012)
15. Yu, K., Tresp, V.: Learning to learn and collaborative filtering. In: NIPS Workshop on Inductive Transfer: 10 Years Later (2005)
16. Zhou, Y., Wilkinson, D., Schreiber, R., Pan, R.: Large-scale parallel collaborative filtering for the netflix prize. In: Fleischer, R., Xu, J. (eds.) AAIM 2008. LNCS, vol. 5034, pp. 337–348. Springer, Heidelberg (2008)

Appendix A

- $\|M\|_2 \le \|M\|_F \le \sqrt{k}\|M\|_2$
- If a matrix M is μ-incoherent, then,

$$M_{max} \le \frac{\mu^2\sqrt{k}}{\sqrt{mn}}\|M\|_F \le \frac{\mu^2 k}{\sqrt{mn}}\|M\|_2 \tag{18}$$

- *Bernstein's Inequality*: Let $X_i,\ i = \{1, 2, \ldots, n\}$ be independent random numbers. Let $|X_i| \le L\ \forall\ i$ w.p. 1. Then we have the following inequalities:

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

$$P\left[\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}[X_i]\ < -t\right] \le \exp\left(\frac{-t^2/2}{\sum_{i=1}^n Var(X_i)+Lt/3}\right) \tag{19}$$

Appendix B

B.1 Initialization Proofs

Proof (Proof of Lemma 2)

$$\begin{aligned}\|M-\widetilde{M}_k^{\Omega}\|_2^2 &= \|U^*\Sigma^*V^{*\dagger}-\widetilde{U}^0\widetilde{\Sigma}^0\widetilde{V}^{0\dagger}\|_2^2\\ &= \|(I-\widetilde{U}^0\widetilde{U}^{0\dagger})U^*\Sigma^*V^{*\dagger}+\widetilde{U}^0(\widetilde{U}^{0\dagger}U^*\Sigma^*V^{*\dagger}-\widetilde{\Sigma}^0\widetilde{V}^{0\dagger})\|_2^2\\ &\overset{(1)}{=} \|(I-\widetilde{U}^0\widetilde{U}^{0\dagger})U^*\Sigma^*V^{*\dagger}\|_2^2+\|\widetilde{U}^0(\widetilde{U}^{0\dagger}U^*\Sigma^*V^{*\dagger}-\Sigma V^{\dagger})\|_2^2\\ &\geq \|(I-\widetilde{U}^0\widetilde{U}^{0\dagger})U^*\Sigma^*V^{*\dagger}\|_2^2=\|\widetilde{U}_{\perp}^{0\dagger}U^*\Sigma^*\|_2^2\geq\sigma_k^{*2}\|\widetilde{U}_{\perp}^{0\dagger}U^*\|_2^2\end{aligned}$$

where, (1) follows as the two terms span orthogonal spaces. Hence,

$$\begin{aligned}\text{dist}(\widetilde{U}^0,U^*) &\leq \frac{1}{\sigma_k^*}\|M-\widetilde{M}_k^{\Omega}\|_2\overset{(2)}{\leq}\frac{1}{\sigma_k^*}\left(CM_{max}\sqrt{\frac{m\alpha^{3/2}}{p}}+2\frac{\|N^{\Omega}\|_2}{p}\right)\\ &\overset{(3)}{\leq} C\mu^2k\frac{\sigma_1^*}{\sigma_k^*}\sqrt{\frac{m\alpha^{3/2}}{pmn}}+\frac{2\|N^{\Omega}\|_2}{p\sigma_k^*}\\ &\leq \frac{1}{64k},\text{ if } p>\frac{C'k^4\mu^4\sigma_1^{*2}}{m\sigma_k^{*2}}\text{ and }\frac{\|N^{\Omega}\|_2}{p}\leq C''\frac{\sigma_k^*}{k}.\end{aligned}$$

where, (2) follows from Lemma 1 and (3) follows from Equation 18

B.2 Proof of Lemma 4

We recall that $M^{(i)}$ is the i^{th} row of the matrix M. Given, U^t, the t^{th} step iterate. The update of $\widehat{V}^{(t+1)}$ is guided by the following equation from 7.

$$\begin{aligned}\widehat{V}^{(t+1)} &= \underset{V\in\mathbb{R}^{n\times k}}{\text{argmin}}\|P_{\Omega}(U^tV^{\dagger})-P_{\Omega}(\widetilde{M})\|_F^2\\ &= \underset{V\in\mathbb{R}^{n\times k}}{\text{argmin}}\sum_{(i,j)\in\Omega}\left(U^{t(i)\dagger}V^{(j)}-U^{*(i)\dagger}\Sigma^*V^{*(j)}-U_N^{(i)\dagger}\Sigma_NV_N^{(j)}\right)^2\end{aligned}$$

Taking the gradient with respect to each $V^{(j)}$ and setting it to 0 for the optimum $V=\widehat{V}^{(t+1)}$, we have the following $\forall\, j\in[n]$:

$$\sum_{i:(i,j)\in\Omega}U^{t(i)}\left(U^{t(i)\dagger}\left(\widehat{V}^{(t+1)}\right)^{(j)}-U^{*(i)\dagger}\Sigma^*V^{*(j)}-U_N^{(i)\dagger}\Sigma_NV_N^{(j)}\right)=0 \tag{20}$$

We further define matrices $B^j,C^j,D^j\in\mathbb{R}^{k\times k}$ and $C_N^j\in\mathbb{R}^{k\times m}$ for $1\leq j\leq n$ as follows:

$$\begin{aligned}B^j&=\frac{1}{p}\sum_{i:(i,j)\in\Omega}U^{t(i)}U^{t(i)\dagger},\;\; C^j=\frac{1}{p}\sum_{i:(i,j)\in\Omega}U^{t(i)}U^{*(i)\dagger},\\ D^j&=U^{t\dagger}U^*,\;\; C_N^j=\frac{1}{p}\sum_{i:(i,j)\in\Omega}U^{t(i)}U^{N(i)\dagger}.\end{aligned}\tag{21}$$

It is useful to note that $B^j \in \mathbb{R}^{k\times k}$ is obtained by taking the j^{th} diagonal elements of B_{pq} (defined in Equation 12), for $1 \le p, q \le k$, i.e. $(B^j)_{pq} = (B_{pq})_{jj}$. The other matrices are defined similarly. Using the above matrices, we have from the previous equation:

$$\begin{aligned}\left(\widehat{V}^{(t+1)}\right)^{(j)} &= D^j \Sigma^* V^{*(j)} - (B^j)^{-1}\left(B^j D^j - C^j\right)\Sigma^* V^{*(j)} + (B^j)^{-1} C_N^j \Sigma_N V_N^{(j)} \\ \widehat{V}^{(t+1)} &= V^* \Sigma^* U^{*\dagger} U^t - F + N_{res}\end{aligned} \tag{22}$$

The last equation above can be easily seen by writing the structure of matrices defined above.

B.3 Proof of Lemma 6

From Lemma C.6 of [6], under the assumptions on p and M specified in the Lemma, we have, $\|B^{-1}\|_2 \le \frac{1}{1-\delta_{2k}}$. Further from the structure of the matrices C^N, S^N and C_N^j, it can be verified that $\|C^N S^N v^N\|_2^2 = \sum_{j=1}^{n} \|C_N^{(j)} \Sigma_N V_N^{(j)}\|_2^2$. Recall that $V_N^{(j)} \in \mathbb{R}^{m\times 1}$ is the jth row of $V_N \in \mathbb{R}^{n\times m}$ (a similar decomposition is used in Equation 22). Thus we have:

$$\begin{aligned}\left\|C^N S^N v^N\right\|_2^2 &= \sum_{j=1}^{n} \left\|C^{(j)} \Sigma_N V_N^{(j)}\right\|_2^2 = \sum_{j=1}^{n} \left\| \frac{1}{p} \sum_{i:(i,j)\in\Omega} U^{t(i)} U_N^{(i)\dagger} \Sigma_N V_N^{(j)} \right\|_2^2 \\ &\le \frac{1}{p^2} \sum_{j=1}^{n} \sum_{i:(i,j)\in\Omega} \|U^{t(i)} N_{ij}\|_2^2 \le \frac{1}{p^2} \sum_{(i,j)\in\Omega} \|U^{t(i)}\|_2^2 |N_{ij}|_2^2 \\ &\le \mu_1^2 k \left(\frac{\|N^\Omega\|_F}{\sqrt{mp}}\right)^2 \le \mu_1^2 k \left(\frac{\|N^\Omega\|_2}{p}\right)^2\end{aligned} \tag{23}$$

This implies that

$$\begin{aligned}\|N_{res}\|_2 \le \|N_{res}\|_F = \|B^{-1} C^N S^N v^N\|_2 &\le \|B^{-1}\|_2 \|C^N S^N v^N\|_2 \\ &\le \frac{\mu_1 \sqrt{k}}{(1-\delta_{2k})} \left(\frac{\|N^\Omega\|_2}{p}\right).\end{aligned} \tag{24}$$

B.4 Proof of Lemma 7

$$\begin{aligned}\frac{1}{\|(R_V^{(t+1)})^{-1}\|_2} &= \sigma_{min}(R_V^{(t+1)}) = \min_{z:\|z\|_2=1} \|R_V^{(t+1)} z\|_2 = \min_{z:\|z\|_2=1} \|V^{(t+1)} R_V^{(t+1)} z\|_2 \\ &= \min_{z:\|z\|_2=1} \|\widehat{V}^{(t+1)} z\|_2 = \min_{z:\|z\|_2=1} \|\left(V^* \Sigma^* (U^*)^\dagger U^t - F + N_{res}\right) z\|_2 \\ &\ge \min_{z:\|z\|_2=1} \left[\|V^* \Sigma^* (U^*)^\dagger U^t z\|_2 - \|F\|_2 - \|N_{res}\|_2\right] \\ &\ge \sigma_k^* \min_{z:\|z\|_2=1} \|(U^*)^\dagger U^t z\|_2 - \|F\|_2 - \|N_{res}\|_2 \\ &= \sigma_k^* \sqrt{1 - \text{dist}(U^t, U^*)^2} - \|F\|_2 - \|N_{res}\|_2\end{aligned}$$

Thus, $\|(R_V^{(t+1)})^{-1}\|_2 \leq \frac{1}{\sigma_k^* \sqrt{1-\text{dist}(U^t,U^*)^2} - \|F\|_2 - \|N_{res}\|_2}$

B.5 Proof of Lemma 8

In this proof, we use the following set of inequalities:

$$\begin{aligned} &\|(B^j)^{-1}\|_2 \leq \frac{1}{1+\delta_{2k}} \\ &\|B^j\|_2 \leq 1+\delta_{2k},\ \|C^j\|_2 \leq 1+\delta_{2k},\ \|D^j\| \leq \|U^*\|_2\|U^t\|_2 = 1. \end{aligned} \tag{25}$$

The above set of equations involve terms that does not depend on the noise and hence are incorporated from Appendix C.3 of [7]. It can be verified that the proof does not change for the noisy case. We omit the derivation here to avoid redundancy.

Lemma 9. *Under the conditions of Theorem 2, w.p. greater that* $1 - 1/n^3$

$$\|C_N^j \Sigma_N V_N^{(j)}\|_2 \leq N_{max}\mu_1\sqrt{km}(1+\delta_{2k})$$

We prove the above lemma at the end of this section. Now, from Equation 22, we have:

$$\left(\widehat{V}^{(t+1)}\right)^{(j)} = D^j \Sigma^* V^{*(j)} - (B^j)^{-1}\left(B^jD^j - C^j\right)\Sigma^* V^{*(j)} + (B^j)^{-1}C_N^j\Sigma_N V_N^{*(j)}$$

Thus,

$$\begin{aligned} \left\|\left(V^{(t+1)}\right)^{(j)}\right\|_2 \leq\ & \|(R^{(t+1)})^{-1}\|_2 \Big[\left(\|D^j\|_2 + \|(B^j)^{-1}\|_2(\|B^jD^j\|_2 + \|C^j\|_2)\right)\|\Sigma^*\|\|V^{*(j)}\|_2 \\ +\ & \|(B^j)^{-1}\|_2\|C_N^j\Sigma_N V_N^{*(j)}\|_2\Big] \end{aligned} \tag{26}$$

Using equations from 25, Lemma 9 and 7, and $\delta_{2k} \leq \frac{1}{C}$, $C > 1$ we have the following:

$$\begin{aligned} \left\|\left(V^{(t+1)}\right)^{(j)}\right\|_2 \leq\ & \|(R^{(t+1)})^{-1}\|_2 \left[\frac{\sigma_1^*\mu\sqrt{k}}{\sqrt{n}}\left(1 + \frac{(2(1+\delta_{2k}))}{1-\delta_{2k}}\right) + \frac{N_{max}\mu_1\sqrt{km}(1+\delta_{2k})}{1-\delta_{2k}}\right] \\ =\ & \|(R^{(t+1)})^{-1}\|_2\left[\frac{4\sigma_1^*\mu\sqrt{k}}{\sqrt{n}} + 2N_{max}\mu_1\sqrt{km}\right] \end{aligned} \tag{27}$$

We now use that for, $\mu_1 = \frac{32\mu\sigma_1^*\sqrt{k}}{\sigma_k^*}$ and further, $N_{max} \leq C_3\frac{\sigma_k^*}{n\sqrt{k}}$. Choosing C_3 appropriately, we have $N_{max}\mu_1\sqrt{km} \leq \frac{2\sigma_1^*\mu\sqrt{k}}{\sqrt{n}}$. Finally, using Lemmata 5 and 6, the fact that dist$(U^*, U^t) \leq$ dist$(U^*, U^0) \leq 0.5$ and using the conditions on $\|N^\Omega\|_2$ from Theorem 2 , we have, $\|(R^{(t+1)})^{-1}\|_2 \leq \frac{4}{\sigma_k^*}$. Using these we have:

$$\left\|\left(V^{(t+1)}\right)^{(j)}\right\|_2 \leq \|(R^{(t+1)})^{-1}\|_2\frac{8\sigma_1^*\mu\sqrt{k}}{\sqrt{n}} \leq \frac{32\sigma_1^*\mu\sqrt{k}}{\sigma_k^*\sqrt{n}} \tag{28}$$

Thus we have, $\mu(V^{(t+1)}) \leq \frac{32\sigma_1^*\mu}{\sigma_k^*} \leq \frac{32\sigma_1^*\mu\sqrt{k}}{\sigma_k^*}$.

Proof of Lemma 9. $\|C_N^j \Sigma_N V_N^{(j)}\|_2 = \max_{x:\|x\|_2=1} x^\dagger C_N^j \Sigma_N V_N^{(j)}$. Given any x such that $\|x\|_2 = 1$, then

$$x^\dagger C_N^j \Sigma_N V_N^{(j)} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} x^\dagger U^{t(i)} U_N^{(i)\dagger} \Sigma_N V_N^{(j)} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} x^\dagger U^{t(i)} N_{ij} \tag{29}$$

We define $\delta_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$, $Z_i = \frac{1}{p}\delta_{ij} x^\dagger U^{t(i)} N_{ij}$ and $Z = \sum_{i=1}^m Z_i$

$$E[Z] = \sum_{i=1}^m E[Z_i] = \sum_{i=1}^m x^\dagger U^{t(i)} N_{ij} \le N_{max} \sum_{i=1}^m \|U^{t(i)}\|_2 = N_{max}\mu_1\sqrt{km} \tag{30}$$

$$\begin{aligned} var(Z) = & \sum_{i=1}^m E[Z_i^2] - (E[Z_i])^2 = \frac{1-p}{p} \sum_{i=1}^m \left(x^\dagger U^{t(i)}\right)^2 N_{ij}^2 \\ \le & \frac{1}{p} N_{max}^2 \sum_{i=1}^m \|U^{t(i)}\|_2^2 = \frac{1}{p} N_{max}^2 \|U^t\|_F^2 = \frac{N_{max}^2 k}{p} \end{aligned} \tag{31}$$

$$\max_i Z_i = \frac{1}{p} \max_i x^t U^{t(i)} N_{ij} \le \frac{N_{max}\mu_1\sqrt{k}}{p\sqrt{m}} \tag{32}$$

From Equations 29, 30, 31, 32 and using Bernstein's inequality in Equation 19, we have the following:

$$P\left(Z \ge N_{max}\mu_1\sqrt{km}(1+\delta_{2k})\right) \le \exp\left(\frac{-\delta_{2k}^2 N_{max}^2 \mu_1^2 km/2}{\frac{N_{max}^2 k}{p} + \frac{N_{max}^2 \mu_1^2 k \delta_{2k}}{3p}}\right) = \exp\left(\frac{-\delta_{2k}^2 \mu_1^2 mp}{2\left(1 + \frac{\mu_1^2 \delta_{2k}}{3}\right)}\right)$$

From the conditions of Theorem 2, $\delta_{2k} \le \frac{\sigma_1^*}{C\sigma_k^*}$, $p > 12\frac{\log n}{m\delta_{2k}}$, using $(1 + \mu_1^2\delta_{2k}/3) \le (1 + \mu_1^2) \le 2\mu_1^2$, we have:

$$P\left(Z \ge N_{max}\mu_1\sqrt{km}(1+\delta_{2k})\right) \le \exp\left(\frac{-12\mu_1^2 \log n}{4\mu_1}\right) = \frac{1}{n^3}.$$

Thus, we have with probability grater that $1 - 1/n^3$, $\forall\ x : \|x\|_2 = 1$, including the maximizing x, we have $x^\dagger C_N^j \Sigma_N V_N^{(j)} \le N_{max}\mu_1\sqrt{km}(1+\delta_{2k})$. Thus, $\|C_N^j \Sigma_N V_N^{(j)}\|_2 \le N_{max}\mu_1\sqrt{km}(1+\delta_{2k})$. □