Robust Alternating Minimization for Matrix Completion

Andrew Singh
andrewsi@andrew.cmu.edu

Carnegie Mellon University
Pittsburgh, PA 15213

November 8, 2020

Abstract

We consider the problem of robust matrix completion, where the objective is to recover a partially observed matrix of low rank that has been corrupted by an adversary. We present a matrix completion method that is robust to corruption from an arbitrary outlier distribution. Our method performs robust linear regression in an alternating minimization scheme, and it is able to exactly recover the underlying matrix even when the corruptions are not independent of the observed entries.

1 Introduction

Background

Matrix completion is the problem of reconstructing a matrix for which only a fraction of the entries are observed. With no prior assumptions on the matrix, this problem is unsolvable since the missing entries could be arbitrary values. However if we assume that the matrix has some underlying structure; that is, the matrix is of low rank, it becomes possible to recover the matrix exactly even when only a small fraction of its entries are observed.

The problem can be formulated as follows. Let $L^* \in \mathbb{R}^{m \times n}$ be a matrix of rank r. Suppose we only observe a p fraction of entries in L^*. The objective of the problem is, given these sparse observations, to find factors $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ such that the reconstruction error $\|L^* - UV^T\|_F$ is minimized.

While there are several established methods for solving this problem when the observed entries are unperturbed, these methods can easily break down when a small fraction of malicious noise has been added to the data. We consider the problem of *robust* matrix completion, where a fraction of our observations have been corrupted by an adversary. The objective of retrieving the underlying matrix remains the same as the clean setting, but the problem has become more difficult due to the sparse corruptions. Indeed, the standard techniques used for matrix completion break down under these noisy conditions.

Matrix completion has several real-world applications, a popular one of which is recommender systems [1]. Consider a movie recommendation service with n users and m movies. We can view the ratings as a $n \times m$ ratings matrix R where R_{ij} is user i's rating for movie j. Since each user has only seen a small fraction of all movies, this matrix is very sparse. The goal of the recommender

system is to predict users' ratings for unseen movies. We can formulate this as a matrix completion problem: assume some low rank r for the ratings matrix R, and then predict the missing entries in the matrix.

Assuming that the ratings matrix is of low rank comes from the idea that each user and movie can be described by r features, where r is the rank. Each user and movie is represented as a length r vector in the factor matrices U, V found by matrix completion. Because the predicted rating of user u on movie v is simply the dot product of their vectors $u \cdot v$, each value in the vector corresponds to a feature, and the magnitude of that value indicating how strongly the user prefers that feature or the movie exhibits that feature. We can think of these features as things like action, comedy, sci-fi, romance, etc, that fully describe a movie or a user's preferences. Note that none of these features are given to the model; all that is given is the ratings data and the number of features to learn, and through matrix completion the model learns the features from the data.

Related Work

Because matrix completion is a non-convex optimization problem, most of the methods for solving it involve gradient descent or alternating minimization. Simon Funk devised a stochastic gradient descent method in 2006 [2] for the Netflix Prize, a competition held by Netflix that challenged participants to predict movie ratings for Netflix's users. This became a very popular method for matrix factorization in practice due to its simplicity and successful results, and it is still used by recommender systems today such as the MovieLens website. The method simply runs stochastic gradient descent on the ratings data, optimizing each feature one at a time until all have been optimized and then repeating the process until convergence.

Another popular method for matrix completion is alternating minimization. Rather than a specific algorithm, alternating minimization is a general framework for matrix completion, and has been applied to other optimization problems as well [3]. Recall that the objective of matrix completion is to find factors U, V such that their product UV^T best approximates the true matrix. The framework consists of two stages: fixing V and optimizing for U, and fixing U and optimizing for V. The algorithm alternates between these two stages until convergence. A key property of the matrix completion problem is that while the overall problem is non-convex, if we fix one of the factors, then the problem actually becomes linear, and we can directly solve for the other factor. When we directly solve for each factor in an alternating minimization scheme, the method is called alternating least squares, and has seen success in practice. Zhou et al. proposed a parallel implementation of the alternating least squares method for the Netflix Prize in 2008 [4]. Funk's SGD and the alternating least squares method achieve very similar performance in practice; the main difference between the two methods is in efficiency, and depends on the specific application and hardware used [1].

The Netflix Prize inspired a large increase in matrix completion research, however it was not until several years later that the *robust* matrix completion problem began to gain traction. One such work in the robust field is by Cherapanamjeri et al., who propose a method based on projected gradient descent called PG-RMC [5]. Their model attempts to explicitly determine the clean component L and the noisy component S of the observed matrix. The algorithm updates S using a hard thresholding function, then performs a gradient update on L and projects it into the subspace of rank r matrices. They showed that their method can exactly recover the true matrix when the number of observed entries is $O(r^2 n \log n)$ and the fraction of corruptions is $O\big(\frac{1}{\mu^2 r}\big)$, where μ is the incoherence of the matrix (see Definition 1).

A critical assumption that the PG-RMC makes is that the corruption matrix S is independent of the

observed entries Ω. This assumption has significant implications because it requires an adversary to choose their corruptions with no knowledge of the observations; it does not account for the situation where the adversary also has access to the observations and can choose their corruptions using this information. We note that our approach does not require this assumption. We show that with an additional factor of $\frac{1}{r}$ in complexity of the corruption fraction ($\epsilon = O\big(\frac{1}{\mu^2 r^2}\big)$ for our approach vs. $\epsilon = O\big(\frac{1}{\mu^2 r}\big)$ for PG-RMC), we are able to exactly recover the underlying matrix L even when the corruptions are not independent of the observations.

Contributions

Our chief contribution in this paper is showing that exact recovery of the underlying matrix is possible even when the corruptions are not independent of the observations. Our algorithm solves this problem with a $O\big(\frac{1}{\mu^2 r^2}\big)$ complexity on the fraction of corruptions, incurring an additional factor of $\frac{1}{r}$ over other robust methods that rely on the independence assumption (see Related Work), which can solve the problem with a corruption complexity of $O\big(\frac{1}{\mu^2 r}\big)$. While our approach employs the Robust Gradient Descent technique proposed by Prasad et al. [6] for robust regression, our algorithm is general and can work with any such robust regressor. In the future, we plan to study this independence assumption more deeply and analyze how methods such as PG-RMC rely on this assumption in their approach.

Paper Organization

In Section 2, we define the problem and our assumptions. In Section 3, we present our algorithm for robust matrix completion and our main theoretical result, and in Section 4, we give an outline of our analysis. In Section 5, we discuss the remaining work to be done. Our full analysis can be found in Appendix A.

2 Problem Setup

We first provide a definition for the matrix incoherence property. Requiring the underlying matrix to be incoherent ensures that its singular vectors are not too sparse; that is, their entries are all of relatively similar magnitude [7]. Incoherence is a standard assumption in the matrix completion literature and is critical to the analysis of these methods, even in the non-robust case.

Definition 1. *A matrix $L = U\Sigma V^T \in \mathbb{R}^{m\times n}$ with rank r is incoherent with parameter μ if*

$$\|U_i\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{m}} \forall i \in [m], \|V_j\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}} \forall j \in [n]$$

where U_i and V_j denote the i^{th} and j^{th} rows of U and V respectively.

We now define the robust matrix completion problem. Let $L^* = U^*\Sigma^*(V^*)^T \in \mathbb{R}^{m\times n}$ be a rank r μ-incoherent matrix with $m \leq n$ and with non-zero singular values $\sigma_1^* \geq \sigma_2^* \geq \dots \sigma_r^*$. Assume the observed entries Ω are chosen via a Bernoulli sampling model such that $Pr[(i,j) \in \Omega] = p$ for each entry (i,j). Let Ω denote the observed entries of L^*, and let $P_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ denote the sampling operator such that

$$P_\Omega(L^*)_{i,j} = \begin{cases} L^*_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Furthermore, assume that an ϵ fraction of entries of $P_\Omega(L^*)$ have been corrupted by an arbitrary corruption distribution. Specifically, let $P_\Omega(M) = P_\Omega(L^*) + S$ where S is a matrix with ϵ fraction

of its entries in Ω set to arbitrary corruption values and the remaining entries set to 0. This setup follows the Huber contamination model. Critically, note that the corruption matrix S need not be independent of the observed entries Ω. The observed matrix is given by $P_\Omega(M)$, and the objective is to solve the following optimization problem:

$$\min_{U,V} \left\| P_\Omega(L^*) - P_\Omega(UV^T) \right\|_F$$

In summary, we make three standard assumptions for this problem:

Assumption 1. ***Rank and incoherence of*** L^****:*** *L^* is a rank r, μ-incoherent matrix.*

Assumption 2. ***Sampling of entries*** Ω***:*** *The observed entries Ω are chosen via a Bernoulli sampling model such that $Pr[(i,j) \in \Omega] = p$ for each entry (i,j) and each row and column of $P_\Omega(L^*)$ has at least a p fraction of entries.*

Assumption 3. ***Sparsity of corruptions*** S***:*** *With $P_\Omega(M) = P_\Omega(L^*) + S$, at most ϵ fraction of entries in Ω are corrupted in each row and column of $P_\Omega(M)$.*

We reiterate that we do not make the independence assumption; in our model, the corruptions are made by an adversary with full knowledge of the observations Ω and matrix $P_\Omega(L^*)$.

3 Algorithm and Main Result

Our approach for solving the robust matrix completion problem is based on alternating minimization, but instead of performing ordinary linear regression in each step, we perform robust regression. As our robust regressor, we employ the Robust Gradient Descent algorithm proposed by Prasad et al. [6] using the Huber Gradient Estimator. The key idea of Robust Gradient Descent is that when performing gradient descent on noisy data, instead of modifying the loss function, we can modify the gradients. In batch gradient descent, a gradient update is simply the mean of the sample gradients. If we simply take the mean of the gradients, a single outlier can have an arbitrarily large effect on the gradient mean, but through robust mean estimation, we can estimate the mean of the true sample gradients. While we use Robust Gradient Descent to perform robust regression, we note that our algorithm can work with any robust regressor. We present the algorithm in its general form as Algorithm 1.

For ease of analysis, we assume that the algorithm receives fresh samples on each iteration, necessitating the partitioning step in line 2. Our initialization comprises two steps. We first hard-threshold and scale the observed matrix $P_\Omega(M)$ and then perform a rank r SVD on the result to obtain our initial iterate $\widehat{U}^0$. The second step performs another hard-threshold on $\widehat{U}^0$ to ensure that it is incoherent, a necessary requirement of our analysis. We then perform optimization over the iterates $\widehat{U}$ and $\widehat{V}$ in the alternating minimization scheme, employing a robust regressor in each phase to

solve the regression problem.

Algorithm 1: Robust Alternating Minimization for Matrix Completion

function `RobustAltMinComplete(`*Robust regressor* `RobustReg`*, observed matrix* $P_\Omega(M) \in \mathbb{R}^{m\times n}$*, rank* r*, corruption fraction* ϵ*, estimate of first singular value* σ_1`):`
  Partition Ω into $2\tau+1$ subsets $\Omega_0, \ldots, \Omega_{2\tau}$ uniformly and randomly
  $\widehat{L} = \frac{1}{p}\mathcal{HT}(P_\Omega(M))$ where $\mathcal{HT}$ sets all elements with magnutude greater than $\frac{4\mu^2 r}{m}\sigma_1$ to zero
  $\widehat{U}^0 = SVD(\widehat{L}, r)$ (i.e., the top r left singular vectors of $\widehat{L}$)
  Set all elements of $\widehat{U}^0$ with magnitude greater than $\frac{2\mu\sqrt{r}}{\sqrt{m}}$ to zero and orthonormalize the columns of $\widehat{U}^0$
  for $t = 0, 1, \ldots, \tau-1$ **do**
    for $j = 0, 1, \ldots, n-1$ **do**
      $\widehat{V}_j^{t+1} = \texttt{RobustReg}\big(\{(\widehat{U}_i^t, P_{\Omega_{t+1}}(M)_{ij})\}_{i=1}^m\big)$
    end for
    for $i = 0, 1, \ldots, m-1$ **do**
      $\widehat{U}_i^{t+1} = \texttt{RobustReg}\big(\{(\widehat{V}_i^t, P_{\Omega_{\tau+t+1}}(M)_{ij})\}_{j=1}^n\big)$
    end for
  end for
  return $\widehat{U}^\tau, \widehat{V}^\tau$

We now present our main result for Algorithm 1 followed by an outline of our analysis.

Theorem 1. *Let assumptions 1, 2, and 3 hold on L^*, Ω, and S respectively. Let the sampling probability p satisfy*

$$p > C\frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2\mu^4 r^{2.5}\log n \log\frac{r\sigma_1^*}{\epsilon'}}{m\delta_{2r}^2}$$

where $\delta_{2r} \le \frac{\sigma_r^*}{12r\sigma_1^*}$ and $C > 0$ is a global constant. Let the corruption fraction ϵ satisfy

$$\epsilon \le \frac{(\sigma_r^*)^2}{58^2 C_3^2 (\sigma_1^*)^4 \mu^2 r^2}$$

where $C_3 > 0$ is a global constant. Then with high probability, for $\tau = C'\log\frac{\frac{13}{11}\sigma_1^*\sqrt{r}}{\epsilon'}$, Algorithm 1 with Robust Gradient Descent using the Huber Gradient Estimator outputs iterates $\widehat{U}^\tau$ and $\widehat{V}^\tau$ that satisfy $\|L^* - \widehat{U}^\tau(\widehat{V}^\tau)^T\|_F \le \epsilon'$.

4 Analysis

Since our algorithm is essentially the original alternating minimization algorithm with robust regression substituted for ordinary least squares, our analysis is a modified version of the original analysis [8]. Our key addition is Lemma 13, where we analyze the error due to corruptions. The corruptions also necessitate a different initialization method than the original algorithm; we employ a hard-thresholding strategy inspired by the PG-RMC algorithm [5].

The proof of Theorem 1 has two components. We first establish an initial distance between the subspaces spanned by the initial iterate U^0 and the optimal factor U^* (Lemma 2). We then show that given this initialization, in each iteration of the algorithm, the distance between the subspaces

spanned by V^t and V^*, decreases geometrically with the iteration t (Theorem 3). This result symmetrically follows for U^t and U^* as well. For our analysis, we use the following definition of distance between subspaces.

Definition 2. *(Definition 4.1 of [8]) Given two matrices $\widehat{U}, \widehat{W} \in \mathbb{R}^{m \times r}$, the (principle angle) distance between the subspaces spanned by the columns of $\widehat{U}$ and $\widehat{W}$ is given by*

$$dist(\widehat{U}, \widehat{W}) = \|U_\perp^T W\|_2 = \|W_\perp^T U\|_2$$

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 any orthonormal bases of the perpendicular spaces $Span(U)^\perp$ and $Span(W)^\perp$ respectively.

Note:

(a) The distance depends only on the spaces spanned by the columns of $\widehat{U}, \widehat{W}$.

(b) If the ranks of $\widehat{U}$ and $\widehat{W}$ (i.e. the dimensions of their spans) are not equal, then $dist(\widehat{U}, \widehat{W}) = 1$

(c) $dist(\widehat{U}, \widehat{W}) = 0$ if and only if they span the same subspace of $\mathbb{R}^m$.

In Lemma 4, we show that iterates that are close to the optimal factors under this definition of distance are also close to the underlying matrix L^* under our original objective $\|L^* - \widehat{U}(\widehat{V})^T\|_F$. We now present the two main components of the analysis, the initialization and the induction argument.

Lemma 2. *Let U^0 be the iterate obtained by the initialization of Algorithm 1. Then with high probability, we have that $dist(U^0, U^*) \leq \frac{1}{2}$ and that U^0 is $K\mu$-incoherent for $K = \frac{13\sigma_1^*}{\sigma_r^*}$.*

This result guarantees a close enough distance between our initial iterate and the optimal, and it establishes the incoherence of the iterates, a necessary property for the induction step. The proof can be found in Appendix A. We now move to the induction argument.

Theorem 3. *Let $\widehat{U}^t$, $\widehat{V}^t$ denote the t^{th} iterates of Algorithm 1. Then with high probability, $\forall 1 \leq t \leq \tau$, the $(t+1)^{th}$ iterates $\widehat{U}^{t+1}$, $\widehat{V}^{t+1}$ satisfy the following properties:*

$$dist(\widehat{V}^{t+1}, V^*) \leq \frac{1}{2} dist(\widehat{U}^t, U^*)$$

$$dist(\widehat{U}^{t+1}, U^*) \leq \frac{1}{2} dist(\widehat{V}^{t+1}, V^*)$$

This result establishes the geometric sequence of distances and shows the convergence of the iterates to the optimal U^*, V^*. We give a brief outline of our analysis for this convergence result, and defer the full proof to Appendix A.

Observe that in the clean setting, where $\epsilon = 0$ and we have full access to $P_\Omega(L^*)$, the alternating updates are given by

$$\widehat{V}^{t+1} = \arg\min_{\widehat{V}} \left\| P_\Omega(L^*) - P_\Omega(\widehat{U}^t \widehat{V}^T) \right\|_F$$

$$\widehat{U}^{t+1} = \arg\min_{\widehat{U}} \left\| P_\Omega(L^*) - P_\Omega(\widehat{U} (\widehat{V}^{t+1})^T) \right\|_F$$

This optimization problem is linear, and the update can be directly written as a closed-form expression. This is what the standard alternating minimization algorithm does [8]. When the corruption fraction $\epsilon > 0$, we substitute ordinary linear regression for *robust* linear regression using Robust Gradient Descent [6]. We can view the result of this update as the original, closed-form update plus an additive error:

$$\widehat{V}^{t+1} = \arg\min_{\widehat{V}} \left\| P_\Omega(L^*) - P_\Omega(\widehat{U}^t \widehat{V}^T) \right\|_F - A$$

where A is the additive error from the robust regression. We proceed with our argument by analyzing the update in this form. Note that even though we do not directly observe $P_\Omega(L^*)$, isolating the additive error arising from corruption in this way allows us to bound each component separately. We follow the analysis of the original alternating minimization algorithm [8] while also taking into account the additional error term A due to the corruptions. We analyze this additonal source of error in Lemma 13, using the guarantees of the Robust Gradient Descent method run with the Huber Gradient Estimator (Theorem 18). Critically, the additional error arising from corruptions is proportional to the distance between the current and optimal iterates. This means that the corruption error contracts every iteration, allowing us to exactly recover the underlying matrix. The complete proof can be found in Appendix A.

5 Future Work

We have several remaining items planned for the future. We will evaluate our algorithm experimentally and compare it against other robust matrix completion methods such as PG-RMC [5]. In addition, we will further study the independence assumption that requires the corruptions to be independent of the observations. We will explore how methods such as PG-RMC exploit this assumption in their analysis. In addition, we will explore why exact recovery of the underlying matrix is possible under conditions with arbitrary corruptions and what properties and assumptions are most critical to this result.

Beyond these items, we would like to broaden our focus and additionally study the problem of "noisy" matrix completion; that is, the case where the true low-rank matrix has some inherent noise as opposed to being corrupted by an adversary. We would like to analyze the performance of our algorithm in this setting and find out if it matches information-theoretic lower bounds.

6 Acknowledgments

I would like to thank Adarsh Prasad and Professor Pradeep Ravikumar for giving me the opportunity to work on this project and for advising me through the research process. Their support and guidance was an essential part of making this project a reality.

References

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

[2] Simon Funk. *Netflix Update: Try This at Home*, 2006. `https://sifter.org/~simon/journal/20061211.html`.

[3] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *ArXiv*, abs/1712.07897, 2017.

[4] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. pages 337–348, 06 2008.

[5] Yeshwanth Cherapanamjeri, K. Gupta, and P. Jain. Nearly optimal robust matrix completion. In *ICML*, 2017.

[6] A. Prasad, Arun Sai Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *ArXiv*, abs/1802.06485, 2018.

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

[8] P. Jain, Praneeth Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC '13*, 2013.

A Appendix

Proof of Theorem 1

Proof. By Theorem 3, after $O(\log(\frac{1}{\epsilon'}))$ iterations, we have

$$dist(U^t, U^*) \leq \epsilon'$$

$$dist(V^{t+1}, V^*) \leq \epsilon'$$

By Lemma 4, we have that $\|L^* - U^t(\widehat{V}^{t+1})^T\|_F \leq \frac{13}{11}\sigma_1^*\sqrt{r}\epsilon'$. The result now follows by setting $\epsilon'' = \frac{13}{11}\sigma_1^*\sqrt{r}\epsilon'$. □

Lemma 4. *Suppose that after t iterations of Algorithm 1, we have that $dist(U^t, U^*) \leq d$. Then we have that*

$$\|L^* - U^t(\widehat{V}^{t+1})^T\|_F \leq \frac{13}{11}\sigma_1^*\sqrt{r}d$$

Proof. By the update equation in Theorem 3, the residual after the t^{th} iteration is

$$\begin{aligned} L^* - U^t(\widehat{V}^{t+1})^T &= L^* - U^t\Big(V^*\Sigma^*(U^*)^TU^t - F - A\Big)^T \\ &= \Big(I - U^t(U^t)^T\Big)L^* + U^tF^T + U^tA^T \end{aligned}$$

Therefore we have,

$$\begin{aligned} \|L^* - U^t(\widehat{V}^{t+1})^T\|_F &\leq \|(I - U^t(U^t)^T)L^*\|_F + \|F\|_F + \|A\|_F \\ &\leq \|(I - U^t(U^t)^T)U^*\Sigma^*V^*\|_F + \|F\|_F + \|A\|_F \end{aligned}$$

Since V^* has orthonormal columns, $\|V^*\|_F = \sqrt{r}$. We have

$$\begin{aligned} \|L^* - U^t(\widehat{V}^{t+1})^T\|_F &\leq \sqrt{r}\|(I - U^t(U^t)^T)U^*\Sigma^*\|_2 + \|F\|_F + \|A\|_F \\ &\leq \sqrt{r}\|(U_\perp^t)^TU^*\Sigma^*\|_2 + \|F\|_F + \|A\|_F \\ &\leq \sqrt{r}(\sigma_1^*)dist(U^t, U^*) + \|F\|_F + \|A\|_F \end{aligned}$$

By Lemma 11 and Lemma 13, and since $dist(U^t, U^*) \leq d$, we have

$$\|L^* - U^t(\widehat{V}^{t+1})^T\|_F \leq \sqrt{r}(\sigma_1^*)d + \left(\frac{\delta_{2r} r\sigma_r^*}{1-\delta_{2r}}\right)d + \left(\left(\frac{13}{11}\right)^{1.5} C_3\sigma_1^*\sqrt{\epsilon r \log r}\right)d$$

For $\delta_{2r} \leq \frac{\sigma_r^*}{12r\sigma_1^*}$ and $\epsilon \leq \frac{(\sigma_r^*)^2}{58^2 C_3^2(\sigma_1^*)^4\mu^2 r^2}$, we have

$$\begin{aligned}\|L^* - U^t(\widehat{V}^{t+1})^T\|_F &\leq \sqrt{r}(\sigma_1^*)d + \frac{(\sigma_r^*)^2}{11\sigma_1^*}d + \frac{\sigma_r^*}{11\sigma_1^*}d \\ &\leq \left(\sigma_1^*\sqrt{r} + \frac{2\sigma_r^*}{11}\right)d \\ &\leq \frac{13}{11}\sigma_1^*\sqrt{r}d\end{aligned}$$

□

Proof of Lemma 2

Proof. By Lemma 5, after step 1 of the initialization we obtain a matrix $\widehat{U}$ such that $dist(\widehat{U}, U^*) \leq \frac{1}{64r}$. Then by applying step 2 of the initialization to $\widehat{U}$, by Lemma 6 we obtain a matrix U^0 such that $dist(U^0, U^*) \leq \frac{1}{2}$ and U^0 is 4μ-incoherent, therefore U^0 is also $K\mu$-incoherent for $K = \frac{13\sigma_1^*}{\sigma_r^*}$. □

Lemma 5. *(Initialization step 1) Let $\widehat{U}$ be the iterate obtained after line 4 of Algorithm 1. Then with high probability, we have,*

$$dist(\widehat{U}, U^*) \leq \frac{1}{64r}$$

Proof. We present an argument similar to the initialization of the PG-RMC algorithm in [5]. Let $\widehat{L} = \frac{1}{p}\mathcal{HT}(P_\Omega(M))$ where $\mathcal{HT}$ is the hard thresholding operator such that for each $(i,j) \in \Omega$,

$$\mathcal{HT}(P_\Omega(M)_{ij}) = \begin{cases} P_\Omega(M)_{ij} & \text{if } |P_\Omega(M)_{ij}| \leq \gamma \\ 0 & \text{otherwise} \end{cases}$$

where γ denotes the threshold value. After step 1 of the initialization, we have that $\widehat{U} = SVD(\widehat{L}, r)$, that is, the top r singular vectors of $\widehat{L}$. By Lemma 7, we have that

$$\|\widehat{L} - L^*\|_2 \leq \frac{1}{r}\left(\frac{1}{140C_3^2\sigma_1^*} + \frac{180\sigma_r^*}{\sqrt{C}}\right)$$

Let $\widehat{U}\Sigma V^T$ be the top r singular components of $\widehat{L}$. Then we also have

$$\begin{aligned}\|\widehat{L} - L^*\|_2^2 &= \|L^* - \widehat{L}\|_2^2 \\ &= \|U^*\Sigma^*(V^*)^T - \widehat{U}\Sigma V^T\|_2^2 \\ &= \|U^*\Sigma^*(V^*)^T - \widehat{U}(\widehat{U})^T U^*\Sigma^*(V^*)^T + \widehat{U}(\widehat{U})^T U^*\Sigma^*(V^*)^T - \widehat{U}\Sigma V^T\|_2^2 \\ &= \left\|\left(I - \widehat{U}(\widehat{U})^T\right)U^*\Sigma^*(V^*)^T + \widehat{U}\left((\widehat{U})^T U^*\Sigma^*(V^*)^T - \Sigma V^T\right)\right\|_2^2 \\ &\geq \left\|\left(I - \widehat{U}(\widehat{U})^T\right)U^*\Sigma^*(V^*)^T\right\|_2^2 \\ &= \|(\widehat{U}_\perp)^T U^*\Sigma^*(V^*)^T\|_2^2 \\ &\geq (\sigma_r^*)^2\|(\widehat{U}_\perp)^T U^*\|_2^2\end{aligned}$$

where the first inequality follows from the fact that the column space of the first two terms is $\widehat{U}_\perp$ while the column space of the last two terms is $\widehat{U}$. By the above two inequalities, we have

$$\|(\widehat{U}_\perp)^T U^*\|_2 \le \frac{1}{\sigma_r^*}\|\widehat{L} - L^*\|_2 \le \frac{1}{r}\left(\frac{1}{140C_3^2\sigma_1^*\sigma_r^*} + \frac{180}{\sqrt{C}}\right) \le \frac{1}{64r}$$

for large enough constants C and C_3. □

Lemma 6. *(Initialization step 2, Lemma C.2 of [8]) Let U^* be a μ-incoherent matrix and U be an orthonormal column matrix such that $dist(U, U^*) \le \frac{1}{64r}$. Let U^c be the matrix obtained from U by setting all entries with magnitude greater than $\frac{2\mu\sqrt{r}}{m}$ to zero. Let $\widetilde{U}$ be an orthonormal basis of U^c. Then we have the following two claims:*

- *$dist(\widetilde{U}, U^*) \le \frac{1}{2}$*
- *$\widetilde{U}$ is 4μ-incoherent*

Proof. Let $d = dist(U, U^*)$. We have that for every i, $\exists X_i \in \text{Span}(U^*), \|X_i\|_2 = 1$ such that $\langle u_i, X_i\rangle \ge \sqrt{1-d^2}$. In addition, since $X_i \in \text{Span}(U^*)$, we have that X_i is μ-incoherent. Therefore $\|X_i\|_\infty \le \frac{\mu\sqrt{r}}{\sqrt{m}}$. Let U_i^c be the vector obtained by setting all elements of U_i with magnitude greater than $\frac{2\mu\sqrt{r}}{m}$ to zero, and let $U_i^{\bar{c}} = U_i - U_i^c$. Observe that for entry U_{ij}, if $|U_{ij}| > \frac{2\mu\sqrt{r}}{\sqrt{m}}$, then we have that $|U_{ij}^c - X_{ij}| = |X_{ij}| \le \frac{\mu\sqrt{r}}{\sqrt{m}} \le |U_{ij} = X_{ij}|$. Therefore we have

$$\|U_i^c\|_2 - X_i \le \|U_i - X_i\|_2 = \sqrt{\|U_i\|_2^2 + \|X_i\|_2^2 - 2\langle U_i, X_i\rangle} \le d\sqrt{2}$$

This also implies the following:

$$\|U_i^c\|_2 \ge \|X_i\|_2 - d\sqrt{2} = 1 - d\sqrt{2}$$

and

$$\|U_i^{\bar{c}}\|_2 \le \sqrt{1 - \|U_i^c\|_2^2} \le \sqrt{2d(\sqrt{2}-d)} \le 2\sqrt{d} \quad \text{for } d < \frac{1}{\sqrt{2}}$$

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

$$\begin{aligned}
\|(u_\perp^*)^T\widetilde{U}\|_2 &= \|(u_\perp^*)^T U^c\Lambda\|_2 \\
&\le \|(u_\perp^*)^T U^c\|_2\|\Lambda\|_2 \\
&\le \left(\|(u_\perp^*)^T U\|_2 + \|(u_\perp^*)^T U^{\bar{c}}\|_2\right)\|\Lambda\|_2 \\
&\le (d + \|U^{\bar{c}}\|_2)\|\Lambda\|_2 \\
&\le (d + \|U^{\bar{c}}\|_F)\|\Lambda\|_2 \\
&\le (d + 2\sqrt{rd})\|\Lambda\|_2 \\
&\le 3\sqrt{rd}\|\Lambda\|_2
\end{aligned}$$

We now bound $\|\Lambda\|_2$. We have

$$\|\Lambda\|_2^2 = \frac{1}{\sigma_{\min}(\Lambda^{-1})^2} = \frac{1}{\sigma_{\min}(\widetilde{U}\Lambda^{-1})^2} = \frac{1}{\sigma_{\min}(U^c)^2} \le \frac{1}{1 - \|U^{\bar{c}}\|_2^2} \le \frac{1}{1-4rd} \le \frac{4}{3}$$

where the last step follows from the fact that $d \leq \frac{1}{64r} \leq \frac{1}{16r}$. Therefore we have

$$\|(u_\perp^*)^T \widetilde{U}\|_2 \leq 3\sqrt{rd}\left(\frac{4}{3}\right) = 4\sqrt{rd} \leq \frac{1}{2}$$

This proves the first claim of the lemma. We now show the incoherence of $\widetilde{U}$. We have

$$\begin{aligned}
\mu(\widetilde{U}) &= \sqrt{\frac{m}{r}} \max_i \|e_i^T \widetilde{U}\|_2 \\
&\leq \sqrt{\frac{m}{r}} \max_i \|e_i^T U^c \Lambda\|_2 \\
&\leq \sqrt{\frac{m}{r}} \max_i \|e_i^T U^c\|_2 \|\Lambda\|_2 \\
&\leq 4\mu\sqrt{r}
\end{aligned}$$

□

Lemma 7. *Let $\widehat{L} = \frac{1}{p}\mathcal{H}\mathcal{T}(P_\Omega(M))$ as defined in Lemma 5, and let $H = \widehat{L} - L^*$. Then we have that*

$$\|H\|_2 \leq \frac{\sigma_r^*}{r}\left(\frac{1}{140C_3^2\sigma_1^*\sigma_r^*} + \frac{180}{\sqrt{C}}\right)$$

Proof. Observe that by our definition of $\widehat{L}$ and H, we can write $H = E_1 + E_2$ for $E_1 = \mathcal{H}\mathcal{T}(M) - L^*$ and $E_2 = -\mathcal{H}\mathcal{T}(M) - \frac{1}{p}P_\Omega(-\mathcal{H}\mathcal{T}(M))$. Indeed, we have

$$\begin{aligned}
E_1 + E_2 &= \mathcal{H}\mathcal{T}(M) - L^* - \mathcal{H}\mathcal{T}(M) - \frac{1}{p}P_\Omega(-\mathcal{H}\mathcal{T}(M)) \\
&= -L^* - \frac{1}{p}P_\Omega(-\mathcal{H}\mathcal{T}(M)) \\
&= \frac{1}{p}\mathcal{H}\mathcal{T}(P_\Omega(M)) - L^* \\
&= \widehat{L} - L^* \\
&= H
\end{aligned}$$

By Lemma 9, we have that $\frac{1}{\beta}E_2$ satisfies Definition 3 for $\beta = \frac{2\sqrt{n}}{\sqrt{p}}\|vec(-\mathcal{H}\mathcal{T}(M))\|_\infty$. We have,

$$\|E_1 + E_2\|_2 \leq \|E_1\|_2 + \beta\|\frac{1}{\beta}E_2\|_2$$

Applying Lemma 8 to the first term and Lemma 9 to the second term, we have that with probability $1 - \frac{1}{n^{10}}$:

$$\begin{aligned}
\|E_1 + E_2\|_2 &\leq \epsilon n\|vec(E_1)\|_\infty + 3\beta \\
&\leq \epsilon n\|vec(E_1)\|_\infty + \frac{6\sqrt{n}}{\sqrt{p}}\|vec(-\mathcal{H}\mathcal{T}(M))\|_\infty
\end{aligned}$$

We now bound $\|vec(E_1)\|_\infty$ and $\|vec(-\mathcal{HT}(M))\|_\infty$. By the hard thresholding step and by incoherence of L^*, we have that $\|vec(E_1)\|_\infty \leq \frac{8\mu^2 r}{m}(3\sigma_1^*)$. We now provide a bound for $\|vec(L)\|_\infty$ that we will use in our bound of $\|vec(-\mathcal{HT}(M))\|_\infty$. Note that $\|vec(L)\|_\infty = \max_{ij}|L_{ij}^*|$. We have

$$|L_{ij}^*| = \left|\sum_{k=1}^r \sigma_k^* U_{ik}^* V_{jk}^*\right| \leq \sum_{k=1}^r \sigma_k^* |U_{ik}^* V_{jk}^*| \leq \sigma_1^* \sum_{k=1}^r |U_{ik}^* V_{jk}^*| \leq \frac{\mu^2 r}{\sqrt{mn}}\sigma_1^*$$

where the last inequality follows from Cauchy-Schwartz and the incoherence of U^*. Now we have that

$$\begin{aligned}
\|vec(-\mathcal{HT}(M))\|_\infty &= \|vec(L^* - \mathcal{HT}(M) - L^*)\|_\infty \\
&\leq \|vec(\mathcal{HT}(M) - L^*)\|_\infty + \|vec(L^*)\|_\infty \\
&= \|vec(E_1)\|_\infty + \|vec(L^*)\|_\infty \\
&\leq \frac{24\mu^2 r}{m}\sigma_1^* + \frac{6\mu^2 r}{m}\sigma_1^* \\
&\leq \frac{30\mu^2 r}{m}\sigma_1^*
\end{aligned}$$

Returning to our bound for $\|E_1 + E_2\|_2$, we now have that

$$\|E_1 + E_2\|_2 \leq 24\epsilon\mu^2 r\sigma_1^* + \frac{180\mu^2 r}{m}\sqrt{\frac{n}{p}}\sigma_1^*$$

For $p > C\frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2 \mu^4 r^{2.5}\log n \log\frac{r\sigma_1^*}{\epsilon'}}{m\delta_{2r}^2}$ and $\epsilon \leq \frac{(\sigma_r^*)^2}{58^2 C_3^2(\sigma_1^*)^4\mu^2 r^2}$, we have

$$\begin{aligned}
\|E_1 + E_2\|_2 &\leq \frac{24(\sigma_r^*)^2}{58^2 C_3^2(\sigma_1^*)^3 r} + \frac{180\sigma_r^*}{r\sqrt{C}} \\
&\leq \frac{1}{140C_3^2\sigma_1^* r} + \frac{180\sigma_r^*}{r\sqrt{C}} \\
&= \frac{\sigma_r^*}{r}\left(\frac{1}{140C_3^2\sigma_1^*\sigma_r^*} + \frac{180}{\sqrt{C}}\right)
\end{aligned}$$

which gives us the result. □

Lemma 8. *(Lemma 3 of [5]) Let $S \in \mathbb{R}^{m\times n}$ be a sparse matrix with row and column sparsity ϵ. Then we have*

$$\|S\|_2 \leq \epsilon \max m, n\|vec(S)\|_\infty$$

Proof. For any pair of unit vectors u and v, we have that

$$\begin{aligned}
v^T S u = \sum_{1\leq i\leq m}\sum_{1\leq j\leq n} v_i u_j S_{ij} &\leq \sum_{1\leq i\leq m}\sum_{1\leq j\leq n} v_i u_j S_{ij}|S_{ij}|\left(\frac{v_i^2 + u_j^2}{2}\right) \\
&\leq \frac{1}{2}\left(\sum_{1\leq i\leq m} v_i^2 \sum_{1\leq j\leq n}|S_{ij}| + \sum_{1\leq j\leq n} u_j^2 \sum_{1\leq i\leq m}|S_{ij}|\right) \\
&\leq \epsilon \max m, n\|vec(S)\|_\infty
\end{aligned}$$

The result now follows by observing that $\|S\|_2 = \max_{u,v,\|u\|_2=1,\|v\|_2=1} u^T S v$. □

Definition 3. *(Definition 1 of [5]) H is a random matrix of size $m \times n$ with each of its entries drawin independently satisfying the following moment conditions:*

$$E[H_{ij}] = 0, \quad |H_{ij} \leq 1|, \quad E[|H_{ij}|^k] \leq \frac{1}{\max m, n}$$

for $i \in [m], j \in [n]$ and $2 \leq k \leq 2 \log n$.

Lemma 9. *(Lemma 4 of [5]) We have the following two claims:*

- *Suppose H satisfies Definition 3. Then with probability at least $1 - \frac{1}{n^{10+\log \alpha}}$, we have that $\|H\|_2 \leq 3\sqrt{\alpha}$.*
- *Let A be a $m \times n$ matrix with $n \geq m$. Suppose $\Omega \subseteq [m] \times [n]$ is obtained by sampling each element with probability $p \in [\frac{1}{4n}, 0.5]$. Then the following matrix H satisfies Definition 3:*

$$H = \frac{\sqrt{p}}{2\sqrt{n}\|vec(A)\|_\infty}\left(A - \frac{1}{p}P_\Omega(A)\right)$$

Proof of Theorem 3

Proof. We proceed by induction on iteration t. The components of the proof are as follows:

- *Base case:* Establish that $dist(U^0, U^*) \leq \alpha$ for some small enough distance α, and establish that U^0 is $K\mu$-incoherent for some K.
- *Induction step (incoherence):* At iteration t, assuming incoherence of U^t, show incoherence of V^{t+1}.
- *Induction step (distance):* At iteration t, assuming $dist(U^t, U^*) \leq \alpha$ and incoherence of U^t, show that $dist(V^{t+1}, V^*)$ decreases by a constant factor. Then using the incoherence induction step to assert incoherence of V^{t+1}, show that $dist(U^{t+1}, U^*)$ decreases by a constant factor.

The base case is given by Lemma 2 for $\alpha \leq \frac{1}{2}$, and the incoherence induction step is given by Lemma 10. With these results in hand, we proceed with the distance induction step.

We follow [8] and use the QR decomposition of the updates instead of analyzing them directly. This change is purely for ease of analysis and has no effect on the algorithm; the updates output exactly the same matrices at the end of each iteration. We have,

$$\widehat{U}^t = U^t R_U^t \quad \text{(QR decomposition)}$$
$$\widehat{V}^{t+1} = \arg\min_{\widehat{V}} \left\| P_\Omega(L^*) - P_\Omega(U^t \widehat{V}^T) \right\|_F - A$$
$$\widehat{V}^{t+1} = V^{t+1} R_V^{t+1} \quad \text{(QR decomposition)}$$
$$\widehat{U}^{t+1} = \arg\min_{\widehat{U}} \left\| P_\Omega(L^*) - P_\Omega(\widehat{U}(V^{t+1})^T) \right\|_F - A$$

where A is the corruption error. Observe that in the clean setting, the update is characterized by a power method update with an error matrix. Specifically,

$$\widehat{V}^{t+1} = V^* \Sigma^* (U^*)^T U^t - F$$
$$V^{t+1} = \widehat{V}^{t+1} (R^{t+1})^{-1}$$

where R^{t+1} is the upper-triangular matrix obtained from the QR-decomposition of $\widehat{V}^{t+1}$ and F is an error matrix. We now define F as defined in Lemma 4.5 of [8]. First, we define the following $nr \times nr$ matrices B, C, D, S.

$$B = \begin{bmatrix} B_{11} & \dots & B_{1r} \\ \vdots & \ddots & \vdots \\ B_{r1} & \dots & B_{rr} \end{bmatrix}, C = \begin{bmatrix} C_{11} & \dots & C_{1r} \\ \vdots & \ddots & \vdots \\ C_{r1} & \dots & C_{rr} \end{bmatrix}, D = \begin{bmatrix} D_{11} & \dots & D_{1r} \\ \vdots & \ddots & \vdots \\ D_{r1} & \dots & D_{rr} \end{bmatrix}, S = \begin{bmatrix} \sigma_1^* I_n & \dots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \dots & \sigma_r^* I_n \end{bmatrix}$$

where for $1 \le k, l \le r$, we have that B_{kl} and C_{kl} are diagonal matrices with

$$(B_{kl})_{jj} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} U_i^t (U_i^t)^T, \quad (C_{kl})_{jj} = \frac{1}{p} \sum_{i:(i,j)\in\Omega} U_i^t (U_i^*)^T$$

and $D_{kl} = \langle U_{*,k}, U_l^* \rangle \mathbb{I}_{n\times n}$ where $U_{*,k}$ is the k-th column of U^t and U_l^* is the l-th left singular vector of the underlying matrix $L^* = U^* \Sigma^* V^{*T}$. In addition, let $v^* = vec(V^*)$. With these definitions in place, we obtain the error matrix F by "de-stacking" the vector $B^{-1}(BD - C)Sv^*$. That is, the j^{th} column of F is given by

$$F_{*,j} = \begin{bmatrix} (B^{-1}(BD - C)Sv^*)_{nj+1} \\ (B^{-1}(BD - C)Sv^*)_{nj+2} \\ \vdots \\ (B^{-1}(BD - C)Sv^*)_{nj+n} \end{bmatrix}$$

and we have $F = [F_{*,1}, F_{*,2} \dots F_{*,r}]$.

Moving to the noisy setting $\epsilon > 0$, we have

$$\widehat{V}^{t+1} = V^* \Sigma^* (U^*)^T U^t - F - A$$
$$V^{t+1} = \widehat{V}^{t+1} (R^{t+1})^{-1}$$

Note that the only difference is the additional error term A due to the corruptions. Multiplying this update equation on the left by $(V_\perp^*)^T$, we get

$$(V_\perp^*)^T V^{t+1} = -(V_\perp^*)^T (F + A)(R^{t+1})^{-1}$$

Therefore we have

$$\begin{aligned} dist(V^{t+1}, V^*) &= \|(V_\perp^*)^T V^{t+1}\|_2 \\ &= \|(V_\perp^*)^T (F + A)(R^{t+1})^{-1}\|_2 \\ &\le \|(F + A)(\Sigma^*)^{-1}\|_2 \|\Sigma^* (R^{t+1})^{-1}\|_2 \\ &\le \Big(\|F(\Sigma^*)^{-1}\|_2 + \|A(\Sigma^*)^{-1}\|_2 \Big) \|\Sigma^* (R^{t+1})^{-1}\|_2 \end{aligned}$$

We bound $\|F(\Sigma^*)^{-1}\|_2$ in Lemma 11, $\|A(\Sigma^*)^{-1}\|_2$ in Lemma 13, and $\|\Sigma^*(R^{t+1})^{-1}\|_2$ in Lemma 12, using incoherence of U^t by our induction hypothesis. Using these three bounds, we have that with high probability:

$$dist(V^{t+1}, V^*) \le \left(\frac{\frac{\sigma_1^*}{\sigma_r^*} \Big(\frac{\delta_{2r} r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 \sigma_1^* \sqrt{\epsilon r \log r}}{\sigma_r^*} \Big)}{\sqrt{1 - dist(U^t, U^*)^2} - \frac{\sigma_1^*}{\sigma_r^*} \Big(\frac{\delta_{2r} r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 \sigma_1^* \sqrt{\epsilon r \log r}}{\sigma_r^*} \Big) dist(U^t, U^*)} \right) dist(U^t, U^*)$$

By our induction hypothesis, we have that $dist(U^t, U^*) \leq \alpha$. So we have

$$dist(V^{t+1}, V^*) \leq \left(\frac{\frac{\sigma_1^*}{\sigma_r^*}\left(\frac{\delta_{2r} r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 \sigma_1^* \sqrt{\epsilon r \log r}}{\sigma_r^*}\right)}{\sqrt{1-\alpha^2} - \frac{\sigma_1^*}{\sigma_r^*}\left(\frac{\delta_{2r} r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 \sigma_1^* \sqrt{\epsilon r \log r}}{\sigma_r^*}\right)\alpha} \right) dist(U^t, U^*)$$

For $\delta_{2r} \leq \frac{\sigma_r^*}{12 r \sigma_1^*}$, we have

$$\begin{aligned} dist(V^{t+1}, V^*) &\leq \left(\frac{\frac{1}{12 - \frac{\sigma_r^*}{r\sigma_1^*}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 (\sigma_1^*)^2 \sqrt{\epsilon r \log r}}{(\sigma_r^*)^2}}{\sqrt{1-\alpha^2} - \alpha\left(\frac{1}{12 - \frac{\sigma_r^*}{r\sigma_1^*}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 (\sigma_1^*)^2 \sqrt{\epsilon r \log r}}{(\sigma_r^*)^2}\right)} \right) dist(U^t, U^*) \\ &\leq \left(\frac{\frac{1}{11} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 (\sigma_1^*)^2 \sqrt{\epsilon r \log r}}{(\sigma_r^*)^2}}{\sqrt{1-\alpha^2} - \alpha\left(\frac{1}{11} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 (\sigma_1^*)^2 \sqrt{\epsilon r \log r}}{(\sigma_r^*)^2}\right)} \right) dist(U^t, U^*) \end{aligned}$$

For $\epsilon \leq \frac{(\sigma_r^*)^2}{58^2 C_3^2 (\sigma_1^*)^4 \mu^2 r^2}$, we have

$$\begin{aligned} dist(V^{t+1}, V^*) &\leq \left(\frac{\frac{1}{11} + \frac{1}{11\sigma_r^*}}{\sqrt{1-\alpha^2} - \alpha\left(\frac{1}{11} + \frac{1}{11\sigma_r^*}\right)} \right) dist(U^t, U^*) \\ &\leq \left(\frac{\frac{2}{11}}{1 - \alpha - \frac{2}{11}\alpha} \right) dist(U^t, U^*) \\ &\leq \frac{2}{11 - 13\alpha} dist(U^t, U^*) \\ &\leq \frac{1}{2} dist(U^t, U^*) \end{aligned}$$

where the last step follows from the fact that $\alpha \leq \frac{1}{2}$ by Lemma 2. Furthermore, by Lemma 10, we have that V_j^{t+1} is $K\mu$-incoherent as well. Therefore by a similar argument as above, we also have that

$$dist(U^{t+1}, U^*) \leq \frac{1}{2} dist(V^{t+1}, V^*)$$

□

Lemma 10. *(Preservation of incoherence) Let U^t be the iterate after the t-th step of the algorithm. If U^t is $K\mu$-incoherent for $K = \frac{13\sigma_1^*}{\sigma_r^*}$, then with high probability, V^{t+1} is also $K\mu$-incoherent.*

Proof. We define matrices $B^j, C^j, D^j \in \mathbb{R}^{r \times r}, 1 \leq j \leq n$:

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

With these definitions, the update equation for V^{t+1} given in Theorem 3 decouples into j equations of the form

$$V_j^{t+1} = (R^{t+1})^{-1}\Big(\big(D^j - (B^j)^{-1}(B^j D^j - C^j)\big)\Sigma^* V_j^* - A_j\Big)$$

for $1 \le j \le n$, where V_j^{t+1} is the j^{th} row of V^{t+1} and V_j^* is the j^{th} row of V^*. We have,

$$\begin{aligned}\|V_j^{t+1}\|_2 &\le \|(R^{t+1})^{-1}\|_2\bigg(\Big(\|D^j\|_2 + \|(B^j)^{-1}\|_2(\|B^j\|_2\|D^j\|_2 + \|C^j\|_2)\Big)\|\Sigma^*\|_2\|V_j^*\|_2 + \|A_j\|_2\bigg) \\ &\le \frac{1}{\sigma_{min}(R^{t+1})}\bigg(\bigg(\|D^j\|_2 + \frac{\|B^j\|_2\|D^j\|_2 + \|C^j\|_2}{\sigma_{min}(B^j)}\bigg)(\sigma_1^*)\frac{\mu\sqrt{r}}{\sqrt{n}} + \bigg(53(\sigma_1^*)^2\mu\sqrt{\frac{13}{11}\epsilon r\log r}\bigg)\frac{\mu\sqrt{r}}{\sqrt{n}}\bigg)\end{aligned}$$

where the last step follows from Lemma 13 and the incoherence of V^*.

By Lemma 12, we have $\sigma_{min}(R^{t+1}) \ge \sigma_r^*\sqrt{1 - dist(U^t, U^*)^2} - \Big(\frac{\sigma_1^*\delta_{2r}r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}\Big)dist(U^t, U^*)$

By Lemma 14, we have $\sigma_{min}(B^j) \ge 1 - \delta_{2r}$ and $\sigma_{max}(B^j) \le 1 + \delta_{2r}$.

By Lemma 15, we have $\sigma_{max}(C^j) \le 1 + \delta_{2r}$.

Lastly, since $D^j = (U^t)^T U^*$, and both U^t and U^* have orthonormal columns, then $\sigma_{max}(D^j) \le 1$.

Bringing these results together, we have

$$\|V_j^{t+1}\|_2 \le \frac{\Big(1 + \frac{(1+\delta_{2r})+(1+\delta_{2r})}{(1-\delta_{2r})}\Big)\sigma_1^*\frac{\mu\sqrt{r}}{\sqrt{n}} + \Big(53(\sigma_1^*)^2\mu\sqrt{\frac{13}{11}\epsilon r\log r}\Big)\frac{\mu\sqrt{r}}{\sqrt{n}}}{\sigma_r^*\sqrt{1 - dist(U^t, U^*)^2} - \Big(\frac{\sigma_1^*\delta_{2r}r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}\Big)dist(U^t, U^*)}$$

For $\delta_{2r} \le \frac{\sigma_r^*}{12r\sigma_1^*} \le \frac{1}{12}$, and since $dist(U^t, U^*) \le \alpha$ by definition of α, we have

$$\|V_j^{t+1}\|_2 \le \left(\frac{4\sigma_1^* + 53(\sigma_1^*)^2\mu\sqrt{\frac{13}{11}\epsilon r\log r}}{\sigma_r^*\sqrt{1-\alpha^2} - \alpha\Big(\frac{\sigma_r^*}{11} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}\Big)}\right)\frac{\mu\sqrt{r}}{\sqrt{n}}$$

For $\epsilon \le \frac{(\sigma_r^*)^2}{58^2C_3^2(\sigma_1^*)^4\mu^2r^2}$, we have

$$\begin{aligned}
\|V_j^{t+1}\|_2 &\leq \left(\frac{4\sigma_1^* + \sigma_r^*}{\sigma_r^*\sqrt{1-\alpha^2} - \left(\frac{\sigma_r^*}{11} + \frac{1}{11}\right)\alpha}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \\
&\leq \left(\frac{4\sigma_1^* + \sigma_r^*}{\sigma_r^*(1-\alpha) - \sigma_r^*(\frac{2}{11}\alpha)}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \\
&\leq \left(\frac{4\sigma_1^*}{\sigma_r^*(1-\frac{13}{11}\alpha)} + \frac{\sigma_r^*}{\sigma_r^*(1-\frac{13}{11}\alpha)}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \\
&\leq \left(\frac{5\sigma_1^*}{\sigma_r^*(1-\frac{13}{11}\alpha)}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \\
&\leq \left(\frac{5\sigma_1^*}{\sigma_r^*(1-\frac{13}{11}\alpha)}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \\
&\leq \left(\frac{13\sigma_1^*}{\sigma_r^*}\right)\frac{\mu\sqrt{r}}{\sqrt{n}}
\end{aligned}$$

Where the last step follows from the fact that $\alpha \leq \frac{1}{2}$ by Lemma 2.

□

Lemma 11. *(Lemma 5.6 of [8]) Let F be the error matrix defined by the update equation in Theorem 3 and let U^t be a $K\mu$-incoherent orthornormal matrix obtained after the $(t-1)^{th}$ update. Then with probability at least $1 - \frac{1}{n^3}$:*

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

Proof. We have,

$$\begin{aligned}
\|F(\Sigma^*)^{-1}\|_2 &\leq \|F(\Sigma^*)^{-1}\|_F \\
&= \|B^{-1}(BD - C)v^*\|_2 \\
&\leq \|B^{-1}\|_2 \|(BD - C)v^*\|_2 \\
&\leq \frac{\delta_{2r} r}{1-\delta_{2r}} dist(U^t, U^*)
\end{aligned}$$

where the last step follows by Lemma 14 and Lemma 16. □

Lemma 12. *(Modified version of Lemma 5.7 of [8]) Let R^{t+1} be the upper-triangular matrix obtained from the QR-decomposition of $\widehat{V}^{t+1}$ and let U^t be a $K\mu$-incoherent orthornormal matrix obtained after the $(t-1)^{th}$ update. Then*

$$\|\Sigma^*(R^{t+1})^{-1}\|_2 \leq \frac{\frac{\sigma_1^*}{\sigma_r^*}}{\sqrt{1 - dist(U^t, U^*)^2} - \frac{\sigma_1^*}{\sigma_r^*}\left(\frac{\delta_{2r} r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5} C_3 \sigma_1^* \sqrt{\epsilon r \log r}}{\sigma_r^*}\right) dist(U^t, U^*)}$$

Proof. Observe that $\|\Sigma^*(R^{t+1})^{-1}\|_2 \leq \frac{\sigma_1^*}{\sigma_{min}(R^{t+1})}$. We have,

$$\begin{aligned}
\sigma_{min}(R^{t+1}) &= \min_{z,\|z\|_2=1} \|R^{t+1}z\|_2 \\
&= \min_{z,\|z\|_2=1} \|V^{t+1}R^{t+1}z\|_2 \\
&= \min_{z,\|z\|_2=1} \|V^*\Sigma^*(U^*)^TU^tz - Fz - Az\|_2 \\
&\geq \min_{z,\|z\|_2=1} \|V^*\Sigma^*(U^*)^TU^tz\|_2 - \|Fz\|_2 - \|Az\|_2 \\
&\geq \min_{z,\|z\|_2=1} \|V^*\Sigma^*(U^*)^TU^tz\|_2 - \|F\|_2 - \|A\|_2 \\
&\geq \sigma_r^*\sigma_{min}((U^*)^TU^t) - \|F\|_2 - \|A\|_2 \\
&\geq \sigma_r^*\sqrt{1 - \|(U_\perp^*)^TU^t\|_2^2} - \sigma_1^*\|F(\Sigma^*)^{-1}\|_2 - \sigma_1^*\|A(\Sigma^*)^{-1}\|_2 \\
&= \sigma_r^*\sqrt{1 - dist(U^t,U^*)^2} - \sigma_1^*\|F(\Sigma^*)^{-1}\|_2 - \sigma_1^*\|A(\Sigma^*)^{-1}\|_2 \\
&\geq \sigma_r^*\sqrt{1 - dist(U^t,U^*)^2} - \frac{\sigma_1^*\delta_{2r}r}{1-\delta_{2r}}dist(U^t,U^*) - \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}dist(U^t,U^*) \\
&= \sigma_r^*\sqrt{1 - dist(U^t,U^*)^2} - \left(\frac{\sigma_1^*\delta_{2r}r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}\right)dist(U^t,U^*)
\end{aligned}$$

where the second to last step follows from Lemmas 11 and 13. Therefore we have,

$$\begin{aligned}
\|\Sigma^*(R^{t+1})^{-1}\|_2 &\leq \frac{\sigma_1^*}{\sigma_r^*\sqrt{1 - dist(U^t,U^*)^2} - \left(\frac{\sigma_1^*\delta_{2r}r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3(\sigma_1^*)^2\sqrt{\epsilon r\log r}}{\sigma_r^*}\right)dist(U^t,U^*)} \\
&= \frac{\frac{\sigma_1^*}{\sigma_r^*}}{\sqrt{1 - dist(U^t,U^*)^2} - \frac{\sigma_1^*}{\sigma_r^*}\left(\frac{\delta_{2r}r}{1-\delta_{2r}} + \frac{\left(\frac{13}{11}\right)^{1.5}C_3\sigma_1^*\sqrt{\epsilon r\log r}}{\sigma_r^*}\right)dist(U^t,U^*)}
\end{aligned}$$

□

Lemma 13. *(Corruption error) Let A be the error matrix resulting from the gradient estimation defined by Theorem 3 and let U^t be a $K\mu$-incoherent orthornormal matrix obtained after the $(t-1)^{th}$ update. Then with high probability, the following two claims hold:*

$$\|A(\Sigma^*)^{-1}\|_2 \leq \frac{\left(\frac{13}{11}\right)^{1.5}C_3\sigma_1^*\sqrt{\epsilon r\log r}}{\sigma_r^*}dist(U^t,U^*)$$

$$\|A_j\|_2 \leq \left(53(\sigma_1^*)^2\mu\sqrt{\frac{13}{11}\epsilon r\log r}\right)\frac{\mu\sqrt{r}}{\sqrt{n}} \quad \forall j$$

Proof. Note that $\|A(\Sigma^*)^{-1}\|_2 \leq \|A\|_2\|(\Sigma^*)^{-1}\|_2 = \frac{1}{\sigma_r^*}\|A\|_2$. It remains to bound $\|A\|_2$. Observe that to produce the next iterate $\widehat{V}^{t+1}$, Algorithm 1 runs robust regression on each row $\widehat{V}_j^{t+1}$. The (x,y) pairs are (U_i^t, M_{ij}) for each i such that M_{ij} is observed, and the parameter is $\widehat{V}_j^{t+1}$. Note

that ϵ fraction of the observed M_{ij} values for each row and column of M are corrupted, and the true labels are the L^*_{ij}. Therefore we have j linear models of the form

$$L^*_{ij} = U^t_i \widehat{V}_j^{(t+1)^*} + \omega_{ij}$$

where $\widehat{V}_j^{(t+1)^*}$ can be viewed as the "oracle" $\widehat{V}_j^{t+1}$, that is, the iterate $\widehat{V}_j^{t+1}$ that minimizes our objective $\|P_\Omega(L^*) - P_\Omega(U^t(\widehat{V}^{t+1})^T)\|_F$. By employing Robust Gradient Descent with the Huber Gradient Estimator as our robust regressor [6], we can apply Theorem 18. By Theorem 18, for sufficiently large number of iterations τ', we have that $\forall 1 \leq j \leq n$, with probability at least $1 - \frac{1}{n^3}$:

$$\|A_j\|_2 \leq C_3 \sqrt{\frac{\sigma_j^2 \|\Sigma_j\|_2 \epsilon \log r}{\tau_{\ell j}}}$$

where σ_j^2 is the variance of the noise ω, Σ_j is the covariance matrix of the data U^t, and $\tau_{\ell j}$ is the minimum eigenvalue of Σ_j. The $\gamma(\tilde{n}, p, \tilde{\delta})$ term in Theorem 18 is a sample-dependent term, and is negligible for sufficiently large n. We take $1 - \kappa = \tau_{\ell j}$ for all j due to (34) of [6].

We bound σ_j^2, the variance of the noise ω. Noting that $\omega_{ij} = L^*_{ij} - U^t_i \left(\widehat{V}_j^{(t+1)^*}\right)^T$, we have that

$$\sum_{i=1}^{m} \sigma_j^2 \leq \left\| L^* - U^t \left(\widehat{V}^{(t+1)^*}\right)^T \right\|_F^2$$

We now derive an upper bound for $\left\| L^* - U^t \left(\widehat{V}^{(t+1)^*}\right)^T \right\|_F$ with an argument similar to Lemma 4, except we consider $\widehat{V}^{(t+1)^*}$ instead of $\widehat{V}^{t+1}$.

Recall that the update equation for $\widehat{V}^{t+1}$ as written in Theorem 3 is

$$\widehat{V}^{t+1} = V^* \Sigma^* (U^*)^T U^t - F - A$$

where A is the corruption error. Since $\widehat{V}^{(t+1)^*}$ is the oracle, that is, the iterate that minimizes our objective under no corruptions, the update equation for $\widehat{V}^{(t+1)^*}$ is

$$\widehat{V}^{(t+1)^*} = V^* \Sigma^* (U^*)^T U^t - F$$

With this modification, we can now proceed with the same argument as Lemma 4. We have

$$\begin{aligned}
\left\| L^* - U^t \left(\widehat{V}^{(t+1)^*}\right)^T \right\|_F &= \left\| L^* - U^t \left(V^* \Sigma^* (U^*)^T U^t - F\right)^T \right\|_F \\
&= \left\| \left(I - U^t (U^t)^T\right) L^* + U^t F^T \right\|_F \\
&\leq \|(I - U^t (U^t)^T) L^*\|_F + \|F\|_F \\
&\leq \|(I - U^t (U^t)^T) U^* \Sigma^* V^*\|_F + \|F\|_F
\end{aligned}$$

Since V^* has orthonormal columns, $\|V^*\|_F = \sqrt{r}$. We have

$$\begin{aligned}
\left\| L^* - U^t \left(\widehat{V}^{(t+1)^*}\right)^T \right\|_F &\leq \sqrt{r} \|(I - U^t (U^t)^T) U^* \Sigma^*\|_2 + \|F\|_F \\
&\leq \sqrt{r} \|(U^t_\perp)^T U^* \Sigma^*\|_2 + \|F\|_F \\
&\leq \sqrt{r} (\sigma_1^*) dist(U^t, U^*) + \|F\|_F
\end{aligned}$$

By Lemma 11, we have

$$\left\|L^* - U^t\left(\widehat{V}^{(t+1)^*}\right)^T\right\|_F \leq \sqrt{r}(\sigma_1^*)dist(U^t, U^*) + \left(\frac{\delta_{2r} r\sigma_r^*}{1-\delta_{2r}}\right)dist(U^t, U^*)$$

For $\delta_{2r} \leq \frac{\sigma_r^*}{12r\sigma_1^*}$, we have

$$\begin{aligned}\left\|L^* - U^t\left(\widehat{V}^{(t+1)^*}\right)^T\right\|_F &\leq \sqrt{r}(\sigma_1^*)dist(U^t, U^*) + \frac{(\sigma_r^*)^2}{11\sigma_1^*}dist(U^t, U^*) \\ &\leq \left(\sigma_1^*\sqrt{r} + \frac{\sigma_r^*}{11}\right)dist(U^t, U^*) \\ &\leq \frac{12}{11}\sigma_1^*\sqrt{r}\big(dist(U^t, U^*)\big)\end{aligned}$$

We now bound $\tau_{\ell j}$ and $\|\Sigma_j\|_2$, the minimum and maximum eigenvalues of Σ_j respectively. Observe that $\Sigma_j = B^j = \frac{1}{p}\sum_{i:(i,j)\in\Omega} U_i^t (U_i^t)^T$. By Lemma 14, we have that with probability at least $1 - \frac{1}{n^3}$, $\sigma_{min}(B^j) \geq 1 - \delta_{2r}$, therefore $\tau_{\ell j} \geq 1 - \delta_{2r}$. By the Bernstein inequality used in Lemma 14, we also have that with probability at least $1 - \frac{1}{n^3}$, $\sigma_{max}(B^j) \leq 1 + \delta_{2r}$ for all j. Therefore $\|\Sigma_j\|_2 \leq 1 + \delta_{2r}$. For $\delta_{2r} \leq \frac{\sigma_r^*}{12r\sigma_1^*} \leq \frac{1}{12}$, we have that $\tau_{\ell j} \geq \frac{11}{12}$ and $\|\Sigma_j\|_2 \leq \frac{13}{12}$ for all j.

With these results in hand, we have

$$\begin{aligned}\|A\|_2 &\leq \|A\|_F \\ &= \sqrt{\sum_{j=1}^m \|A_j\|_2^2} \\ &\leq \sqrt{\sum_{j=1}^m C_3^2\left(\frac{\sigma_j^2\|\Sigma_j\|_2\epsilon\log r}{\tau_{\ell j}}\right)} \\ &\leq \sqrt{C_3^2\frac{13}{11}\epsilon\log r\sum_{j=1}^m \sigma_j^2} \\ &\leq \sqrt{C_3^2\frac{13}{11}\epsilon\log r\Big(\frac{12}{11}\sigma_1^*\sqrt{r}\big(dist(U^t, U^*)\big)\Big)^2} \\ &\leq C_3\Big(\frac{12}{11}\sigma_1^*\sqrt{r}\big(dist(U^t, U^*)\big)\Big)\sqrt{\frac{13}{11}\epsilon\log r} \\ &\leq \left(\frac{13}{11}\right)^{1.5} C_3\sigma_1^*\sqrt{\epsilon r\log r}\Big(dist(U^t, U^*)\Big)\end{aligned}$$

This proves the first claim of the lemma. For the second claim, we derive a bound for σ_j^2 using the incoherence property. Recall that $\omega_{ij} = L_{ij}^* - U_i^t\left(\widehat{V}_j^{(t+1)^*}\right)^T$ for all i, j. As shown in Lemma 7, we have that $|L_{ij}^*| \leq \frac{\mu^2 r}{\sqrt{mn}}\sigma_1^*$ for all i, j due to incoherence of L^*. We know that U^t is $K\mu$-incoherent by assumption, so in order to bound $U_i^t\left(\widehat{V}_j^{(t+1)^*}\right)^T$, we must establish the incoherence of $\widehat{V}^{(t+1)^*}$. We do so using a similar argument to Lemma 10, but considering $\widehat{V}^{(t+1)^*}$ instead of V^{t+1}.

The update equation for each row $\widehat{V}_j^{t+1}$ is given by

$$\widehat{V}_j^{t+1} = \Big(D^j - (B^j)^{-1}(B^j D^j - C^j)\Big)\Sigma^* V_j^*$$

and we have

$$\begin{aligned}\|\widehat{V}_j^{t+1}\|_2 &\leq \Big(\|D^j\|_2 + \|(B^j)^{-1}\|_2(\|B^j\|_2\|D^j\|_2 + \|C^j\|_2)\Big)\|\Sigma^*\|_2\|V_j^*\|_2 \\ &\leq \sigma_1^*\frac{\mu\sqrt{r}}{\sqrt{n}}\left(\|D^j\|_2 + \frac{\|B^j\|_2\|D^j\|_2 + \|C^j\|_2}{\sigma_{min}(B^j)}\right)\end{aligned}$$

where the last step follows from incoherence of V^*.

By Lemma 14, we have $\sigma_{min}(B^j) \geq 1 - \delta_{2r}$ and $\sigma_{max}(B^j) \leq 1 + \delta_{2r}$.

By Lemma 15, we have $\sigma_{max}(C^j) \leq 1 + \delta_{2r}$.

Lastly, since $D^j = (U^t)^T U^*$, and both U^t and U^* have orthonormal columns, then $\sigma_{max}(D^j) \leq 1$.

Bringing these results together, we have

$$\left\|\widehat{V}_j^{(t+1)^*}\right\|_2 \leq \sigma_1^*\frac{\mu\sqrt{r}}{\sqrt{n}}\left(1 + \frac{(1+\delta_{2r}) + (1+\delta_{2r})}{(1-\delta_{2r})}\right)$$

For $\delta_{2r} \leq \frac{\sigma_r^*}{12r\sigma_1^*} \leq \frac{1}{12}$, we have

$$\left\|\widehat{V}_j^{(t+1)^*}\right\|_2 \leq 4\sigma_1^*\frac{\mu\sqrt{r}}{\sqrt{n}}$$

So $\widehat{V}^{(t+1)^*}$ is $4\sigma_1^*\mu$-incoherent. With incoherence of $\widehat{V}^{(t+1)^*}$ established, we can use the Cauchy-Schwartz inequality to assert that for all i, j:

$$\left|U_i^t\left(\widehat{V}_j^{(t+1)^*}\right)^T\right| \leq \left\|U_i^t\right\|_2\left\|\widehat{V}_j^{(t+1)^*}\right\|_2 \leq \left(\frac{13\sigma_1^*}{\sigma_r^*}\right)\frac{\mu\sqrt{r}}{\sqrt{m}}\left(4\sigma_1^*\right)\frac{\mu\sqrt{r}}{\sqrt{n}} = \frac{52(\sigma_1^*)^2\mu^2 r}{\sigma_r^*\sqrt{mn}}$$

Therefore we have that for all i, j:

$$\begin{aligned}|\omega_{ij}| &= \left|L_{ij}^* - U_i^t\left(\widehat{V}_j^{(t+1)^*}\right)^T\right| \\ &\leq |L_{ij}^*| + \left|U_i^t\left(\widehat{V}_j^{(t+1)^*}\right)^T\right| \\ &\leq \frac{\sigma_1^*\mu^2 r}{\sqrt{mn}} + \frac{52(\sigma_1^*)^2\mu^2 r}{\sigma_r^*\sqrt{mn}} \\ &\leq \frac{(\sigma_r^*)^2\sigma_1^*\mu^2 r + 52(\sigma_1^*)^2\mu^2 r}{\sigma_r^*\sqrt{mn}} \\ &\leq \frac{53(\sigma_1^*)^2\sigma_r^*\mu^2 r}{\sigma_r^*\sqrt{mn}} \\ &\leq \frac{53(\sigma_1^*)^2\mu^2 r}{\sqrt{n}}\end{aligned}$$

Therefore $\sigma_j \leq \frac{53(\sigma_1^*)^2\mu^2 r}{\sqrt{n}}$ for all j. Using this new bound for σ_j, we have that for all j:

$$\begin{aligned}\|A_j\|_2 &\leq C_3\sqrt{\frac{\sigma_j^2\|\Sigma_j\|_2\epsilon\log r}{\tau_{\ell j}}} \\ &\leq C_3\left(\frac{53(\sigma_1^*)^2\mu^2 r}{\sqrt{n}}\right)\sqrt{\frac{13}{11}\epsilon\log r} \\ &= \left(53(\sigma_1^*)^2\mu\sqrt{\frac{13}{11}\epsilon r\log r}\right)\frac{\mu\sqrt{r}}{\sqrt{n}}\end{aligned}$$

□

Lemma 14. *(Lemma C.6 of [8]) Let B be defined as in Theorem 3. Then with probability at least $1 - \frac{1}{n^3}$:*

$$\|B^{-1}\|_2 \leq \frac{1}{1 - \delta_{2r}}$$

Proof. Let $X \in \mathbb{R}^{n \times r}$ and let $x = vec(X) \in \mathbb{R}^{nr}$ such that $\|x\|_2 = 1$. We have,

$$\begin{aligned}\|B^{-1}\|_2 &= \frac{1}{\sigma_{min}(B)} \\ &= \frac{1}{min_{x,\|x\|_2=1} x^T B x}\end{aligned}$$

As stated in Lemma 10, define

$$B^j = \frac{1}{p} \sum_{i:(i,j)\in\Omega} U_i^t (U_i^t)^T$$

We have that $\forall x$,

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

The Lemma now follows by the following lower bound on $\sigma_{min}(B^j), \forall j$. Consider any $w \in \mathbb{R}^r$ such that $\|w\|_2 = 1$. We have,

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

Observe that $\mathbb{E}[Z] = w^t U^t (U^t)^T w = w^T w = 1$ and

$$\begin{aligned}\mathbb{E}[Z^2] &= \frac{1}{p} \sum_i \langle w, U_i^t \rangle^4 \\ &\leq \frac{K^2 \mu^2 r}{mp} \sum_i \langle w, U_i^t \rangle^2 \\ &= \frac{K^2 \mu^2 r}{mp}\end{aligned}$$

where the inequality follows from incoherence of U^t. Similarly, $\max_i |\langle w, U_i^t \rangle^2| \leq \frac{K^2 \mu^2 r}{mp}$. Therefore by Bernstein's inequality, we have,

$$Pr(|Z - \mathbb{E}[Z]| \geq \delta_{2r}) \leq \exp\left(- \frac{\frac{\delta_{2r}^2 mp}{2}}{(1 + \frac{\delta_{2r}}{3}) K^2 \mu^2 r} \right) = \exp\left(- \frac{\frac{\delta_{2r}^2 mp}{2}}{(1 + \frac{\delta_{2r}}{3}) (\frac{13\sigma_1^*}{\sigma_r^*})^2 \mu^2 r} \right)$$

For $p > C \frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2 \mu^4 r^{2.5} \log n \log \frac{r\sigma_1^*}{\epsilon'}}{m \delta_{2r}^2}$, by the union bound we have that with probability at least $1 - \frac{1}{n^3}$,

$$w^T B^j w \geq 1 - \delta_{2r} \quad \forall w, j$$

Therefore $\forall j, \sigma_{min}(B^j) \geq 1 - \delta_{2r}$. □

Lemma 15. *(Lemma C.7 of [8]) Let* $C^j = \frac{1}{p}\sum_{i:(i,j)\in\Omega} U_i^t (U_i^*)^T$. *Then with probability at least* $1 - \frac{1}{n^3}$*:*

$$\|C^j\|_2 \leq 1 + \delta_{2r} \quad \forall j$$

Proof. Let $x, y \in \mathbb{R}^r$ be two arbitrary unit vectors. Then we have,

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

Therefore $Z = x^T C^j y = \frac{1}{p}\sum_i \delta_{ij}(x^T U_i^t)(y^T U_i^*)$. Observe that $\mathbb{E}[Z] = x^T (U^t)^T U^* y$ and

$$\begin{aligned}
\mathbb{E}[Z^2] &= \frac{1}{p}\sum_i (x^T U_i^t)^2 (y^T U_i^*)^2 \\
&\leq \frac{\mu^2}{mp} x^T (U^t)^T U^t x \\
&= \frac{\mu^2 r}{mp}
\end{aligned}$$

where the inequality follows from incoherence of U^* and the last step from the fact that U^t is orthornormal and x is a unit vector. In addition, $\max_i |(x^T U_i^t)(y^T U_i^*)| \leq \frac{K^2\mu^2 r}{mp}$. Therefore by Bernstein's inequality, we have,

$$Pr(|Z - \mathbb{E}[Z]| \geq \delta_{2r}) \leq \exp\left(-\frac{\frac{\delta_{2r}^2 mp}{2}}{(1 + \frac{K^2\delta_{2r}}{3})\mu^2 r}\right) = \exp\left(-\frac{\frac{\delta_{2r}^2 mp}{2}}{(1 + (\frac{13\sigma_1^*}{\sigma_r^*})^2\frac{\delta_{2r}}{3})\mu^2 r}\right)$$

For $p > C\frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2 \mu^4 r^{2.5}\log n \log\frac{r\sigma_1^*}{\epsilon'}}{m\delta_{2r}^2}$, by the union bound we have that with probability at least $1 - \frac{1}{n^3}$,

$$x^T C^j y \leq 1 + \delta_{2r} \quad \forall x, y, j$$

Therefore $\forall j, \|C^j\|_2 \leq 1 + \delta_{2r}$. □

Lemma 16. *(Lemma C.8 of [8]) Let* F *be defined as in Theorem 3. Then with probability at least* $1 - \frac{1}{n^3}$*:*

$$\|(BD - C)v^*\|_2 \leq \delta_{2r} r(dist(U^t, U^*))$$

Proof. Let $X \in \mathbb{R}^{n\times r}$ and let $x = vec(X) \in \mathbb{R}^{nr}$ such that $\|x\|_2 = 1$. Let $H^j = (B^j D^j - C^j)$, that is,

$$\begin{aligned}
H^j &= \frac{1}{p}\sum_{i:(i,j)\in\Omega} U_i^t (U_i^t)^T (U^t)^T U^* - U_i^t (U_i^*)^T \\
&= \frac{1}{p}\sum_{i:(i,j)\in\Omega} H_i^j
\end{aligned}$$

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

$$\sum_i H_i^j = (U^t)^T U^t (U^t)^T U^* - (U^t)^T U^* = 0$$

Now observe that

$$\begin{aligned} x^T(BD-C)v^* &= \sum_j (X_j)^T (B^j D^j - C^j) V_j^* \\ &= \frac{1}{p}\sum_{kl}\sum_{(i,j)\in\Omega} X_{jk} V_{jl}^* (H_i^j)_{kl} \end{aligned}$$

Using the previous equation, we have that $\forall (k,l)$,

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

Therefore by Lemma 17, we have that with probability at least $1-\frac{1}{n^3}$,

$$\begin{aligned} x^T(BD-C)v^* &= \sum_j (X_j)^T (B^j D^j - C^j) V_j^* \\ &\le \frac{1}{p}\sum_{kl} \sqrt{\sum_j (X_{jk})^2 (V_{jl}^*)^2} \sqrt{\sum_i (H_i^j)_{kl}^2} \end{aligned}$$

In addition, we have,

$$\begin{aligned} \sum_i (H_i^j)_{kl}^2 &= \sum_i (U_{ik}^t)^2 ((U_i^t)^T (U^t)^T U_l^* - U_{il}^*)^2 \\ &\le \max_i (U_{ik}^t)^2 \sum_i (U_i^t)^T ((U_i^t)^T (U^t)^T U_l^* - U_{il}^*)^2 \\ &= \max_i (U_{ik}^t)^2 (1 - \|U^t U_l^*\|_2^2) \\ &\le \frac{K^2\mu^2 r}{m} dist(U^t, U^*)^2 \end{aligned}$$

where the last step follows by incoherence of U^t. Therefore by the above two equations and incoherence of V^*, we have that with probability at least $1-\frac{1}{n^3}, \forall x$:

$$\begin{aligned} x^T(BD-C)v^* &\le \sum_{kl} \frac{K^2\mu^2 r}{mp} dist(U^t, U^*) \|X_{*,k}\|_2 \\ &\le \frac{K^2\mu^2 r}{mp} dist(U^t, U^*) \sum_{kl} \|X_{*,k}\|_2 \\ &\le \frac{K^2\mu^2 r}{mp} dist(U^t, U^*) \sum_{kl} \sqrt{r} \\ &\le \frac{K^2\mu^2 r^{3.5}}{mp} dist(U^t, U^*) \\ &= \frac{(\frac{13\sigma_1^*}{\sigma_r^*})^2 \mu^2 r^{3.5}}{mp} dist(U^t, U^*) \end{aligned}$$

where we use the fact that $\sum_k \|X_{*,k}\|_2 \leq \sqrt{r}\|x\|_2 = \sqrt{r}$. Then for $p > C\frac{\left(\frac{\sigma_1^*}{\sigma_r^*}\right)^2 \mu^4 r^{2.5} \log n \log \frac{r\sigma_1^*}{\epsilon'}}{m\delta_{2r}^2}$, we have

$$x^T(BD - C)v^* \leq \delta_{2r} r(dist(U^t, U^*))$$

Since $\max_{x,\|x\|_2=1} x^T(BD - C)v^* = \|(BD - C)v^*\|_2$, we have,

$$\|(BD - C)v^*\|_2 \leq \delta_{2r} r(dist(U^t, U^*))$$

□

Lemma 17. *(Lemma C.5 of [8]) Let Ω be a set of indices sampled uniformly at random from $[m] \times [n]$ with each element of $[m] \times [n]$ sampled independently with probability $p \geq \frac{C \log n}{m}$. Then with probability at least $1 - \frac{1}{n^3}, \forall x \in \mathbb{R}^m, y \in \mathbb{R}^n s.t. \sum_i x_i = 0$, we have*

$$\sum_{(i,j)\in\Omega} x_i y_j \leq \left(C\sqrt{(\sqrt{mn})p}\right)\|x\|_2\|y\|_2$$

where $C > 0$ is a global constant.

Theorem 18. *(Robust Linear Regression, Theorem 2 of [6]) Consider the statistical model where each (x, y) is such that $y = X^T\theta^* + \omega$. Assume the number of samples n is large enough such that $\gamma(\tilde{n}, r, \tilde{\delta}) < \frac{C_1\tau_\ell}{\|\Sigma\|_2\sqrt{\log r}}$ and the contamination level is such that*

$$\epsilon < \left(\frac{C_2\tau_\ell}{\|\Sigma\|_2\sqrt{\log r}} - \gamma(\tilde{n}, r, \tilde{\delta})\right)^2$$

for some constants C_1, C_2. Then there are universal constants C_3, C_4 such that if Algorithm 1 (RGD) is initialized at θ^0 with step size $\nu = \frac{2}{\tau_u + \tau_\ell}$ and Algorithm 2 (Huber estimator) as a gradient estimator, then it returns iterates $\{\hat{\theta}\}_{t=1}^T$ such that for a contraction parameter $\kappa < 1$, with probability at least $1 - \delta$,

$$\|\hat{\theta} - \theta^*\|_2 \leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{C_3\sigma\sqrt{\|\Sigma\|_2 \log r}}{1 - \kappa}(\epsilon^{\frac{1}{2}} + \gamma(\tilde{n}, p, \tilde{\delta}))$$

where σ^2 is the variance of the ω, and Σ is the covariance matrix of the covariates x.