Matrix Completion From a Few Entries

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh, *Student Member, IEEE*

***Abstract*—Let M be an $n\alpha \times n$ matrix of rank r, and assume that a uniformly random subset E of its entries is observed. We describe an efficient algorithm, which we call OptSpace, that reconstructs M from $|E| = O(r\,n)$ observed entries with relative root mean square error**

$$\mathrm{RMSE} \leq C(\alpha)\left(\frac{nr}{|E|}\right)^{1/2}$$

with probability larger than $1 - 1/n^3$. Further, if $r = O(1)$ and M is sufficiently unstructured, then OptSpace reconstructs it *exactly* from $|E| = O(n \log n)$ entries with probability larger than $1 - 1/n^3$. This settles (in the case of bounded rank) a question left open by Candès and Recht and improves over the guarantees for their reconstruction algorithm. The complexity of our algorithm is $O(|E| r \log n)$, which opens the way to its use for massive data sets. In the process of proving these statements, we obtain a generalization of a celebrated result by Friedman–Kahn–Szemerédi and Feige–Ofek on the spectrum of sparse random matrices.

***Index Terms*—Gradient descent, low rank, manifold optimization, matrix completion, phase transition, spectral methods.**

I. Introduction

IMAGINE that each of m customers watches and rates a subset of the n movies available through a movie rental service. This yields a dataset of customer-movie pairs $(i,j) \in E \subseteq [m] \times [n]$ and, for each such pair, a rating $M_{ij} \in \mathbb{R}$. The objective of *collaborative filtering* is to predict the rating for the missing pairs in such a way as to provide targeted suggestions.[1] The general question we address here is: Under which conditions do the known ratings provide sufficient information to infer the unknown ones? Can this inference problem be solved efficiently? The second question is particularly important in view of the massive size of actual data sets.

A. Model Definition

A simple mathematical model for such data assumes that the (unknown) matrix of ratings has rank $r \ll m, n$. More precisely, we denote by M the matrix whose entry $(i,j) \in [m] \times [n]$ corresponds to the rating user i would assign to movie j. We assume that there exist matrices U, of dimensions $m \times r$, and V, of dimensions $n \times r$ such that $U^T U = m\mathbf{I}$, $V^T V = n\mathbf{I}$ (here and below $\mathbf{I}$ denotes the identity matrix), and a diagonal matrix Σ, of dimensions $r \times r$ such that

$$M = U \Sigma V^T\,. \tag{1}$$

For justification of these assumptions and background on the use of low rank matrices in information retrieval, we refer to [4]. Since we are interested in very large data sets, we shall strive to prove performance guarantees that are asymptotically optimal for large m, n. However, our main results are completely nonasymptotic and provide concrete bounds for any m, n. Notice that we can assume $m \geq n$, since we can always apply our algorithm to the transpose of the matrix M. Throughout this paper, therefore, we will assume $\alpha \equiv m/n \geq 1$.

We further assume that the factors U, V are unstructured. This notion is formalized by the *incoherence condition* introduced by Candès and Recht [9], and defined in Section I-C.

Out of the $m \times n$ entries of M, a subset $E \subseteq [m] \times [n]$ (the user/movie pairs for which a rating is available) is revealed. We let M^E be the $m \times n$ matrix that contains the revealed entries of M, and is filled with 0's in the other positions

$$M^E_{ij} = \begin{cases} M_{ij}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The set E will be uniformly random given its size $|E|$.

B. Algorithm

A naive algorithm consists of the following projection operation.

Projection: Compute the singular value decomposition (SVD) of M^E (with $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$)

$$M^E = \sum_{i=1}^{\min(m,n)} \sigma_i x_i y_i^T \tag{3}$$

and return the matrix $\mathsf{P}_r(M^E) = (mn/|E|) \sum_{i=1}^{r} \sigma_i x_i y_i^T$ obtained by setting to 0 all but the r largest singular values. Notice that, apart from the rescaling factor $(mn/|E|)$, $\mathsf{P}_r(M^E)$ is the orthogonal projection of M^E onto the set of rank-r matrices. The rescaling factor compensates the smaller average size of the entries of M^E with respect to M.

It turns out that, if $|E| = \Theta(n)$, this algorithm performs very poorly. The reason is that the matrix M^E contains columns and rows with $\Theta(\log n / \log\log n)$ nonzero (revealed) entries. These over-represented rows/columns alter the spectrum of M^E as discussed in Section II. This motivates the definition of the following operation (hereafter, the *degree* of a column or of a row is the number of its revealed entries).

Manuscript received March 18, 2009; revised November 12, 2009. Current version published May 19, 2010. This work was supported in part by a Terman fellowship and in part by an NSF CAREER award (CCF-0743978). The material in this paper was presented in part at ISIT, Seoul, Korea, July 2009.

R. H. Keshavan and S. Oh are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: raghuram@stanford.edu; swoh@stanford.edu).

A. Montanari is with the Department of Electrical Engineering and the Departments of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: montanari@stanford.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2046205

[1]Indeed, in 2006, Netflix made public such a dataset with $m \approx 5 \cdot 10^5$, $n \approx 2 \cdot 10^4$ and $|E| \approx 10^8$ and challenged the research community to predict the missing ratings with root mean square error below 0.8563 [29].

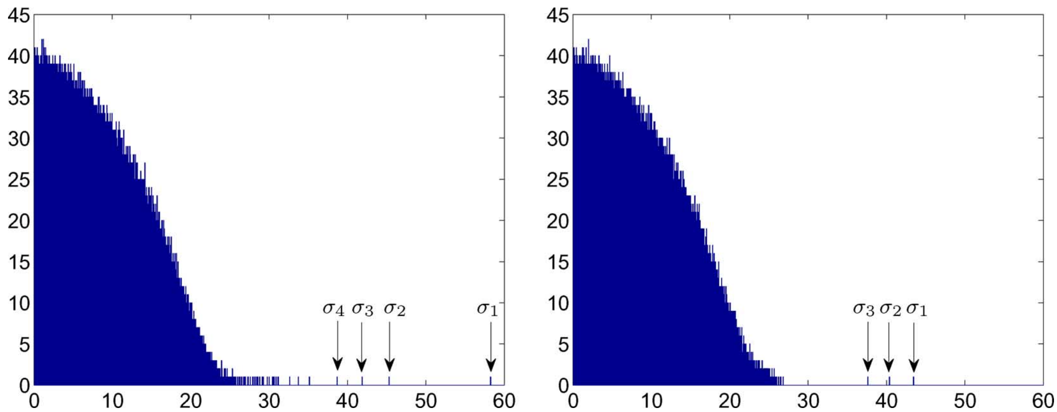

Fig. 1. Histogram of the singular values of a partially revealed matrix M^E before trimming (left) and after trimming (right) for $10^4 \times 10^4$ random rank-3 matrix M with $\epsilon = 30$ and $\Sigma = \mathrm{diag}(1, 1.1, 1.2)$. After trimming, the underlying rank-3 structure becomes clear. Here, the number of revealed entries per row follows a heavy tail distribution with $\mathbb{P}\{N = k\} = \mathrm{const.}/k^3$.

Trimming: Set to zero all columns in M^E with degree larger that $2|E|/n$. Set to 0 all rows with degree larger than $2|E|/m$. Let the matrix thus obtained be $\widetilde{M}^E$.

Fig. 1 shows the singular value distributions of M^E and $\widetilde{M}^E$ for a random rank-3 matrix M. The surprise is that trimming (which amounts to "throwing out information") makes the underlying rank-3 structure much more apparent. This effect becomes even more important when the number of revealed entries per row/column follows a heavy tail distribution, as is the case for real data.

In terms of the above routines, our algorithm has the following structure.

OPTSPACE (matrix M^E)

1:Trim M^E, and let $\widetilde{M}^E$ be the output.

2:Project $\widetilde{M}^E$ to $\mathsf{P}_r(\widetilde{M}^E)$.

3:Clean residual errors by minimizing the discrepancy $F(X, Y)$.

The last step of the above algorithm allows to reduce (or eliminate) small discrepancies between $\mathsf{P}_r(\widetilde{M}^E)$ and M, and is described below. Note that the discarded entries from the trimming step are incorporated back into the algorithm in this cleaning step.

Cleaning: Various implementations are possible, but we found the following one particularly appealing. Given $X \in \mathbb{R}^{m\times r}$, $Y \in \mathbb{R}^{n\times r}$ with $X^TX = m\mathbf{I}$ and $Y^TY = n\mathbf{I}$, we define

$$F(X,Y) \equiv \min_{S\in\mathbb{R}^{r\times r}} \mathcal{F}(X,Y,S) \tag{4}$$

$$\mathcal{F}(X,Y,S) \equiv \frac{1}{2}\sum_{(i,j)\in E}(M_{ij} - (XSY^T)_{ij})^2 . \tag{5}$$

The cleaning step consists in writing $\mathsf{P}_r(\widetilde{M}^E) = X_0S_0Y_0^T$ and minimizing $F(X, Y)$ locally with initial condition $X = X_0$, $Y = Y_0$.

Notice that $F(X, Y)$ is easy to evaluate since it is defined by minimizing the quadratic function $S \mapsto \mathcal{F}(X, Y, S)$ over the low-dimensional matrix S. Further, it depends on the matrices X and Y only through their column spaces. In geometric terms, F is a function defined over the Cartesian product of two Grassmann manifolds (we refer to Section VI for background and references). Optimization over Grassmann manifolds is a well understood topic [15] and efficient algorithms (in particular Newton and conjugate gradient) can be applied. To be definite, we assume that gradient descent with line search is used to minimize $F(X, Y)$.

The implementation proposed here implicitly assumes that the rank r is known. A very simple algorithm for estimating the rank of the matrix M from the revealed entries is introduced in [23]. It is proved there that the algorithm recovers the correct rank with high probability under the hypotheses of Theorem 1.1 (i.e., bounded entries) with the additional assumption that $|E| \geq Crn$.

C. *Incoherence Property*

In order to formalize the notion of incoherence, we write $U = [u_1, u_2, \ldots, u_r]$ and $V = [v_1, v_2, \ldots, v_r]$ for the columns of the two factors, with $||u_i|| = \sqrt{m}$, $||v_i|| = \sqrt{n}$ and $u_i^T u_j = 0$, $v_i^T v_j = 0$ for $i \neq j$ (there is no loss of generality in this, since normalizations can be adsorbed by redefining Σ). We shall further write $\Sigma = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_r)$ with $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_r > 0$. Note that Σ_k is the kth singular value of M scaled by $1/\sqrt{mn}$.

The matrices U, V and Σ will be said to be (μ_0, μ_1)-*incoherent* if they satisfy the following properties:

A0. For all $i \in [m]$, $j \in [n]$, we have $\sum_{k=1}^r U_{i,k}^2 \leq \mu_0 r$, $\sum_{k=1}^r V_{j,k}^2 \leq \mu_0 r$.

A1. For all $i \in [m]$, $j \in [n]$, we have $|\sum_{k=1}^r U_{i,k}(\Sigma_k/\Sigma_1)V_{j,k}| \leq \mu_1 r^{1/2}$.

Apart from a difference in normalization, the first assumption A0 coincides with the corresponding assumption in [9]. Further, [9] make an assumption that $|\sum_{k=1}^r U_{i,k}V_{j,k}| \leq \mu_1 r^{1/2}$. This is analogous to A1 (called also there A1), although it does not coincide with it. The two versions of assumption A1 coincide in the case of equal singular values $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_r$. In the general case, they do not coincide but neither one implies the other. For instance, in case the vectors $(U_{i,1}, \ldots, U_{i,r})$ and $(V_{j,1}, \ldots, V_{j,r})$ are collinear, our condition is weaker, and is implied by the assumption of [9].

The incoherence condition is satisfied with high probability if $M = XY^T$ where the entries of $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ are i.i.d. zero mean uniformly bounded random variables, with incoherence parameter scaling as $(r\log n)^{1/2}$ (this follows from a standard Chernoff bound argument). It is also satisfied with high probability if $M = U\Sigma V^T$ with U and V uniformly random matrices with $U^TU = m\mathbf{I}$ and $V^TV = n\mathbf{I}$, with incoherence parameter scaling as $(r\log n)^{1/2}$. Here and below, $\mathbf{I}$ denotes the identity matrix.

D. Some Notations

In the following, whenever we write that a property A holds with high probability (w.h.p.), we mean that there exists a function $f(n) = f(n;\alpha,\mu_0,\mu_1,\Sigma_{\min},\Sigma_{\max})$ such that $\mathbb{P}(A) \geq 1 - f(n)$ and $f(n) \to 0$. Notice that we will actually prove precise probability bounds on our main results.

Probability is taken with respect to the uniformly random subset $E \subseteq [m]\times[n]$. Define $\epsilon \equiv |E|/\sqrt{mn}$. In the case $m = n$, ϵ corresponds to the average number of revealed entries per row or column. In practice, it is convenient to work with a model in which each entry is revealed independently with probability $\epsilon/\sqrt{mn}$. Elementary tail bounds on binomial random variables imply that (under the independent model) there exists a constant A such that, for all $\epsilon\sqrt{\alpha} \geq 1$

$$\mathbb{P}\Big\{|E| \in [n\epsilon\sqrt{\alpha} - A\sqrt{n\log n}, n\epsilon\sqrt{\alpha} + A\sqrt{n\log n}]\Big\} \geq 1 - \frac{1}{n^{10}}. \quad (6)$$

Since the success of the algorithm is a monotone function of $|E|$ (we can always "throw away" entries) any guarantee proved within one model holds within the other model as well if we allow for a vanishing shift in ϵ. This type of ensemble equivalence is standard and heavily used in random graph theory [6], [25]. Finally, we will use C, C', etc., to denote numerical constants.

Given a vector $x \in \mathbb{R}^n$, $||x||$ will denote its Euclidean norm. For a matrix $X \in \mathbb{R}^{n\times n'}$, $||X||_F$ is its Frobenius norm, and $||X||_2$ its operator norm (i.e., $||X||_2 = \sup_{u\neq 0}||Xu||/||u||$). The standard scalar product between vectors or matrices will sometimes be indicated by $\langle x,y\rangle$ or $\langle X,Y\rangle$, respectively. We shall use the standard combinatorics notation $[N] = \{1,2,\ldots,N\}$ to denote the set of first N integers.

E. Main Results

Notice that computing $\mathsf{P}_r(\widetilde{M}^E)$ only requires to find the first r singular vectors of a sparse matrix. Our main result establishes that this simple procedure achieves arbitrarily small relative root mean square error from $O(nr)$ revealed entries. We define the relative root mean square error as

$$\mathrm{RMSE} \equiv \left[\frac{1}{mnM_{\max}^2}||M - \mathsf{P}_r(\widetilde{M}^E)||_F^2\right]^{1/2}. \quad (7)$$

where we denote by $||A||_F$ the Frobenius norm of matrix A. Notice that the factor $(1/mn)$ corresponds to the usual normalization by the number of entries and the factor $(1/M_{\max}^2)$ corresponds to the maximum size of the matrix entries where M satisfies $|M_{ij}| \leq M_{\max}$ for all i and j.

Theorem 1.1: Assume M to be a rank r matrix of dimension $n\alpha \times n$ that satisfies $|M_{i,j}| \leq M_{\max}$ for all i,j. Then with probability larger than $1 - 1/n^3$

$$\frac{1}{mnM_{\max}^2}||M - \mathsf{P}_r(\widetilde{M}^E)||_F^2 \leq C\,\frac{\alpha^{3/2}rn}{|E|} \quad (8)$$

for some numerical constant C.

This theorem is proved in Section III.

Notice that the top r singular values and singular vectors of the sparse matrix $\widetilde{M}^E$ can be computed efficiently by subspace iteration [5]. Each iteration requires $O(|E|r)$ operations. As proved in Section III, the $(r+1)$th singular value is smaller than one half of the rth one. As a consequence, subspace iteration converges exponentially. A simple calculation shows that $O(\log n)$ iterations are sufficient to ensure the error bound mentioned.

The "cleaning" step in the above pseudocode improves systematically over $\mathsf{P}_r(\widetilde{M}^E)$ and, for large enough $|E|$, reconstructs M exactly. Let M be a $m\times n$ matrix satisfying the incoherence property defined in Section I-C and let Σ_k denote the kth singular value of M scaled by $1/\sqrt{mn}$.

Theorem 1.2: Assume M to be a rank r matrix that satisfies the incoherence conditions A0 and A1 with parameters (μ_0,μ_1). Let $\mu = \max\{\mu_0,\mu_1\}$. Further, assume $\Sigma_{\min} \leq \Sigma_1,\ldots,\Sigma_r \leq \Sigma_{\max}$. Then there exists a numerical constant C' such that, if

$$|E| \geq C'nr\sqrt{\alpha}\left(\frac{\Sigma_{\max}}{\Sigma_{\min}}\right)^2\max\Big\{\mu_0\log n,\, \mu^2r\sqrt{\alpha}\Big(\frac{\Sigma_{\max}}{\Sigma_{\min}}\Big)^4\Big\} \quad (9)$$

then the cleaning procedure in OPTSPACE converges, with probability larger than $1 - 1/n^3$, to the matrix M.

This theorem is proved in Section VI. The basic intuition is that, for $|E| \geq C'(\alpha)nr\,\max\{\log n, r\}$, $\mathsf{P}_r(\widetilde{M}^E)$ is so close to M that the cost function is well approximated by a quadratic function.

Theorem 1.1 is order optimal: the number of degrees of freedom in M is of order nr, without the same number of observations it is impossible to fix them. The extra $\log n$ factor in Theorem 1.2 is due to a coupon-collector effect [9], [20], [21]: it is necessary that E contains at least one entry per row and one per column and this happens only for $|E| \geq Cn\log n$. As a consequence, for a matrix M with bounded rank $r = O(1)$ and bounded condition number $(\Sigma_{\max}/\Sigma_{\min}) = O(1)$, Theorem 1.2 is optimal. It is suboptimal by a polylogarithmic factor for $r = O((\log n)^a)$.

The above guarantee only holds "up to numerical constants" independent of the matrix dimension or incoherence parameters. One might wonder how good OPTSPACE is in practice. While a detailed study is beyond the scope of this paper [23], in Fig. 2, we present the results of a numerical experiment for uniformly random matrices M. We plot the empirical *reconstruction rate* of OPTSPACE as a function of the average number of revealed entries per row. Here, we declare a matrix to be reconstructed if the relative error $||M - \widehat{M}||_F/||M||_F \leq 10^{-4}$. The *reconstruction rate* is the fraction of instances for which the matrix was reconstructed. For comparison, we also plot a lower bound obtained using the algorithm from [31].

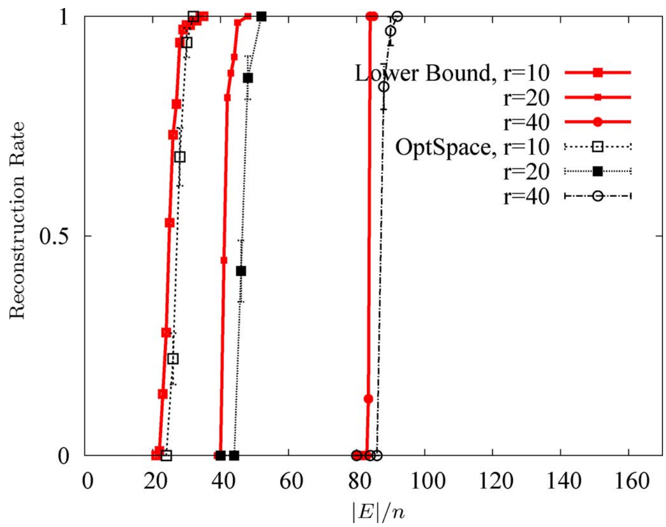

Fig. 2. Reconstruction rates for matrices with dimension $m = n = 500$ using OptSpace for different ranks. The solid curves are lower bounds for the probability of error in reconstructing a rank r matrix using $|E|$ entries revealed uniformly at random. This bound is proved in [31].

F. Related Work

Beyond collaborative filtering, low rank models are used for clustering, information retrieval, machine learning, and image processing. In [16], the NP-hard problem of finding a matrix of minimum rank satisfying a set of affine constraints was addressed through convex relaxation. This problem is analogous to the problem of finding the sparsest vector satisfying a set of affine constraints, which is at the heart of *compressed sensing* [10], [14]. The connection with compressed sensing was emphasized in [30], that provided performance guarantees under appropriate conditions on the constraints.

In the case of collaborative filtering, we are interested in finding a matrix M of minimum rank that matches the known entries $\{M_{ij} : (i,j) \in E\}$. Each known entry thus provides an affine constraint. Candès and Recht [9] introduced the incoherent model for M. Within this model, they proved that, if E is random, the convex relaxation correctly reconstructs M as long as $|E| \geq C\, r\, n^{6/5} \log n$. On the other hand, from a purely information theoretic point of view (i.e., disregarding algorithmic considerations), it is clear that $|E| = O(n\, r)$ observations should allow to reconstruct M with arbitrary precision. Indeed this point was raised in [9]. In [20], the authors proved through a nonalgorithmic argument that (under a probabilistic model) any relative RMSE error δ can be achieved with $|E| \geq C(\delta) r\, n$. Theorem 1.1 confirms this and shows that this objective is achieved by simple trimming plus SVD.

Theorem 1.1 can also be compared with a copious line of work in the theoretical computer science literature [1], [2], [18]. An important motivation in this context is the development of fast algorithms for low-rank approximation. In particular, Achlioptas and McSherry [2] prove a theorem analogous to 1.1, but holding only for $|E| \geq (8 \log n)^4 n$ (in the case of square matrices). A similar theorem follows also from earlier work by Seginer [32], and is also proved in [9]. In Theorem 1.1, we improve over these results by eliminating logarithmic factors in the error estimate as well as conditions on $|E|$. We refer to Section II for further discussion of this point.

The bound on $|E|$ for exact recovery provided by Theorem 1.2 improves upon the corresponding bound obtained in [9] for $r = O(n^{1/5})$. Note that in many practical applications, such as positioning or structure-from-motion [11], [34], r is known and is small (indeed it is comparable with the ambient dimension that is 3). In the regime $r = O(\log n)$, Theorem 1.2 is optimal upto polylogarithmic factors. The theorem is also suboptimal in the following regimes:

1. Large rank. Lower bounds on the minimal number of entries scale linearly in the rank rather than quadratically as in Theorem 1.2. As should be clear from Fig. 2 (and from more extensive simulations in [23]) this appears to be a weakness of our proof technique rather than of the algorithm.
2. Large condition number $(\Sigma_{\max}/\Sigma_{\min})$. Indeed our bound depends explicitly on this quantity, while this is not the case in [9]. This appears to be indeed a limitation of the singular value decomposition step. However, [23] discusses a simple modification of OptSpace, and shows empirically that this overcomes the problem.

Notice that the most complex component of our algorithm is the SVD in step 2. We were able to treat realistic data sets with $n \approx 10^5$. This must be compared with the $O(n^4)$ complexity of semidefinite programming [9].

Cai, Candès and Shen [7] recently proposed a low-complexity first-order procedure to solve the convex problem posed in [9]. Our spectral method is akin to a single step of this procedure, with the important novelty of the trimming step that improves its performance significantly. Our analysis techniques might provide a new tool for characterizing the convex relaxation as well. While this manuscript was under review, several efficient algorithms for solving the low-rank matrix completion problem were proposed. These include Accelerated Proximal Gradient (APG) algorithm [33], Fixed Point Continuation with Approximate SVD (FPCA) [26], Atomic Decomposition for Minimum Rank Approximation (ADMiRA) [24], Soft-Impute [27], Subspace Evolution and Transfer (SET) [13], and Singular Value Projection (SVP) [28].

A short account of our results was presented at the 2009 International Symposium on Information Theory [21]. While the present paper was under completion, Emmanuel Candès and Terence Tao posted online a preprint proving a theorem analogous to 1.2[12]. Their approach is substantially different from ours and based on a careful analysis of the convex relaxation approach studied earlier in [9]. Their result guarantees success with high probability for $|E| \geq C(\mu) nr(\log n)^6$. Apart from the difference in the underlying algorithm, the two results are not directly comparable: 1) the scaling of the number of observations with the problem dimensions provided by Theorem 1.2 is superior at small rank, but not at large rank; 2) the estimate in Theorem 1.2 is sensitive to the condition number of M, while the result of [9] assumes "strong incoherence" conditions for the factors U, V.

G. Open Problems and Future Directions

It is worth pointing out some limitations of our results, and interesting research directions:

1) Optimal RMSE With $O(n)$ Entries: Numerical simulations with the OPTSPACE algorithm suggest that the RMSE decays much faster with the number of observations per degree of freedom ($|E|/nr$), than indicated by (8). This improved behavior is a consequence of the cleaning step in the algorithm. It would be important to characterize the decay of RMSE with ($|E|/nr$).

2) Threshold for Exact Completion: As pointed out, Theorem 1.2 is order optimal in the matrix dimensions m, n for bounded rank, namely $r = O(1)$ and bounded condition number of M, $(\Sigma_{\max}/\Sigma_{\min}) = O(1)$. It would nevertheless be useful to derive quantitatively sharp estimates in this regime. A systematic numerical study was initiated in [20]. It appears that available theoretical estimates (including the recent ones in [12]) are suboptimal for larger values of the rank. We expect that our arguments can be strengthened to prove exact reconstruction for $|E| \geq C'(\alpha, \mu, \Sigma_{\max}/\Sigma_{\min}) nr \log n$ for all values of r.

3) Noisy Observation: In most applications, it is not realistic to assume that the matrix M is exactly of low rank, and that its entries are observed with absolute precision. The robustness of matrix completion under noisy observations was recently studied in [8], [22], but several questions remain open.

4) More General Models: The model studied here and introduced in [9] presents obvious limitations. In applications to collaborative filtering, the subset of observed entries E is far from uniformly random. A recent paper [31] investigates the uniqueness of the solution of the matrix completion problem for general sets E. In applications to fast low-rank approximation, it would be desirable to consider nonincoherent matrices as well (as in [2]).

Finally, in linear algebra applications, one has the option of choosing the set of sampled entries E. It would be interesting to understand how this choice could be optimized.

II. On the Spectrum of Sparse Matrices and the Role of Trimming

The trimming step of the OPTSPACE algorithm is somewhat counter-intuitive in that we seem to be wasting information. In this section we want to clarify its role through a simple example. Before describing the example, let us stress once again two facts: *i)* in the last step of our the algorithm, the trimmed entries are actually incorporated in the cost function and hence the full information is exploited; *ii)* trimming is not the only way to treat over-represented rows/columns in M^E, and probably not the optimal one. One might for instance rescale the entries of such rows/columns. We stick to trimming because we can prove it actually works.

Let us now turn to the example. Assume, for the sake of simplicity $m = n$, and M to be the rank one matrix such that $M_{ij} = 1$ for all 1. Within the independent model, the matrix M^E has i.i.d. entries, with distribution $\mathrm{Bernoulli}(\epsilon/n)$. The number of nonzero entries in a column is $\mathrm{Binomial}(n, \epsilon/n)$ and is independent for different columns. It is not hard to realize that the column with the largest number of entries has more than $C \log n/\log\log n$ entries, with positive probability (this probability can be made as large as we want by reducing C). Let i be the index of this column, and consider the test vector $\underline{e}^{(i)}$ that has the ith entry equal to 1 and all the others equal to 0. By computing $\|M^E \underline{e}^{(i)}\|$, we conclude that the largest singular value of M^E is at least $\sqrt{C\log n/\log\log n}$. In particular, this is very different from the largest singular value of $\mathbb{E}\{M^E\}$ which is ϵ. This suggest that approximating M with the $\mathsf{P}_r(M^E)$ leads to a large error.

Reading the proof of Theorem 1.1 indeed confirms this expectation, as the proof requires showing that the eigenvalues of M are close to the ones of $\mathbb{E}\{\widetilde{M}^E\}$ (notice that we have to assume trimming in this case). Following arguments similar to the ones in the proof, it is indeed possible to show that, in our example

$$\frac{1}{n^2}\,\|M - \mathsf{P}_r(M^E)\|_F^2 \geq C'(\epsilon)\frac{\log n}{\log\log n}. \tag{10}$$

It is clear that this phenomenon is indeed general, as illustrated by Fig. 1. Also, the phenomenon is more severe in real data sets than in the present model, where each entry is revealed independently.

To summarize, Theorem 1.1 simply does not hold without trimming, or a similar procedure to normalize rows/columns of M^E. Trimming allows to overcome the above phenomenon by setting to 0 over-represented rows/columns. On the other hand, this makes the analysis technically nontrivial. Indeed, while M^E is a matrix with independent (but not identically distributed) entries, this is not the case for $\widetilde{M}^E$. As a consequence we cannot rely on standard concentration-of-measure results.

III. Proof of Theorem 1.1 and Technical Results

As explained in the previous section, the crucial idea is to consider the singular value decomposition of the trimmed matrix $\widetilde{M}^E$ instead of the original matrix M^E, as in (3). We shall then redefine $\{\sigma_i\}$, $\{x_i\}$, $\{y_i\}$, by letting

$$\widetilde{M}^E = \sum_{i=1}^{\min(m,n)} \sigma_i x_i y_i^T\,. \tag{11}$$

Here $\|x_i\| = \|y_i\| = 1$, $x_i^T x_j = y_i^T y_j = 0$ for $i \neq j$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Our key technical result is that, apart from a trivial rescaling, these singular values are close to the ones of the full matrix M. Recall here that $\epsilon \equiv |E|/\sqrt{mn}$.

Lemma 3.1: There exists a numerical constant $C > 0$ such that, with probability larger than $1 - 1/n^3$

$$\left|\frac{\sigma_q}{\epsilon} - \Sigma_q\right| \leq C M_{\max}\sqrt{\frac{\alpha}{\epsilon}} \tag{12}$$

where it is understood that $\Sigma_q = 0$ for $q > r$.

This result generalizes a celebrated bound on the second eigenvalue of random graphs [17], [19] and is illustrated in Fig. 1: the spectrum of $\widetilde{M}^E$ clearly reveals the rank-3 structure of M.

As shown in Section V, Lemma 3.1 is a direct consequence of the following estimate.

Lemma 3.2: There exists a numerical constant $C > 0$ such that, with probability larger than $1 - 1/n^3$

$$\left\|\frac{\epsilon}{\sqrt{mn}}M - \widetilde{M}^E\right\|_2 \leq C M_{\max}\sqrt{\alpha\epsilon}. \tag{13}$$

The proof of this lemma is given in Section IV.

We will now prove Theorem 1.1.

Proof: (Theorem 1.1) By triangle inequality

$$\begin{aligned}\left|\left|M-\mathsf{P}_r(\widetilde{M}^E)\right|\right|_2 &\le \left|\left|\frac{\sqrt{mn}}{\epsilon}\widetilde{M}^E-\mathsf{P}_r(\widetilde{M}^E)\right|\right|_2+\left|\left|M-\frac{\sqrt{mn}}{\epsilon}\widetilde{M}^E\right|\right|_2\\ &\le \sqrt{mn}\sigma_{r+1}/\epsilon + CM_{\max}\sqrt{\alpha mn}/\sqrt{\epsilon}\\ &\le 2CM_{\max}\sqrt{\frac{\alpha mn}{\epsilon}}\end{aligned}$$

where we used Lemma 3.2 for the second inequality and Lemma 3.1 for the last inequality. Now, for any matrix A of rank at most $2r$, $||A||_F \le \sqrt{2r}||A||_2$, whence

$$\begin{aligned}\frac{1}{\sqrt{mn}}\left|\left|M-\mathsf{P}_r(\widetilde{M}^E)\right|\right|_F &\le \frac{\sqrt{2r}}{\sqrt{mn}}\left|\left|M-\mathsf{P}_r(\widetilde{M}^E)\right|\right|_2\\ &\le C'M_{\max}\sqrt{\frac{\alpha r}{\epsilon}}.\end{aligned}$$

The result follows by using $|E| = \epsilon\sqrt{mn}$. □

IV. Proof of Lemma 3.2

We want to show that $|x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M)y| \le CM_{\max}\sqrt{\alpha\epsilon}$ for each $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ such that $||x|| = ||y|| = 1$. Our basic strategy (inspired by [17]) will be the following.

(1) Reduce to x, y belonging to discrete sets T_m, T_n.

(2) Bound the contribution of light couples by applying union bound to these discretized sets, with a large deviation estimate on the random variable Z, defined as $Z \equiv \sum_L x_i\widetilde{M}^E_{i,j}y_j - \frac{\epsilon}{\sqrt{mn}}x^TMy$.

(3) Bound the contribution of heavy couples using a bound on the discrepancy of corresponding the graph.

The technical challenge is that a worst-case bound on the tail probability of Z is not good enough, and we must keep track of its dependence on x and y. The definition of *light* and *heavy couples* is provided in the following section.

A. Discretization

We define

$$T_n = \left\{x \in \left\{\frac{\Delta}{\sqrt{n}}\mathbb{Z}\right\}^n : ||x|| \le 1\right\}.$$

Notice that $T_n \subseteq S_n \equiv \{x \in \mathbb{R}^n : ||x|| \le 1\}$. The next remark is proved in [17], [19], and relates the original problem to the discretized one.

Remark 4.1: Let $R \in \mathbb{R}^{m\times n}$ be a matrix. If $|x^TRy| \le B$ for all $x \in T_m$ and $y \in T_n$, then $|x'^TRy'| \le (1-\Delta)^{-2}B$ for all $x' \in S_m$ and $y' \in S_n$.

Hence, it is enough to show that, with high probability, $|x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M)y| \le CM_{\max}\sqrt{\alpha\epsilon}$ for all $x \in T_m$ and $y \in T_n$.

A naive approach would be to apply concentration inequalities directly to the random variable $x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M)y$. This fails because the vectors x, y can contain entries that are much larger than the typical size $O(n^{-1/2})$. We thus separate two contributions. The first contribution is due to *light couples* $L \subseteq [m]\times[n]$, defined as

$$L = \left\{(i,j) : |x_iM_{ij}y_j| \le M_{\max}\left(\frac{\epsilon}{mn}\right)^{1/2}\right\}.$$

The second contribution is due to its complement $\overline{L}$, which we call *heavy couples*. We have

$$\begin{aligned}\left|x^T\left(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M\right)y\right| &\le \left|\sum_{(i,j)\in L} x_i\widetilde{M}^E_{ij}y_j - \frac{\epsilon}{\sqrt{mn}}x^TMy\right|\\ &\quad + \left|\sum_{(i,j)\in\overline{L}} x_i\widetilde{M}^E_{ij}y_j\right|. \qquad (14)\end{aligned}$$

In the next two subsections, we will prove that both contributions are upper bounded by $CM_{\max}\sqrt{\alpha\epsilon}$ for all $x \in T_m$, $y \in T_n$. Applying Remark 4.1 to $|x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M)y|$, this proves the thesis.

B. Bounding the Contribution of Light Couples

Let us define the subset of row and column indices which have not been trimmed as $\mathcal{A}_l$ and $\mathcal{A}_r$

$$\begin{aligned}\mathcal{A}_l &= \{i \in [m] : \deg(i) \le \frac{2\epsilon}{\sqrt{\alpha}}\}\\ \mathcal{A}_r &= \{j \in [n] : \deg(j) \le 2\epsilon\sqrt{\alpha}\}\end{aligned}$$

where $\deg(\cdot)$ denotes the degree (number of revealed entries) of a row or a column. In following the proof, it might be convenient to keep in mind the bipartite graph with left vertices corresponding to rows of M, right vertices to its columns, and edges to the elements of E. The subscripts l and r in $\mathcal{A}_l$ and $\mathcal{A}_r$ refer to the left and right vertices.

Notice that $\mathcal{A} = (\mathcal{A}_l, \mathcal{A}_r)$ is a function of the random set E. It is easy to get a rough estimate of the sizes of $\mathcal{A}_l, \mathcal{A}_r$.

Remark 4.2: There exists C_1 and C_2 depending only on α such that, with probability larger than $1 - 1/n^4$, $|\mathcal{A}_l| \ge m - \max\{e^{-C_1\epsilon}m, C_2\alpha\}$, and $|\mathcal{A}_r| \ge n - \max\{e^{-C_1\epsilon}n, C_2\}$.

For the proof of this claim, we refer to Appendix A. For any $E \subseteq [m]\times[n]$ and $A = (A_l, A_r)$ with $A_l \subseteq [m]$, $A_r \subseteq [n]$, we define $M^{E,A}$ by setting to zero the entries of M that are not in E, those whose row index is not in A_l, and those whose column index not in A_r. Consider the event [see (15), shown at the bottom of the next page], where it is understood that x and y belong, respectively, to T_m and T_n. Note that $\widetilde{M}^E = M^{E,\mathcal{A}}$, and, hence, we want to bound $\mathbb{P}\{\mathcal{H}(E,\mathcal{A})\}$. We proceed as follows:

$$\begin{aligned}\mathbb{P}\{\mathcal{H}(E,\mathcal{A})\} &= \sum_A \mathbb{P}\{\mathcal{H}(E,A),\, \mathcal{A}=A\}\\ &\le \sum_{\substack{|A_l|\ge m(1-\delta),\\ |A_r|\ge n(1-\delta)}} \mathbb{P}\{\mathcal{H}(E,A),\, \mathcal{A}=A\} + \frac{1}{n^4}\\ &\le 2^{(n+m)H(\delta)} \max_{\substack{|A_l|\ge m(1-\delta),\\ |A_r|\ge n(1-\delta)}} \mathbb{P}\{\mathcal{H}(E;A)\} + \frac{1}{n^4}\end{aligned} \qquad (16)$$

with $\delta \equiv \max\{e^{-C_1\epsilon}, C_2/n\}$ and $H(x)$ the binary entropy function.

We are now left with the task of bounding $\mathbb{P}\{\mathcal{H}(E;A)\}$ uniformly over A where $\mathcal{H}$ is defined as in (15). The key step consists in proving the following tail estimate

Lemma 4.3: Let $x \in S_m$, $y \in S_n$, $Z = \sum_{(i,j)\in L} x_i M_{ij}^{E,A} y_j - \frac{\epsilon}{\sqrt{mn}} x^T M y$, and assume $|A_l| \geq m(1-\delta)$, $|A_r| \geq n(1-\delta)$ with δ small enough. Then

$$\mathbb{P}\left(Z > LM_{\max}\sqrt{\epsilon}\right) \leq \exp\left\{-\frac{\sqrt{\alpha}(L-3)n}{2}\right\}.$$

Proof: We begin by bounding the mean of Z as follows (for the proof of this statement we refer to Appendix B).

Remark 4.4: $|\mathbb{E}[Z]| \leq 2M_{\max}\sqrt{\epsilon}$.

For $A = (A_l, A_r)$, let M^A be the matrix obtained from M by setting to zero those entries whose row index is not in A_l, and those whose column index not in A_r. Define the potential contribution of the light couples a_{ij} and independent random variables Z_{ij} as

$$a_{ij} = \begin{cases} x_i M_{ij}^A y_j, & \text{if } |x_i M_{ij}^A y_j| \leq M_{\max}(\epsilon/mn)^{1/2} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_{ij} = \begin{cases} a_{i,j}, & \text{w.p. } \epsilon/\sqrt{mn} \\ 0 & \text{w.p. } 1-\epsilon/\sqrt{mn}, \end{cases}.$$

Let $Z_1 = \sum_{i,j} Z_{ij}$ so that $Z = Z_1 - \frac{\epsilon}{\sqrt{mn}} x^T M y$. Note that $\sum_{i,j} a_{ij}^2 \leq \sum_{i,j}\left(x_i M_{ij}^A y_j\right)^2 \leq M_{\max}^2$. Fix $\lambda = \sqrt{mn}/2M_{\max}\sqrt{\epsilon}$ so that $|\lambda a_{i,j}| \leq 1/2$, where $e^{\lambda a_{ij}} - 1 \leq \lambda a_{ij} + 2(\lambda a_{ij})^2$. It then follows that

$$\begin{aligned}
&\mathbb{E}[e^{\lambda Z}] \\
&\leq \exp\left\{\frac{\epsilon}{\sqrt{mn}}\Big(\sum_{i,j}\lambda a_{i,j} + 2\sum_{i,j}(\lambda a_{i,j})^2\Big) - \frac{\lambda\,\epsilon}{\sqrt{mn}} x^T M y\right\} \\
&\leq \exp\left\{\lambda \mathbb{E}[Z] + \frac{\sqrt{mn}}{2}\right\}.
\end{aligned}$$

The thesis follows by Chernoff bound $\mathbb{P}(Z > a) \leq e^{-\lambda a}\mathbb{E}[e^{\lambda Z}]$ after simple calculus. □

Note that $\mathbb{P}\left(-Z > LM_{\max}\sqrt{\epsilon}\right)$ can also be bounded analogously. We can now finish the upper bound on the light couples contribution. Consider the error event (15). A simple volume calculation shows that $|T_m| \leq (10/\Delta)^m$. We can apply union bound over T_m and T_n to (16) to obtain the equation shown at the bottom of the page. Hence, assuming $\alpha \geq 1$, there exists a numerical constant C' such that, for $C > C'\sqrt{\alpha}$, the first term is of order $e^{-\Theta(n)}$, and this finishes the proof.

C. Bounding the Contribution of Heavy Couples

Let Q be an $m \times n$ matrix with $Q_{ij} = 1$ if $(i,j) \in E$ and $i \in \mathcal{A}_l$, $j \in \mathcal{A}_r$ (i.e., entry (i,j) is not trimmed by our algorithm), and $Q_{ij} = 0$ otherwise. Since $|M_{ij}| \leq M_{\max}$, the heavy couples satisfy $|x_i y_j| \geq \sqrt{\epsilon/mn}$. We then have

$$\begin{aligned}
\left|\sum_{(i,j)\in\overline{L}} x_i \widetilde{M}_{ij}^E y_j\right| &\leq M_{\max}\sum_{(i,j)\in\overline{L}} Q_{ij}|x_i y_j| \\
&\leq M_{\max}\sum_{\substack{(i,j)\in E:\\ |x_i y_j|\geq\sqrt{\epsilon/mn}}} Q_{ij}|x_i y_j|.
\end{aligned}$$

Notice that Q is the adjacency matrix of a random bipartite graph with vertex sets $[m]$ and $[n]$ and maximum degree bounded by $2\epsilon\max(\alpha^{1/2}, \alpha^{-1/2})$. The following remark strengthens a result of [19].

Remark 4.5: Given vectors x, y, let $\overline{L}' = \{(i,j) : |x_i y_j| \geq C\sqrt{\epsilon/mn}\}$. Then there exist a constant C' such that, $\sum_{(i,j)\in\overline{L}'} Q_{ij}|x_i y_j| \leq C'(\sqrt{\alpha} + \frac{1}{\sqrt{\alpha}})\sqrt{\epsilon}$, for all $x \in T_m$, $y \in T_n$ with probability larger than $1 - 1/2n^3$.

For the reader's convenience, a proof of this fact is presented in Appendix C. The analogous result in [19] (for the adjacency matrix of a nonbipartite graph) is proved to hold only with probability larger than $1 - e^{-C\epsilon}$. The stronger statement quoted here can be proved using concentration of measure inequalities. The last remark implies that for all $x \in T_m$, $y \in T_n$, and $\alpha \geq 1$, the contribution of heavy couples is bounded by $CM_{\max}\sqrt{\alpha\epsilon}$ for some numerical constant C with probability larger than $1 - 1/2n^3$.

$$\mathcal{H}(E,A) = \left\{\exists\, x, y : \left|\sum_{(i,j)\in L} x_i M_{ij}^{E,A} y_j - \frac{\epsilon}{\sqrt{mn}} x^T M y\right| > CM_{\max}\sqrt{\alpha\epsilon}\right\} \tag{15}$$

$$\begin{aligned}
\mathbb{P}\{\mathcal{H}(E,\mathcal{A})\} &\leq 2\cdot 2^{(n+m)H(\delta)}\cdot\left(\frac{20}{\Delta}\right)^{n+m} e^{-\frac{(C-3)\sqrt{\alpha}n}{2}} + \frac{1}{n^4} \\
&\leq \exp\left\{\log 2 + (1+\alpha)\left(H(\delta)\log 2 + \log(20/\Delta)\right)n - \frac{(C-3)\sqrt{\alpha}n}{2}\right\} + \frac{1}{n^4}
\end{aligned}$$

V. Proof of Lemma 3.1

Recall the variational principle for the singular values

$$\sigma_q = \min_{H,\dim(H)=n-q+1} \max_{y\in H,\|y\|=1} \|\widetilde{M}^E y\| \tag{17}$$

$$= \max_{H,\dim(H)=q} \min_{y\in H,\|y\|=1} \|\widetilde{M}^E y\| . \tag{18}$$

Here, H is understood to be a linear subspace of $\mathbb{R}^n$.

Using (17) with H the orthogonal complement of $\mathrm{span}(v_1,\dots,v_{q-1})$, we have, by Lemma 3.2

$$\begin{aligned}\sigma_q &\le \max_{y\in H,\|y\|=1} \left|\left|\widetilde{M}^E y\right|\right| \\ &\le \frac{\epsilon}{\sqrt{mn}}\left(\max_{y\in H,\|y\|=1}\left|\left|My\right|\right|\right) \\ &\quad + \max_{y\in H,\|y\|=\|x\|=1}\left|x^T\left(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}}M\right)y\right| \\ &\le \epsilon\Sigma_q + CM_{\max}\sqrt{\alpha\epsilon}\end{aligned}$$

Recall that U and V are normalized such that $\sqrt{mn}\Sigma_q$ is the qth singular value of M. The lower bound is proved analogously, by using (18) with $H = \mathrm{span}(v_1,\dots,v_q)$.

VI. Minimization on Grassmann Manifolds and Proof of Theorem 1.2

The function $F(X,Y)$ defined in (4) and to be minimized in the last part of the algorithm can naturally be viewed as defined on Grassmann manifolds. Here, we recall from [15] a few important facts on the geometry of Grassmann manifold and related optimization algorithms. We then prove Theorem 1.2. Technical calculations are deferred to Sections VII, VIII, and to the appendices.

We recall that, for the proof of Theorem 1.2, it is assumed that $\Sigma_{\min}$, $\Sigma_{\max}$ are bounded away from 0 and ∞. Numerical constants are denoted by C, C' etc. Finally, throughout this section, we use the notation $X^{(i)} \in \mathbb{R}^r$ to refer to the ith row of the matrix $X \in \mathbb{R}^{m\times r}$ or $X \in \mathbb{R}^{n\times r}$.

A. Geometry of the Grassmann Manifold

Denote by $\mathsf{O}(d)$ the orthogonal group of $d\times d$ matrices. The Grassmann manifold is defined as the quotient $\mathsf{G}(n,r) \simeq \mathsf{O}(n)/\mathsf{O}(r)\times\mathsf{O}(n-r)$. In other words, a point on the manifold is the equivalence class of an $n\times r$ orthogonal matrix A

$$[A] = \{AQ : Q \in \mathsf{O}(r)\} . \tag{19}$$

For consistency with the rest of the paper, we will assume the normalization $A^T A = n\,\mathbf{I}$ where $\mathbf{I}$ denotes the identity matrix. To represent a point in $\mathsf{G}(n,r)$, we will use an explicit representative of this form. More abstractly, $\mathsf{G}(n,r)$ is the manifold of r-dimensional subspaces of $\mathbb{R}^n$.

It is easy to see that $F(X,Y)$ depends on the matrices X, Y only through their equivalence classes $[X]$, $[Y]$. We will, therefore, interpret it as a function defined on the manifold $\mathsf{M}(m,n) \equiv \mathsf{G}(m,r)\times\mathsf{G}(n,r)$

$$F : \mathsf{M}(m,n) \to \mathbb{R} \tag{20}$$

$$([X],[Y]) \mapsto F(X,Y) . \tag{21}$$

In the following, a point in this manifold will be represented as a pair $\mathbf{x} = (X,Y)$, with X an $n\times r$ orthogonal matrix and Y an $m\times r$ orthogonal matrix. Boldface symbols will be reserved for elements of $\mathsf{M}(m,n)$ or of its tangent space, and we shall use $\mathbf{u} = (U,V)$ for the point corresponding to the matrix $M = U\Sigma V^T$ to be reconstructed.

Given $\mathbf{x} = (X,Y) \in \mathsf{M}(m,n)$, the tangent space at $\mathbf{x}$ is denoted by $\mathsf{T}_{\mathbf{x}}$ and can be identified with the vector space of matrix pairs $\mathbf{w} = (W,Z)$, $W \in \mathbb{R}^{m\times r}$, $Z \in \mathbb{R}^{n\times r}$ such that $W^T X = Z^T Y = 0$. The "canonical" Riemann metric on the Grassmann manifold corresponds to the usual scalar product $\langle W, W'\rangle \equiv \mathrm{Tr}(W^T W')$. The induced scalar product on $\mathsf{T}_{\mathbf{x}}$ between $\mathbf{w} = (W,Z)$ and $\mathbf{w}' = (W',Z')$ is $\langle \mathbf{w},\mathbf{w}'\rangle = \langle W,W'\rangle + \langle Z,Z'\rangle$.

This metric induces a canonical notion of distance on $\mathsf{M}(m,n)$ which we denote by $d(\mathbf{x}_1,\mathbf{x}_2)$ (geodesic or arc-length distance). If $\mathbf{x}_1 = (X_1,Y_1)$ and $\mathbf{x}_2 = (X_2,Y_2)$ then

$$d(\mathbf{x}_1,\mathbf{x}_2) \equiv \sqrt{d(X_1,X_2)^2 + d(Y_1,Y_2)^2} \tag{22}$$

where the arc-length distances $d(X_1,X_2)$, $d(Y_1,Y_2)$ on the Grassmann manifold can be defined explicitly as follows. Let $\cos\theta = (\cos\theta_1,\dots,\cos\theta_r)$, $\theta_i \in [-\pi/2,\pi/2]$ be the singular values of $X_1^T X_2/m$. Then

$$d(X_1,X_2) = \|\theta\|_2 . \tag{23}$$

The θ_i's are called the "principal angles" between the subspaces spanned by the columns of X_1 and X_2. It is useful to introduce two equivalent notions of distance

$$d_{\mathrm{c}}(X_1,X_2) = \frac{1}{\sqrt{n}} \min_{Q_1,Q_2\in\mathsf{O}(r)} \|X_1Q_1 - X_2Q_2\|_F \quad \text{(chordal distance)} \tag{24}$$

$$d_p(X_1,X_2) = \frac{1}{\sqrt{2n}}\|X_1X_1^T - X_2X_2^T\|_F \quad \text{(projection distance)}. \tag{25}$$

Notice that d_{c} and d_p do not depend on the specific representatives X_1, X_2, but only on the equivalence classes $[X_1]$ and $[X_2]$. Distances on $\mathsf{M}(m,n)$ are defined through Pythagorean theorem, e.g., $d_{\mathrm{c}}(\mathbf{x}_1,\mathbf{x}_2) = \sqrt{d_{\mathrm{c}}(X_1,X_2)^2 + d_{\mathrm{c}}(Y_1,Y_2)^2}$.

Remark 6.1: The geodesic, chordal and projection distance are equivalent, namely

$$\begin{aligned}\frac{1}{\pi}d(X_1,X_2) &\le \frac{1}{\sqrt{2}}\,d_{\mathrm{c}}(X_1,X_2) \le d_p(X_1,X_2) \\ &\le d_{\mathrm{c}}(X_1,X_2) \le d(X_1,X_2).\end{aligned} \tag{26}$$

For the reader's convenience, a proof of this fact is presented in Appendix D.

An important remark is that geodesics with respect to the canonical Riemann metric admit an explicit and efficiently computable form. Given $\mathbf{u} \in \mathsf{M}(m,n)$, $\mathbf{w} \in \mathsf{T}_{\mathbf{u}}$ the corresponding geodesic is a curve $t \mapsto \mathbf{x}(t)$, with $\mathbf{x}(t) = \mathbf{u} + \mathbf{w}t + O(t^2)$ which minimizes arc-length. If $\mathbf{u} = (U,V)$ and $\mathbf{w} = (W,Z)$ then $\mathbf{x}(t) = (X(t),Y(t))$ where $X(t)$ can be expressed in terms of the singular value decomposition $W = L\Theta R^T$ [15]

$$X(t) = UR\cos(\Theta t)R^T + L\sin(\Theta t)R^T \tag{27}$$

which can be evaluated in time of order $O(nr)$. An analogous expression holds for $Y(t)$.

B. Explicit Formulae for the Gradient

The gradient of F at $\mathbf{x}$ is the vector $\operatorname{grad} F(\mathbf{x}) \in \mathsf{T}_{\mathbf{x}}$ such that, for any smooth curve $t \mapsto \mathbf{x}(t) \in \mathsf{M}(m,n)$ with $\mathbf{x}(t) = \mathbf{x} + \mathbf{w}\,t + O(t^2)$, one has

$$F(\mathbf{x}(t)) = F(\mathbf{x}) + \langle \operatorname{grad} F(\mathbf{x}), \mathbf{w}\rangle\, t + O(t^2)\,. \tag{28}$$

In order to write an explicit representation of the gradient of our cost function F, it is convenient to introduce the projector operator

$$\mathcal{P}_E(M)_{ij} = \begin{cases} M_{ij}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

The two components of the gradient are then

$$\operatorname{grad} F(\mathbf{x})_X = \mathcal{P}_E(XSY^T - M)YS^T - XQ_X \tag{30}$$
$$\operatorname{grad} F(\mathbf{x})_Y = \mathcal{P}_E(XSY^T - M)^T XS - YQ_Y \tag{31}$$

where S is the minimizer in (4) and $Q_X, Q_Y \in \mathbb{R}^{r\times r}$ are determined by the condition $\operatorname{grad} F(\mathbf{x}) \in \mathsf{T}_{\mathbf{x}}$. This yields

$$Q_X = \frac{1}{m} X^T \mathcal{P}_E(XSY^T - M)YS^T \tag{32}$$
$$Q_Y = \frac{1}{n} Y^T \mathcal{P}_E(XSY^T - M)^T XS\,. \tag{33}$$

C. Algorithm

At this point, the gradient descent algorithm is fully specified. It takes as input the factors of $\mathsf{P}_r(\widetilde{M}^E)$, to be denoted as $\mathbf{x}_0 = (X_0, Y_0)$, and minimizes a regularized cost function

$$\widetilde{F}(X,Y) = F(X,Y) + \rho\, G(X,Y) \tag{34}$$
$$\equiv F(X,Y) + \rho \sum_{i=1}^{m} G_1\left(\frac{\|X^{(i)}\|^2}{3\mu_0 r}\right) + \rho \sum_{j=1}^{n} G_1\left(\frac{\|Y^{(j)}\|^2}{3\mu_0 r}\right) \tag{35}$$

where $X^{(i)}$ denotes the ith column of X^T, and $Y^{(j)}$ the jth column of Y^T. The role of the regularization is to force $\mathbf{x}$ to remain incoherent during the execution of the algorithm.

$$G_1(z) = \begin{cases} 0, & \text{if } z \le 1 \\ e^{(z-1)^2} - 1, & \text{if } z \ge 1. \end{cases} \tag{36}$$

We will take $\rho = n\epsilon$. Notice that $G(X,Y)$ is again naturally defined on the Grassmann manifold, i.e., $G(X,Y) = G(XQ, YQ')$ for any $Q, Q' \in \mathsf{O}(r)$.

Let [see (37), shown at the bottom of the page]. We have $G(X,Y) = 0$ on $\mathcal{K}(3\mu_0)$. Notice that $\mathbf{u} \in \mathcal{K}(\mu_0)$ by the incoherence property. Also, by the following remark proved in Appendix D, we can assume that $\mathbf{x}_0 \in \mathcal{K}(3\mu_0)$.

Remark 6.2: Let $U, X \in \mathbb{R}^{n\times r}$ with $U^TU = X^TX = n\mathbf{I}$ and $U \in \mathcal{K}(\mu_0)$ and $d(X,U) \le \delta \le \frac{1}{16}$. Then there exists $X'' \in \mathbb{R}^{n\times r}$ such that $X''^T X'' = n\mathbf{I}$, $X'' \in \mathcal{K}(3\mu_0)$ and $d(X'', U) \le 4\delta$. Further, such an X'' can be computed in a time of $O(nr^2)$.

GRADIENT DECENT (matrix M^E, factors $\mathbf{x}_0$)

1: For $k = 0, 1, \ldots$ do:

2: Compute $\mathbf{w}_k = \operatorname{grad} \widetilde{F}(\mathbf{x}_k)$.

3: Let $t \mapsto \mathbf{x}_k(t)$ be the geodesic with $\mathbf{x}_k(t) = \mathbf{x}_k + \mathbf{w}_k t + O(t^2)$.

4: Minimize $t \mapsto \widetilde{F}(\mathbf{x}_k(t))$ for $t \ge 0$, subject to $d(\mathbf{x}_k(t), \mathbf{x}_0) \le \gamma$.

5: Set $\mathbf{x}_{k+1} = \mathbf{x}_k(t_k)$ where t_k is the minimum location.

6: End For.

In the above, γ must be set in such a way that $d(\mathbf{u}, \mathbf{x}_0) \le \gamma$. The next remark determines the correct scale.

Remark 6.3: Let $U, X \in \mathbb{R}^{m\times r}$ with $U^TU = X^TX = m\mathbf{I}$, $V, Y \in \mathbb{R}^{n\times r}$ with $V^TV = Y^TY = n\mathbf{I}$, and $M = U\Sigma V^T$, $\widehat{M} = XSY^T$ for $\Sigma = \operatorname{diag}(\Sigma_1, \ldots, \Sigma_r)$ and $S \in \mathbb{R}^{r\times r}$. If $\Sigma_1, \ldots, \Sigma_r \ge \Sigma_{\min}$, then

$$d_p(U,X) \le \frac{1}{\sqrt{2\alpha n}\Sigma_{\min}} \|M - \widehat{M}\|_F$$
$$d_p(V,Y) \le \frac{1}{\sqrt{2\alpha n}\Sigma_{\min}} \|M - \widehat{M}\|_F. \tag{38}$$

As a consequence of this remark and Theorem 1.1, we can assume that $d(\mathbf{u}, \mathbf{x}_0) \le C(\frac{\Sigma_{\max}}{\Sigma_{\min}}) \frac{\mu_1 r\sqrt{\alpha}}{\sqrt{\epsilon}}$. We shall then set $\gamma = C'(\frac{\Sigma_{\max}}{\Sigma_{\min}}) \frac{\mu_1 r\sqrt{\alpha}}{\sqrt{\epsilon}}$ (the value of C' is set in the course of the proof).

Before passing to the proof of Theorem 1.2, it is worth discussing a few important points concerning the gradient descent algorithm.

$$\mathcal{K}(\mu') \equiv \left\{(X,Y) \text{ such that } \|X^{(i)}\|^2 \le \mu' r,\ \|Y^{(j)}\|^2 \le \mu' r \text{ for all } i \in [m],\, j \in [n]\right\} \tag{37}$$

(i) The appropriate choice of γ might seem to pose a difficulty. In reality, this parameter is introduced only to simplify the proof. We will see that the constraint $d(\mathbf{x}_k(t), \mathbf{x}_0) \le \gamma$ is, with high probability, never saturated.

(ii) Indeed, the line minimization instruction 4 (which might appear complex to implement) can be replaced by a standard step selection procedure, such as the one in [3].

(iii) Similarly, there is no need to know the actual value of μ_0 in the regularization term. One can start with $\mu_0 = 1$ and then repeat the optimization doubling it at each step.

(iv) The Hessian of F can be computed explicitly as well. This opens the way to quadratically convergent minimization algorithms (e.g., the Newton method).

D. Proof of Theorem 1.2

The proof of Theorem 1.2 breaks down in two lemmas. The first one implies that, in a sufficiently small neighborhood of $\mathbf{u}$, the function $\mathbf{x} \mapsto F(\mathbf{x})$ is well approximated by a parabola.

Lemma 6.4: There exists numerical constants C_0, C_1, C_2 such that the following happens. Assume $\epsilon \ge C_0\mu_0\sqrt{\alpha}\, r \max\{\log n; \mu_0 r\sqrt{\alpha}(\Sigma_{\max}/\Sigma_{\min})^4\}$ and $\delta \le \Sigma_{\min}/C_0\Sigma_{\max}$. Then

$$C_1\sqrt{\alpha}\Sigma_{\min}^2\, d(\mathbf{x},\mathbf{u})^2 + C_1\sqrt{\alpha}\,||S-\Sigma||_F^2 \le \frac{1}{n\epsilon}\, F(\mathbf{x}) \le C_2\sqrt{\alpha}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2 \quad (39)$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \le \delta$, with probability at least $1 - 1/n^4$. Here, $S \in \mathbb{R}^{r\times r}$ is the matrix realizing the minimum in (4).

The second Lemma implies that, with high probability, $\mathbf{x} \mapsto F(\mathbf{x})$ does not have any other stationary point (apart from $\mathbf{u}$) within such a neighborhood.

Lemma 6.5: There exists numerical constants C_0, C such that the following happens. Assume

$$\epsilon \ge C_0\mu_0 r\sqrt{\alpha}(\Sigma_{\max}/\Sigma_{\min})^2 \max\{\log n; \mu_0 r\sqrt{\alpha}(\Sigma_{\max}/\Sigma_{\min})^4\}$$

and $\delta \le \Sigma_{\min}/C_0\Sigma_{\max}$. Then

$$||\mathrm{grad}\, \widetilde{F}(\mathbf{x})||^2 \ge C\, n\epsilon^2\Sigma_{\min}^4\, d(\mathbf{x},\mathbf{u})^2$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \le \delta$, with probability at least $1 - 1/n^4$.

We can now prove Theorem 1.2.

Proof: (Theorem 1.2) Let $\delta > 0$ be such that Lemma 6.4 and Lemma 6.5 are verified, and C_1, C_2 be defined as in Lemma 6.4. We further assume $\delta \le \sqrt{(e^{1/9}-1)/C_2}$. Take ϵ large enough such that, $d(\mathbf{u},\mathbf{x}_0) \le \min(1, (C_1/C_2)^{1/2}(\Sigma_{\min}/\Sigma_{\max}))\delta/10$. Further, set the algorithm parameter to $\gamma = \delta/4$.

We make the following claims:

1. $\mathbf{x}_k \in \mathcal{K}(4\mu_0)$ for all k.
   Indeed $\mathbf{x}_0 \in \mathcal{K}(3\mu_0)$, where $\widetilde{F}(\mathbf{x}_0) = F(\mathbf{x}_0) \le C_2\sqrt{\alpha}n\epsilon\Sigma_{\max}^2\,\delta^2$. The claim follows because $\widetilde{F}(\mathbf{x}_k)$ is nonincreasing and $\widetilde{F}(\mathbf{x}) \ge \rho\, G(X,Y) \ge n\epsilon\sqrt{\alpha}\Sigma_{\max}^2(e^{1/9}-1)$ for $\mathbf{x} \notin \mathcal{K}(4\mu_0)$, where we choose ρ to be $n\epsilon\sqrt{\alpha}\Sigma_{\max}^2$.
2. $d(\mathbf{x}_k,\mathbf{u}) \le \delta/10$ for all k.
   Since we set $\gamma = \delta/4$, by triangular inequality, we can assume to have $d(\mathbf{x}_k,\mathbf{u}) \le \delta/2$. Since $d(\mathbf{x}_0,\mathbf{u})^2 \le (C_1\Sigma_{\min}^2/C_2\Sigma_{\max}^2)(\delta/10)^2$, we have $\widetilde{F}(\mathbf{x}) \ge F(\mathbf{x}) \ge F(\mathbf{x}_0)$ for all $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{u}) \in [\delta/10, \delta]$. Since $\widetilde{F}(\mathbf{x}_k)$ is nonincreasing and $\widetilde{F}(\mathbf{x}_0) = F(\mathbf{x}_0)$, the claim follows.

Notice that, by the last observation, the constraint $d(\mathbf{x}_k(t),\mathbf{x}_0) \le \gamma$ is never saturated, and, therefore, our procedure is just gradient descent with exact line search. Therefore, by [3] this must converge to the unique stationary point of $\widetilde{F}$ in $\mathcal{K}(4\mu_0) \cap \{\mathbf{x} : d(\mathbf{x},\mathbf{u}) \le \delta/10\}$, which, by Lemma 6.5, is $\mathbf{u}$. □

VII. Proof of Lemma 6.4

A. A Random Graph Lemma

The following Lemma will be used several times in the following.

Lemma 7.1: There exist two numerical constants C_1, C_2 such that the following happens. If $\epsilon \ge C_1 \log n$ then, with probability larger than $1 - 1/n^5$

$$\sum_{(i,j)\in E} x_i y_j \le \frac{C_2\epsilon}{n\sqrt{\alpha}}\, ||x||_1||y||_1 + C_2\sqrt{\alpha\epsilon}||x||_2\, ||y||_2\,. \quad (40)$$

for all $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$.

Proof: Write $x_i = x_0 + x_i'$ where $\sum_i x_i' = 0$. Then

$$\sum_{(i,j)\in E} x_i y_j = x_0 \sum_{j\in[n]} \deg(j) y_j + \sum_{(i,j)\in E} x_i' y_j \quad (41)$$

where we recall that $\deg(j) = \{i \in [m] : \text{such that } (i,j) \in E\}$. Further $|x_0| = |\sum_i x_i/m| \le ||x||_1/m$. The first term is upper bounded by

$$x_0 \max_{j\in n} \deg(j)||y||_1 \le \max_{j\in n} \deg(j)||x||_1||y||_1/m\,. \quad (42)$$

For $\epsilon \ge C_1 \log n$, with probability larger than $1 - 1/2n^5$, the maximum degree is bounded by $(9/C_1)\sqrt{\alpha}\epsilon$ which is of same order as the average degree. Therefore, this term is at most $C_2\sqrt{\alpha}\epsilon||x||_1||y||_1/m$.

The second term is upper bounded by $C_2\sqrt{\alpha\epsilon}||x'||_2||y||_2$ using Theorem 1.1 in [19] or, equivalently, Theorem 3.1 in the case $r = 1$ and $M_{\max} = 1$. It can be shown to hold with probability larger than $1 - 1/2n^5$ with a large enough numerical constant C_2. The thesis follows because $||x'||_2 \le ||x||_2$. □

B. Preliminary Facts and Estimates

This subsection contains some remarks that will be useful in the proof of Lemma 6.5 as well.

Let $\mathbf{w} = (W,Z) \in \mathsf{T}_{\mathbf{u}}$, and $t \mapsto (X(t),Y(t))$ be the geodesic such that $(X(t),Y(t)) = (U,V) + (W,Z)t + O(t^2)$. By setting $(X,Y) = (X(1),Y(1))$, we establish a one-to-one correspondence between the points $\mathbf{x}$ as in the statement and a neighborhood of the origin in $\mathsf{T}_{\mathbf{u}}$. If we let $W = L\Theta R^T$ be the singular value decomposition of W (with $L^TL = m\mathbf{I}$ and $R^TR = \mathbf{I}$), the explicit expression for geodesics in (27) yields

$$X = U+\overline{W}\,,\qquad \overline{W} = UR(\cos\Theta-\mathbf{I})R^T+L\sin\Theta R^T\,. \quad (43)$$

An analogous expression can obviously be written for $Y = V + \overline{Z}$. Note that without loss of generality, $\Theta_{ii} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Indeed, let $X = UR\cos\Theta R^T + L\sin\Theta R^T$, where $\Theta_{ii} = k_i\pi + \Phi_{ii}$, $k_i \in \mathbb{Z}$ and $\Phi_{ii} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Then, $X = UR\cos\Phi DR^T + L\sin\Phi DR^T$, where D is a diagonal matrix with $D_{ii} = (-1)^{k_i}$. Then, as representative of a point on the Grassmann manifold $\mathsf{G}(m,r)$, X is equivalent to $XRDR^T = UR\cos\Phi R^T + L\sin\Phi R^T$. Hence, in the following we can assume, without loss of generality, that $\Theta_{ii} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Notice that, by the equivalence between chordal and canonical distance, Remark 6.1, we have

$$\frac{1}{m}\|\overline{W}\|_F^2 + \frac{1}{n}\|\overline{Z}\|_F^2 \le 2\,d(\mathbf{u},\mathbf{x})^2\,. \quad (44)$$

Remark 7.2: If $\mathbf{u} \in \mathcal{K}(\mu_0)$ and $\mathbf{x} \in \mathcal{K}(4\mu_0)$, then $(\overline{W},\overline{Z}) \in \mathcal{K}(10\mu_0)$ and $\mathbf{w} = (W,Z) \in \mathcal{K}(5\pi^2\mu_0/2)$.

Proof: The first fact follows from $\|\overline{W}^{(i)}\|^2 \le 2\|X^{(i)}\|^2 + 2\|U^{(i)}\|^2$. In order to prove $\mathbf{w} \in \mathcal{K}(5\pi^2\mu_0/2)$, we notice that

$$\|W^{(i)}\|^2 = \|\Theta L^{(i)}\|^2 \le \frac{\pi^2}{4}\|\sin\Theta L^{(i)}\|^2$$
$$\le \frac{\pi^2}{4}\|X^{(i)} - R\cos\Theta R^T U^{(i)}\|^2 \le \frac{\pi^2}{2}\Big(\|X^{(i)}\|^2 + \|U^{(i)}\|^2\Big)\,.$$

The claim follows by showing a similar bound for $\|Z^{(i)}\|^2$. □

We next prove a simple *a priori* estimate.

Remark 7.3: There exist numerical constants C_1, C_2 such that the following holds with probability larger than $1 - 1/n^5$. If $\epsilon \ge C_1\log n$, then for any $(X,Y) \in \mathcal{K}(\mu)$ and $S \in \mathbb{R}^{r\times r}$ [see (45), shown at the bottom of the page].

Proof: Using Lemma 7.1, $\sum_{(i,j)\in E}(XSY^T)_{ij}^2$ is upper bounded by

$$\begin{aligned}
&\sigma_{\max}(S)^2\sum_{a,b}\sum_{(i,j)\in E}X_{ia}^2Y_{jb}^2\\
&\quad\le \frac{C_2\epsilon}{n\sqrt{\alpha}}\sigma_{\max}(S)^2\sum_{i,j}\|X^{(i)}\|^2\|Y^{(j)}\|^2\\
&\qquad + C_2\sigma_{\max}(S)^2\sqrt{\alpha\epsilon}\Big(\sum_i\|X^{(i)}\|^4\Big)^{1/2}\Big(\sum_j\|Y^{(j)}\|^4\Big)^{1/2}\\
&\quad\le \frac{C_2\epsilon}{n\sqrt{\alpha}}\sigma_{\max}(S)^2\sum_{i,j}\|X^{(i)}\|^2\|Y^{(j)}\|^2\\
&\qquad +C_2\sigma_{\max}(S)^2\sqrt{\alpha\epsilon}\mu r\Big(\sum_i\|X^{(i)}\|^2\Big)^{1/2}\Big(\sum_j\|Y^{(j)}\|^2\Big)^{1/2}\\
&\quad\le C_2\|S\|_2^2\sqrt{\alpha}\,n\epsilon\Big(\frac{1}{m}\|X\|_F^2 + \frac{1}{n}\|Y\|_F^2\Big)^2\\
&\qquad + C_2\|S\|_2^2\,\alpha\mu r\,n\sqrt{\epsilon}\Big(\frac{1}{m}\|X\|_F^2 + \frac{1}{n}\|Y\|_F^2\Big)
\end{aligned}$$

where in the second step we used the incoherence condition. The last step follows from the inequalities $2ab \le \alpha(a/\alpha + b)^2$ and $2ab \le \sqrt{\alpha}(a^2/\alpha + b^2)$. □

C. The Proof

Proof: (Lemma 6.4) Denote by $S \in \mathbb{R}^{r\times r}$ the matrix realizing the minimum in (4). We will start by proving a lower bound on $F(\mathbf{x})$ of the form

$$\begin{aligned}\frac{1}{n\epsilon}F(\mathbf{x}) \ge C_1\sqrt{\alpha}\,\Sigma_{\min}^2\,d(\mathbf{x},\mathbf{u})^2 + C_1\sqrt{\alpha}\,\|S-\Sigma\|_F^2&\\
-C_1'\sqrt{\alpha}\Sigma_{\max}d(\mathbf{x},\mathbf{u})^2\|S-\Sigma\|_F&\quad (46)\end{aligned}$$

and an upper bound as in (39). Together, for $d(\mathbf{x},\mathbf{u}) \le \delta \le 1$, these imply $\|S-\Sigma\|_F^2 \le C\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2$, where the lower bound in (39) follows for $\delta \le \Sigma_{\min}/C_0\Sigma_{\max}$.

In order to prove the bound (46), we write $X = U + \overline{W}$, $Y = V + \overline{Z}$, and

$$\begin{aligned}&F(X,Y)\\
&\ = \frac{1}{2}\sum_{(i,j)\in E}(U(S-\Sigma)V^T + US\overline{Z}^T + \overline{W}SV^T + \overline{W}S\overline{Z}^T)_{ij}^2\\
&\ \ge \frac{1}{4}A^2 - \frac{1}{2}B^2\end{aligned}$$

where we used the inequality $(1/2)(a+b)^2 \ge (a^2/4) - (b^2/2)$, and defined

$$\begin{aligned}A^2 &\equiv \sum_{(i,j)\in E}(U(S-\Sigma)V^T + US\overline{Z}^T + \overline{W}SV^T)_{ij}^2\\
B^2 &\equiv \sum_{(i,j)\in E}(\overline{W}S\overline{Z}^T)_{ij}^2\,.\end{aligned}$$

Using Remark 7.3, and (44), we get

$$\begin{aligned}B^2 &\le C\sqrt{\alpha}n\epsilon\,\|S\|_2^2\left(d(\mathbf{x},\mathbf{u})^2 + \frac{\mu_0 r\sqrt{\alpha}}{\sqrt{\epsilon}}\right)d(\mathbf{x},\mathbf{u})^2\\
&\le 2C\sqrt{\alpha}n\epsilon\big(\Sigma_{\max}^2 + \|S-\Sigma\|_F^2\big)\left(\delta^2 + \frac{\mu_0 r\sqrt{\alpha}}{\sqrt{\epsilon}}\right)d(\mathbf{x},\mathbf{u})^2\end{aligned}$$

where the second inequality follows from the inequality $\sigma_{\max}(S)^2 \le 2\Sigma_{\max}^2 + 2\,\|S-\Sigma\|_F^2$.

$$\sum_{(i,j)\in E}(XSY^T)_{ij}^2 \le C_2\|S\|_2^2\sqrt{\alpha}\,n\epsilon\left(\frac{1}{m}\|X\|_F^2 + \frac{1}{n}\|Y\|_F^2\right)\left(\frac{1}{m}\|X\|_F^2 + \frac{1}{n}\|Y\|_F^2 + \frac{\mu r\sqrt{\alpha}}{\sqrt{\epsilon}}\right) \quad (45)$$

By Theorem 4.1 in [9], for $\epsilon \geq C\mu_0\sqrt{\alpha}\, r\log n$, we have $A^2 \geq (1/2)\mathbb{E}\{A^2\}$ with probability larger than $1 - 1/n^5$, for all $A \in \{UY^T + XV^T : X \in \mathbb{R}^{m\times r}, Y \in \mathbb{R}^{n\times r}\}$. Further

$$\begin{aligned}
&\mathbb{E}\{A^2\}\\
&= \frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)V^T + US\overline{Z}^T + \overline{W}SV^T\|_F^2\\
&= \frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)V^T\|_F^2 + \frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2\\
&\quad + \frac{\epsilon}{\sqrt{mn}}\|\overline{W}SV^T\|_F^2\\
&\quad + \frac{2\epsilon}{\sqrt{mn}}\langle US\overline{Z}^T, \overline{W}SV^T\rangle + \frac{2\epsilon}{\sqrt{mn}}\langle U(S-\Sigma)V^T, \overline{W}SV^T\rangle\\
&\quad + \frac{2\epsilon}{\sqrt{mn}}\langle US\overline{Z}^T, U(S-\Sigma)V^T\rangle.
\end{aligned}$$

Let us call the absolute value of the six terms on the right hand side $E_1, \dots E_6$. A simple calculation yields

$$E_1 = n\epsilon\sqrt{\alpha}\|S-\Sigma\|_F^2 \tag{47}$$

$$\begin{aligned}
E_2 + E_3 &\geq n\epsilon\sqrt{\alpha}\sigma_{\min}(S)^2\Big(\frac{1}{m}\|\overline{W}\|_F^2 + \frac{1}{n}\|\overline{Z}\|_F^2\Big)\\
&\geq C'\sigma_{\min}(S)^2 n\epsilon\sqrt{\alpha}d(\mathbf{x},\mathbf{u})^2\,.
\end{aligned} \tag{48}$$

The absolute value of the fourth term can be written as

$$\begin{aligned}
E_4 &= \frac{2\epsilon}{n\sqrt{\alpha}}|\langle US\overline{Z}^T, \overline{W}SV^T\rangle|\\
&\leq \frac{2\epsilon}{n\sqrt{\alpha}}\,\sigma_{\max}(S)^2\|\overline{W}^TU\|_F\|V^T\overline{Z}\|_F\\
&\leq \frac{2\epsilon\alpha}{n\sqrt{\alpha}}\sigma_{\max}(S)^2\Big(\frac{1}{\alpha^2}\|\overline{W}^TU\|_F^2 + \|V^T\overline{Z}\|_F^2\Big)\,.
\end{aligned}$$

In order proceed, consider (43). Since by tangency condition $U^TL = 0$, we have $U^T\overline{W} = mR(\cos\Theta - 1)R^T$, where

$$\begin{aligned}
\|U^T\overline{W}\|_F &= m\|\cos\theta - 1\| = \frac{m}{2}\,\|4\sin^2(\theta/2)\|\\
&\leq \frac{m}{2}\,\|2\sin(\theta/2)\|^2
\end{aligned} \tag{49}$$

(here, $\theta = (\theta_1, \dots, \theta_r)$ is the vector containing the diagonal elements of Θ). A similar calculation reveals that $\|\overline{W}\|_F^2 = m\|2\sin(\theta/2)\|^2$, thus proving $\|U^T\overline{W}\|_F^2 \leq \|\overline{W}\|_F^4/4 \leq Cm\delta^2\|\overline{W}\|_F^2$. The bound $\|V^T\overline{Z}\|_F^2 \leq Cn\delta^2\|\overline{Z}\|_F^2$ is proved in the same way, thus yielding

$$E_4 \leq Cn\epsilon\sqrt{\alpha}\sigma_{\max}(S)^2\delta^2\, d(\mathbf{x},\mathbf{u})^2\,. \tag{50}$$

By a similar calculation

$$\begin{aligned}
E_5 &= \frac{2\epsilon}{\sqrt{\alpha}}\mathrm{Tr}\{(S-\Sigma)S^T\overline{W}^TU\}\\
&\leq \frac{2\epsilon}{\sqrt{\alpha}}\sigma_{\max}((S-\Sigma)S^T)\|\overline{W}^TU\|_F\\
&\leq n\epsilon\sqrt{\alpha}\sigma_{\max}(S)\|S-\Sigma\|_F d(\mathbf{u},\mathbf{x})^2
\end{aligned}$$

and analogously

$$E_6 \leq n\epsilon\sqrt{\alpha}\sigma_{\max}(S)\|S-\Sigma\|_F d(\mathbf{u},\mathbf{x})^2\,.$$

Combining these estimates, and using $A^2 \geq \mathbb{E}\{A^2\}/2$, we get

$$\begin{aligned}
\frac{1}{n\epsilon}A^2 \geq\, & C_1\sqrt{\alpha}\|S-\Sigma\|_F^2 + C_1\sqrt{\alpha}\sigma_{\min}(S)^2 d(\mathbf{u},\mathbf{x})^2\\
&- C_2\sqrt{\alpha}\sigma_{\max}(S)^2\delta^2 d(\mathbf{u},\mathbf{x})^2\\
&- C_2\sqrt{\alpha}\sigma_{\max}(S)\,\|S-\Sigma\|_F\, d(\mathbf{u},\mathbf{x})^2
\end{aligned}$$

for some numerical constants C_1, $C_2 > 0$. Using the bounds $\sigma_{\min}(S)^2 \geq \Sigma_{\min}^2/2 - \|S-\Sigma\|_F^2$, $\sigma_{\max}(S)^2 \leq 2\Sigma_{\max}^2 + 2\,\|S-\Sigma\|_F^2$, and the assumption $d(\mathbf{x},\mathbf{u}) \leq \delta$ for $\delta \leq \Sigma_{\min}/C_0\Sigma_{\max}$, we get the claim (46).

We are now left with the task of proving the upper bound in (39). We can set $\Sigma = S$, thus obtaining

$$\begin{aligned}
F(X,Y) &\leq \frac{1}{2}\sum_{(i,j)\in E}(U\Sigma\overline{Z}^T + \overline{W}\Sigma V^T + \overline{W}\Sigma\overline{Z}^T)_{ij}^2\\
&\leq \widehat{A}^2 + \widehat{B}^2
\end{aligned}$$

where we defined

$$\begin{aligned}
\widehat{A}^2 &\equiv \sum_{(i,j)\in E}(U\Sigma\overline{Z}^T + \overline{W}\Sigma V^T)_{ij}^2\\
\widehat{B}^2 &\equiv \sum_{(i,j)\in E}(\overline{W}\Sigma\overline{Z}^T)_{ij}^2\,.
\end{aligned}$$

Bounds for these two quantities are derived as for A^2 and B^2. More precisely, by Theorem 4.1 in [9], we have $\widehat{A}^2 \leq 2\mathbb{E}\{\widehat{A}^2\}$ with probability at least $1 - 1/n^5$ and

$$\begin{aligned}
\mathbb{E}\{\widehat{A}^2\} &= \frac{\epsilon}{n\sqrt{\alpha}}\|\overline{W}\Sigma V^T + U\Sigma\overline{Z}^T\|_F^2\\
&= \frac{2\epsilon}{n\sqrt{\alpha}}\|\overline{W}\Sigma V^T\|_F^2 + \frac{2\epsilon}{n\sqrt{\alpha}}\|U\Sigma\overline{Z}^T\|_F^2\\
&\leq 2\sqrt{\alpha}n\epsilon\Sigma_{\max}^2\Big(\frac{1}{m}\|\overline{W}\|_F^2 + \frac{1}{n}\|\overline{Z}\|_F^2\Big)\\
&\leq 4\sqrt{\alpha}n\epsilon\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2\,.
\end{aligned}$$

$\widehat{B}^2$ is bounded similar to B^2 and we get

$$\widehat{B}^2 \leq C'\sqrt{\alpha}n\epsilon\Sigma_{\max}^2 d(\mathbf{u},\mathbf{x})^2\,.$$

□

VIII. Proof of Lemma 6.5

As in the proof of Lemma 6.4, see Section VII-B, we let $t \mapsto \mathbf{x}(t) = (X(t), Y(t))$ be the geodesic starting at $\mathbf{x}(0) = \mathbf{u}$ with velocity $\dot{\mathbf{x}}(0) = \mathbf{w} = (W, Z) \in \mathsf{T}_{\mathbf{u}}$. We also define $\mathbf{x} = \mathbf{x}(1) = (X, Y)$ with $X = U + \overline{W}$ and $Y = V + \overline{Z}$. Let $\widehat{\mathbf{w}} = \dot{\mathbf{x}}(1) = (\widehat{W}, \widehat{Z})$ be its velocity when passing through $\mathbf{x}$. An explicit expression is obtained in terms of the singular value decomposition of W and Z. If we let $W = L\Theta R^T$, and differentiate (27) with respect to t at $t = 1$, we obtain

$$\widehat{W} = -UR\Theta\sin\Theta\, R^T + L\Theta\cos\Theta\, R^T\,. \tag{51}$$

An analogous expression holds for $\widehat{Z}$. Since $L^T U = 0$, we have $||\widehat{W}||_F^2 = m||\Theta\sin\Theta||_F^2 + m||\Theta\cos\Theta||_F^2 = m||\theta||^2$. Hence[2]

$$\frac{1}{m}||\widehat{W}||_F^2 + \frac{1}{n}||\widehat{Z}||_F^2 = d(\mathbf{x},\mathbf{u})^2. \tag{52}$$

In order to prove the thesis, it is, therefore, sufficient to lower bound $\langle \mathrm{grad}\,\widetilde{F}(\mathbf{x}), \widehat{\mathbf{w}}\rangle$. In the following we will indeed show that

$$\langle \mathrm{grad}\,F(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq C\sqrt{\alpha}\, n\epsilon\Sigma_{\min}^2\, d(\mathbf{x},\mathbf{u})^2$$

and $\langle \mathrm{grad}\,G(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq 0$, which together imply the thesis by Cauchy–Schwarz inequality.

Let us prove a few preliminary estimates.

Remark 8.1: With the above definitions, $\widehat{\mathbf{w}} \in \mathcal{K}((11/2)\pi^2\mu_0)$.

Proof: Since $\Theta = \mathrm{diag}(\theta_1,\dots,\theta_r)$ with $|\theta_i| \leq \pi/2$, we get

$$\begin{aligned}||\widehat{W}^{(i)}||^2 &\leq 2||\Theta\sin\Theta R^T U^{(i)}||^2 + 2||\Theta\cos\Theta L^{(i)}||^2\\ &\leq \frac{\pi^2}{2}||U^{(i)}||^2 + 2||W^{(i)}||^2.\end{aligned} \tag{53}$$

By assumption we have $||U^{(i)}||^2 \leq \mu_0 r$ and by Remark 7.2 we have $||W^{(i)}||^2 \leq 5\pi^2\mu_0 r/2$. □

One important fact that we will use is that $\widehat{W}$ is well approximated by W or by $\overline{W}$, and $\widehat{Z}$ is well approximated by Z or by $\overline{Z}$. Using (43) and (51), we get

$$||\widehat{W}||_F^2 = ||W||_F^2 = m||\theta||^2 \tag{54}$$

$$||\overline{W}||_F^2 = m||2\sin\theta/2||^2 \tag{55}$$

$$\langle \widehat{W}, \overline{W}\rangle = m\sum_{a=1}^{r}\theta_a\sin\theta_a \tag{56}$$

$$\langle \widehat{W}, W\rangle = m\sum_{a=1}^{r}\theta_a^2\cos\theta_a \tag{57}$$

[2]Indeed, this conclusion could have been reached immediately, since $t \mapsto \mathbf{x}(t)$ is a geodesic parametrized proportionally to the arclength in th interval $t \in [0,1]$.

and therefore

$$||\widehat{W} - \overline{W}||_F^2 = m\sum_{a=1}^{r}[(2\sin(\theta_a/2))^2 + \theta_a^2 - 2\theta_a\sin\theta_a] \tag{58}$$

$$\begin{aligned}&\leq m\sum_{a=1}^{r}(\theta_a - 2\sin(\theta_a/2))^2 \leq \frac{m}{24^2}||\theta||^4\\ &\leq \frac{m}{24^2}d(\mathbf{u},\mathbf{x})^4.\end{aligned} \tag{59}$$

Analogously

$$\begin{aligned}||\widehat{W} - W||_F^2 &= n\sum_{a=1}^{r}[2\theta_a^2 - 2\theta_a^2\cos\theta_a]\\ &\leq m\,||\theta||^4 \leq m\,d(\mathbf{u},\mathbf{x})^4\end{aligned} \tag{60}$$

where we used the inequality $2(1-\cos x) \leq x^2$. The last inequality implies in particular

$$||U^T\widehat{W}||_F = ||U^T(W - \widehat{W})||_F \leq md(\mathbf{u},\mathbf{x})^2\,. \tag{61}$$

Similar bounds hold of course for $Z, \widehat{Z}, \overline{Z}$ (for instance we have $||V^T\widehat{Z}||_F \leq nd(\mathbf{u},\mathbf{x})^2$). Finally, we shall use repeatedly the fact that $||S - \Sigma||_F^2 \leq C\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2$, which follows from Lemma 6.4. This, in turn, implies

$$\sigma_{\max}(S) \leq \Sigma_{\max} + C\,\Sigma_{\max}d(\mathbf{x},\mathbf{u}) \leq 2\,\Sigma_{\max} \tag{62}$$

$$\sigma_{\min}(S) \geq \Sigma_{\min} - C\,\Sigma_{\max}d(\mathbf{x},\mathbf{u}) \geq \frac{1}{2}\,\Sigma_{\min} \tag{63}$$

where we used the hypothesis $d(\mathbf{x},\mathbf{u}) \leq \delta = \Sigma_{\min}/C_0\Sigma_{\max}$.

A. Lower Bound on $\mathrm{grad}\,F(\mathbf{x})$

Recalling that $\mathcal{P}_E$ is the projector defined in (29), and using the expression (30) and (31), for the gradient, we have (64), shown at the bottom of the page, where we defined

$$\begin{aligned}A &= \langle \mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T), (US\widehat{Z}^T + \widehat{W}SV^T)\rangle\\ B_1 &= |\langle \mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T), (\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle|\\ B_2 &= |\langle \mathcal{P}_E(U(S-\Sigma)V^T + \overline{W}S\overline{Z}^T), (US\widehat{Z}^T + \widehat{W}SV^T)\rangle|\\ B_3 &= |\langle \mathcal{P}_E(U(S-\Sigma)V^T + \overline{W}S\overline{Z}^T), (\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle|\,.\end{aligned}$$

At this point the proof becomes very similar to the one in the previous section and consists in lower bounding A and upper bounding B_1, B_2, B_3.

$$\begin{aligned}\langle \mathrm{grad}\,F(\mathbf{x}), \widehat{\mathbf{w}}\rangle &= \langle \mathcal{P}_E(XSY^T - M), (XS\widehat{Z}^T + \widehat{W}SY^T)\rangle\\ &= \langle \mathcal{P}_E(U(S-\Sigma)V^T + US\overline{Z}^T + \overline{W}SV^T + \overline{W}S\overline{Z}^T), (US\widehat{Z}^T + \widehat{W}SV^T + \overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle\\ &\geq A - B_1 - B_2 - B_3\end{aligned} \tag{64}$$

1) Lower Bound on A: Using Theorem 4.1 in [9] we obtain, with probability larger than $1-1/n^5$

$$\begin{aligned}A &\geq \frac{\epsilon}{2\sqrt{mn}}\langle (US\overline{Z}^T+\overline{W}SV^T),(US\widehat{Z}^T+\widehat{W}SV^T)\rangle\\ &\geq \frac{1}{2}A_0-\frac{1}{2}B_0\end{aligned}$$

where (see the equation shown at the bottom of the page). The term A_0 is lower bounded analogously to $\mathbb{E}\{A^2\}$ in the proof of Lemma 6.4, see (48) and (50)

$$\begin{aligned}A_0 &= \frac{\epsilon}{2\sqrt{mn}}\|US\overline{Z}^T+\overline{W}SV^T\|_F^2\\ &= \frac{\epsilon}{2\sqrt{mn}}\|US\overline{Z}^T\|_F^2+\frac{\epsilon}{2\sqrt{mn}}\|\overline{W}SV^T\|_F^2\\ &\quad+\frac{\epsilon}{2\sqrt{mn}}\langle US\overline{Z}^T,\overline{W}SV^T\rangle\\ &\geq Cn\epsilon(\sqrt{\alpha}\,\sigma_{\min}(S)^2-\sqrt{\alpha}\delta^2\sigma_{\max}(S)^2)d(\mathbf{x},\mathbf{u})^2\\ &\geq C\sqrt{\alpha}\,n\epsilon\,\Sigma_{\min}^2\,d(\mathbf{x},\mathbf{u})^2\end{aligned}$$

where we used the bounds (62), (63) and the hypothesis $d(\mathbf{x},\mathbf{u})\leq\delta=\Sigma_{\min}/C_0\Sigma_{\max}$.

As for the second term we notice that

$$\frac{B_0^2}{A_0}\leq n\epsilon\sqrt{\alpha}\Big(\frac{1}{m}\|S(\overline{W}-\widehat{W})\|_F^2+\frac{1}{n}\|S^T(\overline{Z}-\widehat{Z})\|_F^2\Big) \quad (65)$$

$$\leq n\epsilon\sqrt{\alpha}\,\sigma_{\max}(S)^2\Big(\frac{1}{m}\|\overline{W}-\widehat{W}\|_F^2+\frac{1}{n}\|\overline{Z}-\widehat{Z}\|_F^2\Big)$$

$$\leq Cn\epsilon\sqrt{\alpha}\,\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^4 \quad (66)$$

where, in the last step, we used the estimate (59) and the analogous one for $\|\overline{Z}-\widehat{Z}\|_F^2$. Therefore, for $d(\mathbf{x},\mathbf{u})\leq\delta\leq\Sigma_{\min}/C_0\Sigma_{\max}$ and C_0 large enough $A_0>2B_0$, where

$$A\geq C\sqrt{\alpha}\,n\epsilon\,\Sigma_{\min}^2\,d(\mathbf{x},\mathbf{u})^2\,. \quad (67)$$

2) Upper Bound on B_1: We begin by noting that B_1 can be bounded above by the sum of four terms of the form $B_1'=|\langle\mathcal{P}_E(US\overline{Z}^T),\overline{W}S\widehat{Z}^T\rangle|$. We show that $B_1'<A/100$. The other terms are bounded similarly.

Using Remark 7.3, we have

$$\begin{aligned}&\|\mathcal{P}_E(\overline{W}S\widehat{Z}^T)\|_F^2\\ &\leq C\frac{\epsilon}{\sqrt{mn}}\|\overline{W}\|_F^2\|S\widehat{Z}^T\|_F^2+C'\sqrt{\epsilon}\mu_0 r\sqrt{\alpha}\Sigma_{\max}\|\overline{W}\|_F\|S\widehat{Z}\|_F\\ &\leq 2C\frac{\epsilon}{\sqrt{mn}}\|\overline{W}\|_F^2\|S\overline{Z}^T\|_F^2+2C\frac{\epsilon}{\sqrt{mn}}\|\overline{W}\|_F^2\|S(\widehat{Z}-\overline{Z})^T\|_F^2\\ &\quad+C'\sqrt{\epsilon}\mu_0 r\sqrt{\alpha}\Sigma_{\max}\|\overline{W}\|_F\|S\overline{Z}\|_F\\ &\quad+C'\sqrt{\epsilon}\mu_0 r\sqrt{\alpha}\Sigma_{\max}\|\overline{W}\|_F\|S(\widehat{Z}-\overline{Z})\|_F\\ &\leq C''A\left(\delta^2+\frac{\mu_0 r\sqrt{\alpha}\Sigma_{\max}}{\sqrt{\epsilon}\Sigma_{\min}}\right)\end{aligned}$$

where we have used $\frac{\epsilon m}{\sqrt{mn}}\|S\overline{Z}^T\|_F^2\leq 3A_0\leq 12A$ from Section VIII-B1. Therefore, we have

$$\begin{aligned}B_1'^2 &\leq \|\mathcal{P}_E(US\overline{Z}^T)\|_F^2\|\mathcal{P}_E(\overline{W}S\widehat{Z}^T)\|_F^2\\ &\leq C\frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2 A\left(\delta^2+\frac{\mu_0 r\sqrt{\alpha}\Sigma_{\max}}{\sqrt{\epsilon}\Sigma_{\min}}\right)\\ &\leq C'A^2\left(\delta^2+\frac{\mu_0 r\sqrt{\alpha}\Sigma_{\max}}{\sqrt{\epsilon}\Sigma_{\min}}\right).\end{aligned}$$

The claim follows for δ and ϵ as in the hypothesis.

3) Upper Bound on B_2: We have

$$\begin{aligned}B_2 &\leq |\langle\mathcal{P}_E(US\widehat{Z}^T+\widehat{W}SV^T),\overline{W}S\overline{Z}^T\rangle|\\ &\quad+|\langle\mathcal{P}_E(US\widehat{Z}^T),U(S-\Sigma)V^T\rangle|\\ &\quad+|\langle\mathcal{P}_E(\widehat{W}SV^T),U(S-\Sigma)V^T\rangle|\\ &\equiv B_2'+B_2''+B_2'''\,.\end{aligned}$$

We claim that each of these three terms is smaller than $A/30$, where $B_2\leq A/10$.

The upper bound on B_2' is obtained similarly to the one on B_1 to get $B_2'\leq A/30$.

Consider now B_2''. By Theorem 4.1 in [9]

$$\begin{aligned}B_2'' &\leq \frac{\epsilon}{\sqrt{mn}}|\langle U(S-\Sigma)V^T,US\widehat{Z}^T\rangle|\\ &\quad+C\frac{\epsilon}{\sqrt{mn}}\sqrt{\frac{\mu_0 nr\alpha\log n}{n\sqrt{\alpha}\epsilon}}\|U(S-\Sigma)V^T\|_F\|US\widehat{Z}\|_F.\end{aligned}$$

To bound the second term, observe

$$\begin{aligned}\|US\widehat{Z}^T\|_F &\leq \|US\overline{Z}^T\|_F+\|US(\widehat{Z}-\overline{Z})^T\|_F\\ &\leq \|US\overline{Z}^T\|_F+\Sigma_{\max}\sqrt{mn}\,d(\mathbf{x},\mathbf{u})^2.\end{aligned}$$

Also, $\frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2\leq 3A_0\leq 12A$ from Section VIII-B2. Combining these, we have that the second term in B_2'' is smaller than $A/60$ for ϵ as in the hypothesis.

To bound the first term in B_2''

$$\begin{aligned}|\langle U(S-\Sigma)V^T,US\widehat{Z}^T\rangle| &= |\langle U(S-\Sigma)(Y-V)^T,US\widehat{Z}^T\rangle|\\ &\leq \|U(S-\Sigma)\overline{Z}^T\|_F\|US\overline{Z}^T\|_F\\ &\quad+\|U(S-\Sigma)\overline{Z}^T\|_F\|US(\overline{Z}-\widehat{Z}\|_F.\end{aligned}$$

$$A_0=\frac{\epsilon}{2\sqrt{mn}}\|US\overline{Z}^T+\overline{W}SV^T\|_F^2$$

$$B_0=\frac{\epsilon}{2\sqrt{mn}}\|US\overline{Z}^T+\overline{W}SV^T\|_F\|US(\overline{Z}-\widehat{Z})^T+(\overline{W}-\widehat{W})SV^T\|_F$$

Therefore

$$\begin{aligned}B_2'' &\le \frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)\overline{Z}\|_F\|US\overline{Z}\|_F \\ &\quad + \epsilon n\sqrt{\alpha}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^4 + A/60 \\ &\le \frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)\overline{Z}\|_F\|US\overline{Z}\|_F + A/40\end{aligned}$$

for $d(\mathbf{x},\mathbf{u}) \le \delta$ as in the hypothesis.

We are now left with upper bounding $\widetilde{B}_2'' \equiv \frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)\overline{Z}\|_F\|US\overline{Z}\|_F$

$$\begin{aligned}\widetilde{B}_2''^2 &\le \left(\frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)\overline{Z}^T\|_F^2\right)\left(\frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2\right) \\ &\le \left(\epsilon n\sqrt{\alpha}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^4\right)\left(\frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2\right).\end{aligned}$$

Also from the lower bound on A, we have, $\frac{\epsilon}{\sqrt{mn}}\|US\overline{Z}^T\|_F^2 \le 3A_0 \le 12A$. Using $d(\mathbf{x},\mathbf{u}) \le \delta$, we have $\widetilde{B}_2'' \le A/120$ for δ as in the hypothesis. This proves the desired result. The bound on B_2''' is calculated analogously.

4) Upper Bound on B_3: Finally, for the last term it is sufficient to use a crude bound (see the equation shown at the bottom of the page). The terms of the form $\|\mathcal{P}_E(\overline{W}S\widehat{Z}^T)\|_F$ are all estimated as in Section VIII-B2. Also, by Theorem 4.1 of [9]

$$\begin{aligned}\|\mathcal{P}_E(U(S-\Sigma)V^T)\|_F &\le C\frac{\epsilon}{\sqrt{mn}}\|U(S-\Sigma)V^T\|_F^2 \\ &\le Cn\epsilon\sqrt{\alpha}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2.\end{aligned}$$

Combining these estimates with the δ and the ϵ in the hypothesis, we get $B_3 \le A/10$

B. Lower Bound on $\operatorname{grad} G(\mathbf{x})$

By the definition of G in (35), we have

$$\begin{aligned}\langle \operatorname{grad} G(\mathbf{x}), \widehat{\mathbf{w}}\rangle &= \frac{1}{\mu_0 r}\sum_{i=1}^m G_1'\left(\frac{\|X^{(i)}\|^2}{3\mu_0 r}\right)\langle X^{(i)}, \widehat{W}^{(i)}\rangle \\ &\quad + \frac{1}{\mu_0 r}\sum_{j=1}^n G_1'\left(\frac{\|Y^{(i)}\|^2}{3\mu_0 r}\right)\langle Y^{(i)}, \widehat{Z}^{(i)}\rangle. \quad (68)\end{aligned}$$

It is, therefore, sufficient to show that if $\|X^{(i)}\|^2 > 3\mu_0 r$, then $\langle X^{(i)}, \widehat{W}^{(i)}\rangle > 0$, and if $\|Y^{(j)}\|^2 > 3\mu_0 r$, then $\langle Y^{(j)}, \widehat{Z}^{(j)}\rangle > 0$. We will just consider the first statement, the second being completely symmetrical.

From the explicit expressions (43) and (51), we get

$$X^{(i)} = R\left\{\cos\Theta\, R^T U^{(i)} + \sin\Theta\, L^{(i)}\right\} \quad (69)$$

$$\widehat{W}^{(i)} = R\left\{\Theta\cos\Theta L^{(i)} - \Theta\sin\Theta R^T U^{(i)}\right\}. \quad (70)$$

From the first expression, it follows that

$$\|\sin\Theta\, L^{(i)}\|^2 \le \|X^{(i)}\|^2 + \|\cos\Theta\, R^T U^{(i)}\|^2 \le 5\,\mu_0 r.$$

On the other hand, by taking the difference of (69) and (70), we have

$$\begin{aligned}&\|X^{(i)} - \widehat{W}^{(i)}\| \\ &\le \|(\sin\Theta - \Theta\cos\Theta)L^{(i)}\| + \|(\cos\Theta + \Theta\sin\Theta)R^T U^{(i)}\| \\ &\le \max_i(\theta_i^2)\|\sin\Theta L^{(i)}\| + \frac{\pi}{2}\|U^{(i)}\| \le \delta\sqrt{4\mu_0 r} + \frac{\pi}{2}\sqrt{\mu_0 r}.\end{aligned}$$

where we used the inequality $(\sin\omega - \omega\cos\omega) \le \omega^2\sin\omega$ valid for $\omega \in [0,\pi/2]$. For δ small enough, we have, therefore, $\|X^{(i)} - \widehat{W}^{(i)}\| \le (99/100)\sqrt{3\mu_0 r}$. To conclude, for $\|X^{(i)}\| \ge 3\mu_0 r$

$$\begin{aligned}\langle X^{(i)}, \widehat{W}^{(i)}\rangle &\ge \|X^{(i)}\|^2 - \|X^{(i)}\|\,\|X^{(i)} - \widehat{W}^{(i)}\| \\ &\ge \|X^{(i)}\|(\sqrt{3\mu_0 r} - (99/100)\sqrt{3\mu_0 r}) \ge 0.\end{aligned}$$

Appendix A
Proof of Remark 4.2

The proof is a generalization of analogous result in [19], which is proved to hold only with probability larger than $1 - e^{-C\epsilon}$. The stronger statement quoted here can be proved using concentration of measure inequalities.

First, we apply Chernoff bound to the event $\left\{|\overline{\mathcal{A}}_l| > \max\{e^{-C_1\epsilon}m, C_2\alpha\}\right\}$. In the case of large ϵ, when $\epsilon > 3\sqrt{\alpha}\log n$, we have $\mathbb{P}\left\{|\overline{\mathcal{A}}_l| > C_2\alpha\right\} \le 1/2n^4$, for $C_2 \ge \max\{e, 26/\alpha\}$. In the case of small ϵ, when $\epsilon \le 3\sqrt{\alpha}\log n$, we have $\mathbb{P}\left\{|\overline{\mathcal{A}}_l| > \max\{e^{-C_1\epsilon}m, C_2\alpha\}\right\} \le 1/2n^4$, for $C_1 \le 1/600\sqrt{\alpha}$ and $C_2 \ge 130$. Here, we made a moderate assumption of $\epsilon \ge 3\sqrt{\alpha}$, which is typically in the region of interest.

Analogously, we can prove that $\mathbb{P}\left\{|\overline{\mathcal{A}}_r| > \max\{e^{-C_1\epsilon}n, C_2\}\right\} \le 1/2n^4$, which finishes the proof of Remark 4.2.

$$B_3 \le 4\Big(\|\mathcal{P}_E(\overline{W}S\widehat{Z}^T)\|_F + \|\mathcal{P}_E(\widehat{W}S\overline{Z}^T)\|_F\Big)\Big(\|\mathcal{P}_E(U(S-\Sigma)V^T)\|_F + \|\mathcal{P}_E(\overline{W}S\overline{Z}^T)\|_F\Big)$$

APPENDIX B
PROOF OF REMARK 4.4

The expectation of the contribution of light couples, when each edge is independently revealed with probability $\epsilon/\sqrt{mn}$, is

$$\mathbb{E}[Z] = \frac{\epsilon}{\sqrt{mn}}\left(\sum_{(i,j)\in L} x_i M^A_{ij} y_j - x^T M y\right)$$

where we define M^A by setting to zero the rows of M whose index is not in A_l and the columns of M whose index is not in A_r.

In order to bound $\sum_L x_i M^A_{ij} y_j - x^T M y$, we write

$$\begin{aligned}\left|\sum_{(i,j)\in L} x_i M^A_{ij} y_j - x^T M y\right| &= \left|x^T\left(M^A - M\right)y - \sum_{(i,j)\in \overline{L}} x_i M^A_{ij} y_j\right| \\ &\le \left|x^T\left(M^A - M\right)y\right| + \left|\sum_{(i,j)\in \overline{L}} x_i M^A_{ij} y_j\right|.\end{aligned}$$

Note that $|(M^A - M)_{ij}|$ is nonzero only if $i \notin A_l$ or $j \notin A_r$, in which case $|(M^A - M)_{ij}| \le M_{\max}$. Also, by Remark 4.2, there exists $\delta = \max\{e^{-C_1\epsilon}, C_2/n\}$ such that $|i : i \notin A_l| \le \delta m$ and $|j : j \notin A_r| \le \delta n$. Denoting by $\mathbb{I}(\,\cdot\,)$ the indicator function, we have

$$\begin{aligned}&\left|x^T\left(M^A - M\right)y\right| \\ &\le \sum_{ij} |x_i||y_j|\left(\mathbb{I}(i\notin A_l) + \mathbb{I}(j \notin A_r)\right)M_{\max} \\ &= \left(\sum_i |x_i|\mathbb{I}(i\notin A_l)\sum_j |y_j| + \sum_j |y_j|\mathbb{I}(j\notin A_r)\sum_i |x_i|\right)M_{\max} \\ &\le \left(\sqrt{\delta m}\sqrt{n} + \sqrt{\delta n}\sqrt{m}\right)M_{\max} \\ &\le M_{\max}\sqrt{\frac{mn}{\epsilon}}\,.\end{aligned}$$

for $\delta \le \frac{1}{4\epsilon}$. We can bound the second term as follows:

$$\begin{aligned}\left|\sum_{(i,j)\in\overline{L}} x_i M^A_{ij} y_j\right| &\le \sum_{(i,j)\in\overline{L}} \frac{\left|x_i M^A_{ij} y_j\right|^2}{\left|x_i M^A_{ij} y_j\right|} \\ &\le \frac{1}{M_{\max}}\sqrt{\frac{mn}{\epsilon}} \sum_{(i,j)\in\overline{L}} \left|x_i M^A_{ij} y_j\right|^2 \\ &\le \frac{1}{M_{\max}}\sqrt{\frac{mn}{\epsilon}} \sum_{i\in[m],j\in[n]} \left|x_i M^A_{ij} y_j\right|^2 \\ &\le M_{\max}\sqrt{\frac{mn}{\epsilon}}\end{aligned}$$

where the second inequality follows from the definition of heavy couples.

Hence, summing up the two contributions, we get

$$|\mathbb{E}\,[Z]| \le 2M_{\max}\sqrt{\epsilon}\,.$$

APPENDIX C
PROOF OF REMARK 4.5

We can associate to the matrix Q a bipartite graph $\mathcal{G} = ([m],[n],\mathcal{E})$. The proof is similar to the one in [17], [19] and is based on two properties of the graph $\mathcal{G}$:

1. *Bounded degree.* The graph $\mathcal{G}$ has maximum degree bounded by a constant times the average degree
$$\deg(i) \le \frac{2\epsilon}{\sqrt{\alpha}} \tag{71}$$
$$\deg(j) \le 2\epsilon\sqrt{\alpha} \tag{72}$$
for all $i \in [m]$ and $j \in [n]$.
2. *Discrepancy.* We say that $\mathcal{G}$ (equivalently, the adjacency matrix Q) has the discrepancy property if, for any $A \subseteq [m]$ and $B \subseteq [n]$, one of the following is true [see (73)–(74), shown at the bottom of the page] for two numerical constants ξ_1, ξ_2 (independent of n and ϵ). Here $e(A,B)$ denotes the number of edges between A and B and $\mu(A,B) = |A||B||E|/mn$ denotes the average number of edges between A and B before trimming.

1. $$\frac{e(A,B)}{\mu(A,B)} \le \xi_1 \tag{73}$$
2. $$e(A,B)\ln\left(\frac{e(A,B)}{\mu(A,B)}\right) \le \xi_2 \max\{|A|/\sqrt{\alpha}, |B|\sqrt{\alpha}\}\ln\left(\frac{\sqrt{mn}}{\max\{|A|/\sqrt{\alpha}, |B|\sqrt{\alpha}\}}\right) \tag{74}$$

We will prove, later in this section, that the discrepancy property holds with high probability.

Let us partition row and column indices with respect to the value of x_u and y_v

$$A_i = \{u \in [m] : \frac{\Delta}{\sqrt{m}} 2^{i-1} \le |x_u| < \frac{\Delta}{\sqrt{m}} 2^i\}$$
$$B_j = \{v \in [n] : \frac{\Delta}{\sqrt{n}} 2^{j-1} \le |y_v| < \frac{\Delta}{\sqrt{n}} 2^j\}$$

for $i \in \{1, 2, \ldots, \lceil \ln(\sqrt{m}/\Delta)/\ln 2 \rceil\}$, and $j \in \{1, 2, \ldots, \lceil \ln(\sqrt{n}/\Delta)/\ln 2 \rceil\}$, and we denote the size of subsets A_i and B_j by a_i and b_j, respectively. Furthermore, we define $e_{i,j}$ to be the number of edges between two subsets A_i and B_j, and we let $\mu_{i,j} = a_i b_j (\epsilon/\sqrt{mn})$. Notice that all indices u of non zero x_u fall into one of the subsets A_i's defined above, since, by discretization, the smallest nonzero element of $x \in T_m$ in absolute value is at least $\Delta/\sqrt{m}$. The same applies for the entries of $y \in T_n$.

By grouping the summation into A_i's and B_j's, we get

$$\begin{aligned}\sum_{\substack{(u,v):\\ |x_u y_v| \ge \frac{C\sqrt{\epsilon}}{\sqrt{mn}}}} Q_{uv}|x_u y_v| &\le \sum_{(i,j):2^{i+j} \ge \frac{4C\sqrt{\epsilon}}{\Delta^2}} e_{i,j} \frac{\Delta 2^i}{\sqrt{m}} \frac{\Delta 2^j}{\sqrt{n}} \\ &= \Delta^2 \sum a_i b_j \frac{\epsilon}{\sqrt{mn}} \frac{e_{i,j}}{\mu_{i,j}} \frac{2^i}{\sqrt{m}} \frac{2^j}{\sqrt{n}} \\ &= \Delta^2 \sqrt{\epsilon} \sum \underbrace{a_i \frac{2^{2i}}{m}}_{\alpha_i} \underbrace{b_j \frac{2^{2j}}{n}}_{\beta_j} \underbrace{\frac{e_{i,j}\sqrt{\epsilon}}{\mu_{i,j} 2^{i+j}}}_{\sigma_{i,j}}.\end{aligned}$$

Note that, by definition, we have

$$\sum_i \alpha_i \le 4\|x\|^2/\Delta^2 \tag{75}$$
$$\sum_i \beta_i \le 4\|y\|^2/\Delta^2. \tag{76}$$

We are now left with task of bounding $\sum \alpha_i \beta_j \sigma_{i,j}$, for Q that satisfies bounded degree property and discrepancy property.

Define

$$\mathcal{C}_1 \equiv \left\{(i,j) : 2^{i+j} \ge \frac{4C\sqrt{\epsilon}}{\Delta^2} \text{ and } (A_i, B_j) \text{ satisfies (73)}\right\} \tag{77}$$
$$\mathcal{C}_2 \equiv \left\{(i,j) : 2^{i+j} \ge \frac{4C\sqrt{\epsilon}}{\Delta^2} \text{ and } (A_i, B_j) \text{ satisfies (74)}\right\} \backslash \mathcal{C}_1. \tag{78}$$

We need to show that $\sum_{(i,j)\in\mathcal{C}_1\cup\mathcal{C}_2} \alpha_i\beta_j\sigma_{i,j}$ is bounded.

For the terms in $\mathcal{C}_1$ this bound is easy. Since summation is over pairs of indices (i,j) such that $2^{i+j} \ge \frac{4C\sqrt{\epsilon}}{\Delta^2}$, it follows from bounded degree property that $\sigma_{i,j} \le \xi_1\Delta^2/4C$. By (75) and (76), we have $\sum_{\mathcal{C}_1} \alpha_i\beta_j\sigma_{i,j} \le (\xi_1\Delta^2/4C)(2/\Delta)^4 = O(1)$.

For the terms in $\mathcal{C}_2$, the bound is more complicated. We assume $a_i \le \alpha b_j$ for simplicity and the other case can be treated in the same manner. By change of notation the second discrepancy condition becomes

$$\begin{aligned}&e_{i,j} \log\left(\frac{e_{i,j}}{\mu_{i,j}}\right) \\ &\quad \le \xi_2 \max\{a_i/\sqrt{\alpha}, b_j\sqrt{\alpha}\} \log\left(\frac{\sqrt{mn}}{\max\{a_i/\sqrt{\alpha}, b_j\sqrt{\alpha}\}}\right).\end{aligned} \tag{79}$$

We start by changing variables on both sides of (79)

$$\frac{e_{i,j} a_i b_j \epsilon}{\mu_{i,j}\sqrt{mn}} \log\left(\frac{e_{i,j}}{\mu_{i,j}}\right) \le \xi_2 b_j \sqrt{\alpha} \log\left(\frac{2^{2j}}{\beta_j}\right).$$

Now, multiply each side by $2^i/b_j\sqrt{\epsilon}2^j$ to get

$$\sigma_{i,j}\alpha_i \log\left(\frac{e_{i,j}}{\mu_{i,j}}\right) \le \frac{\xi_2 2^i}{\sqrt{\epsilon}2^j}\left[\log(2^{2j}) - \log\beta_j\right]. \tag{80}$$

To achieve the desired bound, we partition the analysis into five cases.

1. $\sigma_{i,j} \le 1$: By (75) and (76), we have $\sum \alpha_i\beta_j\sigma_{i,j} \le (2/\Delta)^4 = O(1)$.
2. $2^i > \sqrt{\epsilon}2^j$: By the bounded degree property in (72), we have $e_{i,j} \le a_i 2\epsilon/\sqrt{\alpha}$, which implies that $e_{i,j}/\mu_{i,j} \le 2n/b_j$. For a fixed i we have, $\sum_j \beta_j\sigma_{i,j}\mathbb{I}(2^i > \sqrt{\epsilon}2^j) \le 2\sqrt{\epsilon}\sum_j 2^{j-i}\mathbb{I}(2^i > \sqrt{\epsilon}2^j) \le 4$. Then, $\sum \alpha_i\beta_j\sigma_{i,j} \le 16/\Delta^2 = O(1)$.
3. $\log(e_{i,j}/\mu_{i,j}) > \frac{1}{4}\left[\log(2^{2j}) - \log\beta_j\right]$: From (80), it immediately follows that $\sigma_{i,j}\alpha_i \le \frac{4\xi_2 2^i}{\sqrt{\epsilon}2^j}$. Due to case 2, we can assume $2^i \le \sqrt{\epsilon}2^j$, which implies that for a fixed j we have the following inequality : $\sum_i \sigma_{i,j}\alpha_i \le 4\xi_2\sum_i \frac{2^i}{\sqrt{\epsilon}2^j}\mathbb{I}(2^i \le \sqrt{\epsilon}2^j) \le 8\xi_2$. Then it follows by (76) that $\sum \alpha_i\beta_j\sigma_{i,j} \le 32\xi_2/\Delta^2 = O(1)$.
4. $\log(2^{2j}) \ge -\log\beta_j$: Due to case 3, we can assume $\log(e_{i,j}/\mu_{i,j}) \le \frac{1}{4}\left[\log(2^{2j}) - \log\beta_j\right]$, which implies that $\log(e_{i,j}/\mu_{i,j}) \le \log(2^j)$. Further, since we are not in case 1, we can assume $1 < \sigma_{i,j} = e_{i,j}\sqrt{\epsilon}/\mu_{i,j}2^{i+j}$. Combining those two inequalities, we get $2^i \le \sqrt{\epsilon}$.
   Since in defining $\mathcal{C}_2$ we excluded $\mathcal{C}_1$, if $(i,j) \in \mathcal{C}_2$ then $\log(e_{i,j}/\mu_{i,j}) \ge 1$. Applying (80) we get
   $$\begin{aligned}\sigma_{i,j}\alpha_i &\le \sigma_{i,j}\alpha_i \log(e_{i,j}/\mu_{i,j}) \\ &\le (\xi_2 2^{i-j}/\sqrt{\epsilon})\left[\log(2^{2j}) - \log\beta_j\right] \le 4\xi_2 2^i/\sqrt{\epsilon}.\end{aligned}$$
   Combining above two results, it follows that $\sum_i \sigma_{i,j}\alpha_i \le 4\xi_2\sum_i \frac{2^i}{\sqrt{\epsilon}}\mathbb{I}(2^i \le \sqrt{\epsilon}) \le 8\xi_2$. Then, we have the desired bound : $\sum \alpha_i\beta_j\sigma_{i,j} \le \frac{32\xi_2}{\Delta^2} = O(1)$.
5. $\log(2^{2j}) < -\log\beta_j$: It follows, since we are not in case 3, that $\log(e_{i,j}/\mu_{i,j}) \le \frac{1}{4}\left[\log(2^{2j}) - \log\beta_j\right] \le -\log\beta_j$. Hence, $e_{i,j}/\mu_{i,j} \le 1/\beta_j$. This implies that $\sigma_{i,j} = e_{i,j}\sqrt{\epsilon}/\mu_{i,j}2^{i+j} \le \sqrt{\epsilon}/\beta_j 2^{i+j}$. Since the summation is over pairs of indices (i,j) such that $2^{i+j} \ge 4C\sqrt{\epsilon}/\Delta^2$, we have $\sum_j \sigma_{i,j}\beta_j \le \frac{\Delta^2}{2C}$. Then it follows that $\sum \alpha_i\beta_j\sigma_{i,j} \le \frac{2}{C} = O(1)$.

Analogous analysis for the set of indices (i,j) such that $a_i > \alpha b_j$ will give us similar bounds. Summing up the results, we get

$$nA^{-T}A^{-1} = X'^T X' = Q_1(n\mathbf{I} - (U - X'Q_1)^T U + U^T(U - X'Q_1) + (U - X'Q_1)^T(U - X'Q_1))Q_1^T \tag{83}$$

that there exists a constant $C' \leq \frac{32}{\Delta^4} + \frac{4\xi_1}{C\Delta^2} + \frac{32}{\Delta^2} + \frac{128\xi_2}{\Delta^2} + \frac{4}{C}$, such that

$$\sum_{(i,j):2^{i+j} \geq \frac{4C\sqrt{\epsilon}}{\Delta^2}} \alpha_i \beta_j \sigma_{i,j} \leq C' .$$

This finishes the proof of Remark 4.5.

Lemma C.1: The adjacency matrix Q has discrepancy property with probability at least $1 - 1/2n^3$.

Proof: The proof is a generalization of analogous result in [17], [19] which is proved to hold only with probability larger than $1 - e^{-C\epsilon}$. The stronger statement quoted here is a result of the observation that, when we trim the graph the number of edges between any two subsets does not increase. Define Q_0 to be the adjacency matrix corresponding to original random matrix M^E before trimming. If the discrepancy assumption holds for Q_0, then it also holds for Q, since $e^Q(A,B) \leq e^{Q_0}(A,B)$, for $A \subseteq [m]$ and $B \subseteq [n]$.

Now we need to show that the desired property is satisfied for Q_0. This is proved for the case of nonbipartite graph in [19, Section 2.2.5], and analogous analysis for bipartite graph shows that for all subsets $A \subseteq [m]$ and $B \subseteq [n]$, with probability at least $1 - 1/2(mn)^p$, the discrepancy condition holds with $\xi_1 = 2e$ and $\xi_2 = (3p + 12)(\alpha^{1/2} + \alpha^{-1/2})$. Since we assume $\alpha \geq 1$, taking p to be $3/2$ proves our claim. □

APPENDIX D
PROOF OF REMARKS 6.1, 6.2 AND 6.3

Proof: (Remark 6.1.) Let $\theta = (\theta_1, \ldots, \theta_p)$, $\theta_i \in [-\pi/2, \pi/2]$ be the principal angles between the planes spanned by the columns of X_1 and X_2. It is known that $d_c(X_1, X_2) = \|2\sin(\theta/2)\|_2$ and $d_p(X_1, X_2) = \|\sin\theta\|_2$. The thesis follows from the elementary inequalities

$$\frac{1}{\pi}\alpha \leq \sqrt{2}\sin(\alpha/2) \leq \sin\alpha \leq 2\sin(\alpha/2) \tag{81}$$

valid for $\alpha \in [0, \pi/2]$. □

Proof: (Remark 6.2) Given $X \in \mathbb{R}^{n\times r}$, define X' by

$$X'^{(i)} = \frac{X^{(i)}}{\|X^{(i)}\|}\min\left(\|X^{(i)}\|, \sqrt{\mu_0 r}\right)$$

for all $i \in [n]$.

Let A be a matrix for extracting the orthonormal basis of the columns of X'. That is $A \in \mathbb{R}^{r\times r}$ such that $X'' = X'A$ and $X''^T X'' = n\mathbf{I}$. Without loss of generality, A can be taken to be a symmetric matrix. In the following, let $\sigma_i = \sigma_i(A^{-1})$ for all $i \in [n]$. Note that by construction $d(U, X') \leq d(U, X) \leq \delta$. Hence, there is a $Q_1 \in \mathsf{O}(r)$ such that

$$\|U - X'Q_1\|_F^2 \leq n\delta^2. \tag{82}$$

We start by writing (83), shown at the top of the page.

Using (82), we have

$$\begin{aligned}\|(U - X'Q_1)^T U\|_F &\leq \|U\|_2 \|(U - X'Q_1)\|_F \\ &\leq n\delta\end{aligned}$$

and

$$\|(U - X'Q_1)^T(U - X'Q_1)\|_F \leq n\delta^2.$$

Therefore, using (83)

$$\sigma_1^2 \leq 1 + 2\delta + \delta^2 \tag{84}$$

$$\sigma_r^2 \geq 1 - 2\delta - \delta^2. \tag{85}$$

From (84), (85), and $\delta \leq 1/16$, we get $\sigma_1 \leq \sqrt{3}$ and $\sigma_r \geq 1/\sqrt{3}$. Since $\|X''^{(i)}\|^2 = \|X'^{(i)}A\|^2 \leq 3\mu_0 r$ for all $i \in [n]$, we have that $X'' \in \mathcal{K}(3\mu_0)$.

We next prove that $d(X', X'') \leq 3\delta$ which implies the thesis by triangular inequality

$$\begin{aligned}d(X', X'')^2 &= \frac{1}{n}\min_{Q\in\mathsf{O}(r)} \|X' - X''Q\|_F^2 \\ &\leq \frac{1}{n}\|X''A^{-1} - X''\|_F^2 \\ &= \|A^{-1} - \mathbf{I}\|_F^2 \\ &\leq \|(A^{-1} - \mathbf{I})(A^{-1} + \mathbf{I})\|_F^2 \\ &\leq \|A^{-T}A^{-1} - \mathbf{I}\|_F^2 \\ &\leq 9\delta^2\end{aligned}$$

where the last inequality is from (83). □

Proof: (Remark 6.3.) We start by observing that

$$d_p(V, Y) = \frac{1}{\sqrt{n}}\min_{A\in\mathbb{R}^{r\times r}} \|V - YA\|_F . \tag{86}$$

Indeed the minimization on the right hand side can be performed explicitly (as $\|V - YA\|_F^2$ is a quadratic function of A) and the minimum is achieved at $A = Y^T V/n$. The inequality follows by simple algebraic manipulations.

Take $A = S^T X^T U\Sigma^{-1}/m$. Then

$$\|V - YA\|_F = \sup_{B, \|B\|_F \leq 1} \langle B, (V - YA)\rangle \tag{87}$$

$$= \sup_{B, \|B\|_F \leq 1} \langle B^T, \frac{1}{m}\Sigma^{-1}U^T(U\Sigma V^T - XSY^T)\rangle \tag{88}$$

$$= \frac{1}{m}\sup_{B, \|B\|_F \leq 1} \langle U\Sigma^{-1}B^T, (M - \widehat{M})\rangle \tag{89}$$

$$\leq \frac{1}{m}\sup_{B, \|B\|_F \leq 1} \|U\Sigma^{-1}B^T\|_F \, \|M - \widehat{M}\|_F . \tag{90}$$

On the other hand

$$\begin{aligned}\|U\Sigma^{-1}B^T\|_F^2 &= \mathrm{Tr}(B\Sigma^{-1}U^T U\Sigma^{-1}B^T) \\ &= m\mathrm{Tr}(B^T B\Sigma^{-2}) \leq m\Sigma_{\min}^{-2}\|B\|_F^2\end{aligned}$$

whereby the last inequality follows from the fact that Σ is diagonal. Together, (86) and (90), this implies our claim. $\square$

ACKNOWLEDGMENT

The authors would like to thank E. Candès and B. Recht for stimulating discussions on the subject of this paper.

REFERENCES

[1] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia, "Spectral analysis of data," in *Proc. 33rd Annu. ACM Symp. Theory of Computing*, New York, 2001, pp. 619–626.

[2] D. Achlioptas and F. McSherry, "Fast computation of low-rank matrix approximations," *J. ACM*, vol. 54, no. 2, p. 9, 2007.

[3] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.

[4] M. W. Berry, Z. Drmać, and E. R. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM Rev.*, vol. 41, no. 2, pp. 335–362, 1999.

[5] M. W. Berry, "Large scale sparse singular value computations," *Int. J. Supercomput. Appl.*, vol. 6, pp. 13–49, 1992.

[6] B. Bollobás*, Graph Theory: An Introductory Course*. New York: Springer-Verlag, 1979.

[7] J.-F. Cai, E. J. Candès, and Z. Shen, A Singular Value Thresholding Algorithm for Matrix Completion, 2008 [Online]. Available: arXiv:0810.3286

[8] E. J. Candès and Y. Plan, Matrix Completion with Noise, 2009 [Online]. Available: arXiv:0903.3131

[9] E. J. Candès and B. Recht, Exact Matrix Completion via Convex Optimization, 2008 [Online]. Available: arxiv:0805.4471

[10] E. J. Candès, J. K. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 489–509, 2006.

[11] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to sfm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1051–1063, Aug. 2004.

[12] E. J. Candès and T. Tao, The Power of Convex Relaxation: Near-Optimal Matrix Completion, 2009 [Online]. Available: arXiv:0903.1476

[13] W. Dai and O. Milenkovic, Set: An Algorithm for Consistent Matrix Completion, 2009 [Online]. Available: arXiv:0909.2705

[14] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, 2006.

[15] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matr. Anal. Appl.*, vol. 20, pp. 303–353, 1999.

[16] M. Fazel, "Matrix Rank Minimization with Applications," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2002.

[17] J. Friedman, J. Kahn, and E. Szemerédi, "On the second eigenvalue in random regular graphs," in *Proc. 21st Annu. ACM Symp. Theory of Computing*, Seattle, WA, May 1989, pp. 587–598.

[18] A. Frieze, R. Kannan, and S. Vempala, "Fast monte-carlo algorithms for finding low-rank approximations," *J. ACM*, vol. 51, no. 6, pp. 1025–1041, 2004.

[19] U. Feige and E. Ofek, "Spectral techniques applied to sparse random graphs," *Random Struct. Alg.*, vol. 27, no. 2, pp. 251–275, 2005.

[20] R. H. Keshavan, A. Montanari, and S. Oh, "Learning low rank matrices from $O(n)$ entries," presented at the Allerton Conf. on Commun., Control and Computing, Sep. 2008.

[21] R. H. Keshavan, A. Montanari, and S. Oh, Matrix Completion from a Few Entries, 2009 [Online]. Available: arXiv:0901.3150

[22] R. H. Keshavan, A. Montanari, and S. Oh, Matrix Completion from Noisy Entries, 2009 [Online]. Available: arXiv:0906.2027

[23] R. H. Keshavan and S. Oh, A Gradient Descent Algorithm on the Grassmann Manifold for Matrix Completion, 2009 [Online]. Available: arXiv:0910.5260

[24] K. Lee and Y. Bresler, Admira: Atomic Decomposition for Minimum Rank Approximation, 2009 [Online]. Available: arXiv:0905.0044

[25] T. Luczak, "On the equivalence of two basic models of random graphs," in *Proc. 3rd Int. Seminar on Random Graphs and Probabilistic Methods in Combinatorics*, 1990, pp. 151–157.

[26] S. Ma, D. Goldfarb, and L. Chen, Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization, 2009 [Online]. Available: arXiv:0905.1643

[27] R. Mazumder, T. Hastie, and R. Tibshirani, Spectral Regularization Algorithms for Learning Large Incomplete Matrices, 2009 [Online]. Available: http://www-stat.stanford.edu/~hastie/Papers/SVD_JMLR.pdf

[28] R. Meka, P. Jain, and I. S. Dhillon, Guaranteed Rank Minimization via Singular Value Projection, 2009 [Online]. Available: arXiv:0909.5457

[29] *Netflix Prize* [Online]. Available: http://www.netflixprize.com/

[30] B. Recht, M. Fazel, and P. Parrilo, Guaranteed Minimum Rank Solutions of Matrix Equations via Nuclear Norm Minimization, 2007 [Online]. Available: arxiv:0706.4138

[31] A. Singer and M. Cucuringu, Uniqueness of Low-Rank Matrix Completion by Rigidity Theory, 2009 [Online]. Available: arXiv:0902.3846

[32] Y. Seginer, "The expected norm of random matrices," *Combin., Probab., Comput.*, vol. 9, pp. 149–166, 2000.

[33] K. Toh and S. Yun, An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Least Squares Problems 2009 [Online]. Available: http://www.math.nus.edu.sg/~matys

[34] Z. Wang, S. Zheng, S. Boyd, and Y. Ye, Further Relaxations of the sdp Approach to Sensor Network Localization, Tech. Rep. 2006.

Raghunandan H. Keshavan received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, in 2007. He is currently pursuing the Ph.D. degree at the Information Systems Laboratory (ISL), Department of Electrical Engineering, Stanford University, Stanford, CA. His research is under the supervision of Prof. A. Montanari in the field of machine learning and probabilistic graphical models.

Andrea Montanari received the Laurea degree in physics in 1997 and the Ph.D. degree in theoretical physics in 2001 from Scuola Normale Superiore, Pisa, Italy.

He was a Postdoctoral fellow at the Laboratoire de Physique Théorique de l'Ecole Normale Supérieure (LPTENS), Paris, France, and the Mathematical Sciences Research Institute, Berkeley, CA. Since 2002, he has been Chargé de Recherche (with Centre National de la Recherche Scientifique, CNRS) at LPTENS. In September 2006, he joined Stanford University, Stanford, CA, as Assistant Professor in the Department of Electrical Engineering and the Deprtment of Statistics.

Prof. Montanari was a corecipient of the ACM SIGMETRICS best paper award in 2008. He received the CNRS bronze medal for theoretical physics in 2006 and the National Science Foundation CAREER Award in 2008.

Sewoong Oh (S'07) received the B.S. degree in electrical engineering from Seoul National University, Korea, in 2002, and the M.S. degree in electrical engineering in 2007 from Stanford University, Stanford, CA, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests are machine learning, inference in graphical models, and coding theory. He is a recipient of the Samsung scholarship.