Matrix Completion from a Few Entries

Raghunandan H. Keshavan[*], Andrea Montanari[*†], and Sewoong Oh[*]

March 17, 2009

Abstract

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

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

Further, if $r = O(1)$, M can be reconstructed *exactly* from $|E| = O(n \log n)$ entries. These results apply beyond random matrices to general low-rank incoherent matrices.

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.

1 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.

1.1 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 $n \times r$, and a diagonal matrix Σ, of dimensions $r \times r$ such that

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

[*]Department of Electrical Engineering, Stanford University

[†]Departments of Statistics, Stanford University

[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 [Net].

For justification of these assumptions and background on the use of low rank matrices in information retrieval, we refer to [BDJ99]. Since we are interested in very large data sets, we shall focus on the limit $m, n \to \infty$ with $m/n = \alpha$ bounded away from 0 and ∞.

We further assume that the factors U, V are unstructured. This notion is formalized by the *incoherence condition* introduced by Candés and Recht [CR08], and defined in Section 2. In particular the incoherence condition is satisfied with high probability if $M = U\Sigma V^T$ with U and V uniformly random matrices with $U^T U = m\mathbf{1}$ and $V^T V = n\mathbf{1}$. Alternatively, incoherence holds if the entries of U and V are i.i.d. bounded random variables.

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

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

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

1.2 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 $\mathsf{T}_r(M^E) = (mn/|E|)\sum_{i=1}^r \sigma_i x_i y_i^T$ obtained by setting to 0 all but the r largest singular values. Notice that, apart from the rescaling factor $(mn/|E|)$, $\mathsf{T}_r(M^E)$ 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 $\Theta(\log n/\log\log n)$ non-zero (revealed) entries. The largest singular values of M^E are of order $\Theta(\sqrt{\log n/\log\log n})$. The corresponding singular vectors are highly concentrated on high-weight column or row indices (respectively, for left and right singular vectors). Such singular vectors are an artifact of the high-weight columns/rows and do not provide useful information about the hidden entries of M. 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).

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 $2|E|/m$.

Figure 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 for real data.

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

SPECTRAL MATRIX COMPLETION(matrix M^E)
1: Trim M^E, and let $\widetilde{M}^E$ be the output;
2: Project $\widetilde{M}^E$ to $\mathsf{T}_r(\widetilde{M}^E)$;
3: Clean residual errors by minimizing the discrepancy $F(X,Y)$.

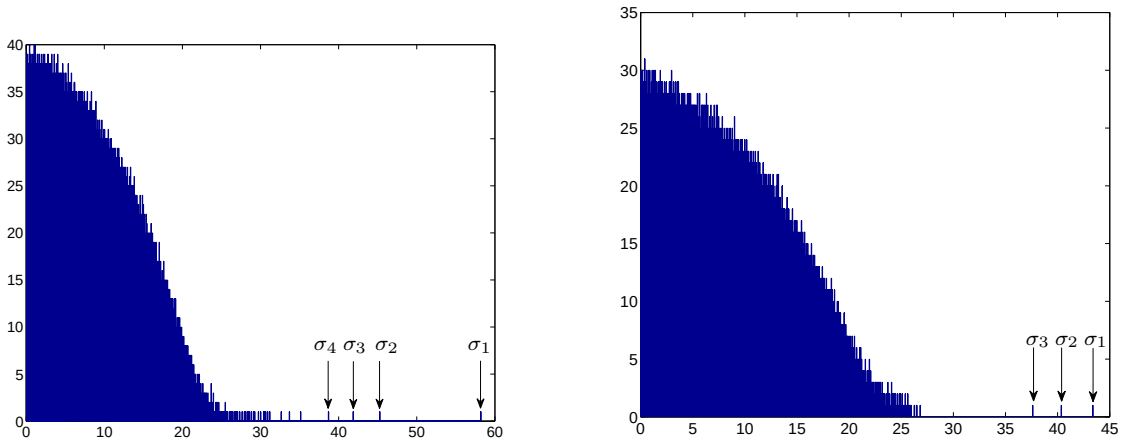

Figure 1: Histogram of the singular values of a partially revealed matrix M^E before trimming (left) and after trimming (right) for $10^4 \times 10^4$ random rank-3 matrix M with $\epsilon = 30$ and $\Sigma = \mathrm{diag}(1, 1.1, 1.2)$. After trimming the underlying rank-3 structure becomes clear. Here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

The last step of the above algorithm allows to reduce (or eliminate) small discrepancies between $\mathsf{T}_r(\widetilde{M}^E)$ and M, and is described below.

Cleaning. Various implementation 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 $X^TX = m\mathbf{1}$ and $Y^TY = n\mathbf{1}$, 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 $\mathsf{T}_r(\widetilde{M}^E) = X_0S_0Y_0^T$ 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 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 5 for background and references). Optimization over Grassmann manifolds is a well understood topic [EAS99] 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 $F(X,Y)$.

Finally, the implementation proposed here implicitly assumes that the rank r is known. In practice this is a non-issue. Since $r \ll n$, a loop over the value of r can be added at little extra cost. For instance, in collaborative filtering applications, r ranges between 10 and 30.

1.3 Main results

Notice that computing $\mathsf{T}_r(\widetilde{M}^E)$ 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

$$\text{RMSE} \equiv \left[\frac{1}{mnr}||M - \mathsf{T}_r(\widetilde{M}^E)||_{\text{F}}^2\right]^{1/2} . \tag{6}$$

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. The factor $(1/r)$ is instead necessary because (as described in Section 2), the typical size of the entries of M is $\sqrt{r}$.

Theorem 1.1. *Assume M to be a rank $r \le n^{1/2}$ matrix that satisfies the incoherence condition A2 (in particular, this is the case for random orthogonal matrices U, V). Then with high probability*

$$\frac{1}{mnr}||M - \mathsf{T}_r(\widetilde{M}^E)||_{\text{F}}^2 \le C(\alpha)\,\frac{nr}{|E|}\,. \tag{7}$$

The proof is provided in Section 3.

Notice that the top r singular values and singular vectors of the sparse matrix $\widetilde{M}^E$ can be computed efficiently by subspace iteration [Ber92]. Each iteration requires $O(|E|r)$ operations. As proved in Section 3, the $(r+1)$-th singular value is smaller than one half of the r-th 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 $\mathsf{T}_r(\widetilde{M}^E)$ and, for large enough $|E|$, reconstructs M exactly.

Theorem 1.2. *Assume M to be a rank $r \le n^{1/2}$ matrix that satisfies the incoherence conditions A1 and A2. Further, assume $\Sigma_{\min} \le \Sigma_1, \dots, \Sigma_r \le \Sigma_{\max}$ with $\Sigma_{\min}, \Sigma_{\max}$ bounded away from 0 and ∞. Then there exists $C'(\alpha)$ such that, if*

$$|E| \ge C'(\alpha) nr\, \max\{\log n, r\}\,, \tag{8}$$

then the cleaning procedure in Spectral Matrix Completion *converges, with high probability, to the matrix M.*

The proof is provided in Section 5. The basic intuition is that, for $|E| \ge C'(\alpha)nr\, \max\{\log n, r\}$, $T_r(\widetilde{M}^E)$ is so close to M that the cost function is well approximated by a quadratic function.

Theorem 1.1 is optimal: the number of degrees of freedom in M is of order nr, without the same number of observations is impossible to fix them. The extra $\log n$ factor in Theorem 1.2 is due to a coupon-collector effect [CR08, KMO08, KOM09]: it is necessary that E contains at least one entry per row and one per column and this happens only for $|E| \ge Cn \log n$. As a consequence, for rank r bounded, Theorem 1.2 is optimal. It is suboptimal by a polylogarithmic factor for $r = O(\log n)$.

1.4 Related work

Beyond collaborative filtering, low rank models are used for clustering, information retrieval, machine learning, and image processing. In [Faz02], the NP-hard problem of finding a matrix of minimum rank satisfying a set of affine constraints was addresses 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* [Don06, CRT06]. The connection with compressed sensing was emphasized in [RFP07], 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 [CR08] 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 [CR08] and proved in [KMO08], through a counting argument.

The present paper describes an efficient algorithm that reconstructs a rank-r matrix from $O(n\, r)$ random observations. 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 [CR08].

Cai, Candès and Shen [CCS08] recently proposed a low-complexity procedure to solve the convex program posed in [CR08]. Our spectral method is akin to a single step of this procedure, with the important novelty of the trimming step that improves significantly its performances. Our analysis techniques might provide a new tool for characterizing the convex relaxation as well.

Theorem 1.1 can also be compared with a copious line of work in the theoretical computer science literature [FKV04, AFK+01, AM07]. An important motivation in this context is the development of fast algorithms for low-rank approximation. In particular, Achlioptas and McSherry [AM07] 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 short account of our results was submitted to the 2009 International Symposium on Information Theory [KOM09]. While the present paper was under completion, Cándes and Tao posted online a preprint proving a theorem analogous to 1.2 [CT09]. Once more, their approach is substantially different from ours.

1.5 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 Spectral Matrix Completion algorithm suggest that the RMSE decays much faster with the number of observations per degree of freedom ($|E|/nr$), than indicated by Eq. (7). This improved behavior is a product 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 for r bounded. It would nevertheless be useful to derive quantitatively sharp estimates in this regime. A systematic numerical study was initiated in [KMO08]. It appears that available theoretical estimates (including the recent ones in [CT09]) are for larger values of the rank, we expect that our arguments can be strenghtened to prove exact reconstruction for $|E| \geq C'(\alpha) nr \log n$ for all values of r.

3. More general models. The model studied here and introduced in [CR08] presents obvious limitations. In applications to collaborative filtering, the subset of observed entries E is far from uniformly random. A recent paper [SC09] 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 non-incoherent matrices as well (as in [AM07]).

2 Incoherence property and some notations

The matrix M to be reconstructed takes the form (1) where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$. We write $U = [u_1, u_2, \dots, u_r]$ and $V = [v_1, v_2, \dots, v_r]$ 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 $\Sigma = \mathrm{diag}(\Sigma_1, \dots, \Sigma_r)$ with $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_r \geq 0$.

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

A1. There exists a constant $\mu_0 > 0$ such that for all $i \in [m]$, $j \in [n]$ we have $\sum_{k=1}^r U_{i,k}^2 \leq \mu_0 r$, $\sum_{k=1}^r V_{i,k}^2 \leq \mu_0 r$.

A2. There exists μ_1 such that $|\sum_{k=1}^r U_{i,k}\Sigma_k V_{j,k}| \leq \mu_1 r^{1/2}$.

Apart from difference in normalization, these assumptions coincide with the ones in [CR08].

In the proof of Theorem 1.1, we only require the second incoherence assumption A2. In the following, whenever we write that a property A holds with high probability (w.h.p.), we mean that there exists a function $f(n) = f(n; \alpha, \mu_1)$ such that $\mathbb{P}(A) \geq 1 - f(n)$ and $f(n) \to 0$ for μ_1 bounded away from 0 and ∞. In the case of exact completion (i.e. in the proof of Theorem 1.2) $f(\,\cdot\,)$ can also depend on μ_0, $\Sigma_{\min}$, $\Sigma_{\max}$, and $f(n) \to 0$ for $\mu_0, \Sigma_{\min}, \Sigma_{\max}$ bounded away from 0 and ∞.

Probability is taken with respect to the uniformly random subset $E \subseteq [m] \times [n]$. It is convenient to work with a model in which each entry is revealed independently with probability $\epsilon/\sqrt{mn}$. Since, with high probability $|E| \in [\epsilon\sqrt{\alpha}\, n - A\sqrt{n\log n}, \epsilon\sqrt{\alpha}\, n - A\sqrt{n\log n}]$, any guarantee on the algorithm performances that holds within one model, holds within the other model as well if we allow for a vanishing shift in ϵ. Finally, we will use C, C' etc. to denote generic constants that depend uniquely on α, μ_1 and, when proving Theorem 1.2, μ_0, $\Sigma_{\min}$, $\Sigma_{\max}$.

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. Finally, we use the standard combinatorics notation $[N] = \{1, 2, \dots, N\}$ to denote the set of first N integers.

3 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 Eq. (3). We shall then redefine $\{\sigma_i\}$, $\{x_i\}$, $\{y_i\}$, by letting

$$\widetilde{M}^E = \sum_{i=1}^{\min(m,n)} \sigma_i x_i y_i^T \,. \tag{9}$$

Here $||x_i|| = ||y_i|| = 1$, $x_i^T x_j = y_i^T y_j = 0$ for $i \neq j$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

We will prove Theorem 1.1 by controlling the operator norm of $M - \mathsf{T}_r(\widetilde{M}^E)$.

Theorem 3.1. *There exists $C(\alpha) < \infty$ such that, with high probability,*

$$\frac{1}{\sqrt{mn}}\big|\big|M - \mathsf{T}_r(\widetilde{M}^E)\big|\big|_2 \leq C\left(\frac{r}{\epsilon}\right)^{1/2} \,. \tag{10}$$

This result is proved later in this section. The proof of Theorem 1.1 is a direct application the of above theorem.

Proof. (Theorem 1.1) Since $M - \mathsf{T}_r(\widetilde{M}^E)$ has rank at most $2r$, it immediately follows that

$$\begin{aligned} \frac{1}{\sqrt{mnr}}\left|\left|M - \mathsf{T}_r(\widetilde{M}^E)\right|\right|_F &\leq \sqrt{2r}\left(\frac{1}{\sqrt{mnr}}\left|\left|M - \mathsf{T}_r(\widetilde{M}^E)\right|\right|_2\right) \\ &\leq C\left(\frac{r}{\epsilon}\right)^{1/2}, \end{aligned}$$

which implies the thesis. □

To prove Theorem 3.1, we will use following key technical results.

Lemma 3.2. *There exists a constant $C > 0$ such that, with high probability*

$$\left|x^T\left(\frac{\epsilon}{\sqrt{mn}}M - \widetilde{M}^E\right)y\right| \leq C\sqrt{r\epsilon}\,, \tag{11}$$

for any $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ such that $||x|| = ||y|| = 1$.

The proof of this lemma is given in Section 4. The next lemma, which is a direct consequence of Lemma 3.2, relates the singular values of the trimmed matrix $\widetilde{M}^E$ to the singular values of M.

Lemma 3.3. *There exists a constant $C > 0$ such that, with high probability*

$$\left|\frac{\sigma_q}{\epsilon} - \Sigma_q\right| \leq C\left(\frac{r}{\epsilon}\right)^{1/2}, \tag{12}$$

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

This lemma is proved later in this section. We can now prove Theorem 3.1.

Proof. (Theorem 3.1) Consider $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ such that $||x|| = ||y|| = 1$. Then,

$$\begin{aligned} \left|x^T(M - \mathsf{T}_r(\widetilde{M}^E))y\right| &\leq \left|x^T\left(M - \frac{\sqrt{mn}}{\epsilon}\widetilde{M}^E\right)y\right| + \left|x^T\left(\frac{\sqrt{mn}}{\epsilon}\widetilde{M}^E - \mathsf{T}_r(\widetilde{M}^E)\right)y\right| \\ &\leq C\sqrt{mn}\left(\frac{r}{\epsilon}\right)^{\frac{1}{2}} + \sqrt{mn}\frac{\sigma_{r+1}}{\epsilon} \\ &\leq C'\sqrt{mn}\left(\frac{r}{\epsilon}\right)^{\frac{1}{2}}, \end{aligned}$$

where we used Lemma 3.2 for the second inequality and Lemma 3.3 for the last inequality. This implies that $\frac{1}{\sqrt{mn}}\left|\left|M - \mathsf{T}_r(\widetilde{M}^E)\right|\right|_2 \leq C'\,(r/\epsilon)^{\frac{1}{2}}$. □

We end this section with the proof of Lemma 3.3.

Proof. (Lemma 3.3) Recall the variational characterization of the singular values.

$$\sigma_q = \min_{H,\dim(H)=n-q+1}\ \max_{y\in H,||y||=1} ||\widetilde{M}^E y|| \tag{13}$$

$$= \max_{H,\dim(H)=q}\ \min_{y\in H,||y||=1} ||\widetilde{M}^E y||\,. \tag{14}$$

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

Using Eq. (13) with H the orthogonal complement of $\mathrm{span}(v_1, \ldots, v_{q-1})$, we have, by Lemma 3.2,

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

In order to get a lower bound, we use Eq. (14) with $H = \mathrm{span}(v_1, \ldots, v_q)$. By Lemma 3.2 we have

$$\begin{aligned}
\sigma_q &\geq \min_{y \in H, ||y||=1} ||\widetilde{M}^E y|| \\
&\geq \frac{\epsilon}{\sqrt{mn}} \left(\min_{y \in H, ||y||=1} ||My|| \right) - \max_{y \in H, ||y||=||x||=1} \left| x^T \left(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}} M \right) y \right| \\
&\geq \epsilon \Sigma_q - C\sqrt{r\epsilon}
\end{aligned}$$

Combining the upper and lower bounds, we get the desired result. □

4 Proof of Lemma 3.2

We want to show that $|x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}} M)y| \leq C\sqrt{r\epsilon}$ for each $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ such that $||x|| = ||y|| = 1$. Following the technique of [FKS89], this will be done by first reducing ourselves to x, y belonging to finite sets. We define

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

Notice that $T_n \subseteq S_n \equiv \{x \in \mathbb{R}^n : ||x|| \leq 1\}$. The next two remarks are proved in [FKS89, FO05], and relate the original problem to the discretized one.

Remark 4.1. *Let $R \in \mathbb{R}^{m \times n}$ be a matrix. If $|x^T R y| \leq B$ for all $x \in T_m$ and $y \in T_n$, then $|x'^T R y'| \leq (1-\Delta)^{-2} B$ for all $x' \in S_m$ and $y' \in S_n$.*

Remark 4.2. $|T_m| \leq (10/\Delta)^m$.

Hence it is enough to show that $|x^T(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}} M)y| \leq C\sqrt{r\epsilon}$ for all $x \in T_m$ and $y \in T_n$. There are two parts to the proof of this claim. One bounds the contribution of *light couples* $L \subseteq [m] \times [n]$, defined as

$$L = \left\{ (i,j) : |x_i M_{ij} y_j| \leq \left(\frac{r\epsilon}{mn} \right)^{1/2} \right\},$$

and the other bounds the contribution of its complement $\overline{L}$, which we call *heavy couples*. We have

$$\left| x^T \left(\widetilde{M}^E - \frac{\epsilon}{\sqrt{mn}} M \right) y \right| \leq \left| \sum_{(i,j) \in L} x_i \widetilde{M}^E_{ij} y_j - \frac{\epsilon}{\sqrt{mn}} x^T M y \right| + \left| \sum_{(i,j) \in \overline{L}} x_i \widetilde{M}^E_{ij} y_j \right| \tag{15}$$

In the next two subsections, we will prove that the first contribution is upper bounded by $C_1\sqrt{r\epsilon}$ and the second by $C_2\sqrt{r\epsilon}$ 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.

4.1 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. 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 4.3. *There exists C_1 and C_2 depending only on α such that, with probability larger than $1 - 1/n^3$, $|\mathcal{A}_l| \ge m - \max\{e^{-C_1\epsilon}m, C_2\alpha\}$, 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

$$\mathcal{H}(E, A) = \left\{ \exists\, x, y \,:\, \left| \sum_{(i,j)\in L} x_i M^{E,A}_{ij} y_j - \frac{\epsilon}{\sqrt{mn}} x^T M y \right| > C_1\sqrt{r\epsilon} \right\}\,, \tag{16}$$

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

$$\begin{aligned}\mathbb{P}\left\{\mathcal{H}(E,\mathcal{A})\right\} &= \sum_A \mathbb{P}\left\{\mathcal{H}(E,A),\, \mathcal{A} = A\right\}\\ &\le \sum_{\substack{|A_l|\ge m(1-\delta),\\ |A_r|\ge n(1-\delta)}} \mathbb{P}\left\{\mathcal{H}(E,A),\, \mathcal{A} = A\right\} + \frac{1}{n^3}\\ &\le 2^{(n+m)H(\delta)} \max_{\substack{|A_l|\ge m(1-\delta),\\ |A_r|\ge n(1-\delta)}} \mathbb{P}\left\{\mathcal{H}(E;A)\right\} + \frac{1}{n^3}\,,\end{aligned} \tag{17}$$

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 $\mathbb{P}\left\{\mathcal{H}(E;A)\right\}$ uniformly over A where $\mathcal{H}$ is defined as in Eq. (16). The key step consists in proving the following tail estimate

Lemma 4.4. *Let $x \in S_m$, $y \in S_n$, $Z = \sum_{(i,j)\in L} x_i M^{E,A}_{ij} y_j - \frac{\epsilon}{\sqrt{mn}} x^T M y$, and assume $|A_l| \ge m(1-\delta)$, $|A_r| \ge n(1-\delta)$ with δ small enough. Then*

$$\mathbb{P}\left(Z > L\sqrt{r\epsilon}\right) \le \exp\Big\{ -\frac{n\alpha^{1/2}}{2}\Big(L - 2\mu_1^2 - \mu_1\Big)\Big\}\,.$$

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

Remark 4.5. $|\mathbb{E}\left[Z\right]| \le \left(\mu_1 + \mu_1^2\right)\sqrt{r\epsilon}$.

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

$$\begin{aligned} a_{ij} &= \begin{cases} x_i M^A_{ij} y_j & \text{if } |x_i M^A_{ij} y_j| \le (r\epsilon/mn)^{1/2}, \\ 0 & \text{otherwise,} \end{cases} \\ Z_{ij} &= \begin{cases} a_{i,j} & \text{w.p. } \epsilon/\sqrt{mn}, \\ 0 & \text{w.p. } 1-\epsilon/\sqrt{mn}, \end{cases} \end{aligned}$$

Let $Z_1 = \sum_{i,j} Z_{ij}$ so that $Z = Z_1 - \frac{\epsilon}{\sqrt{mn}} x^T M y$. Note that $\sum_{i,j} a_{ij}^2 \le \sum_{i,j} \left(x_i M^A_{ij} y_j \right)^2 \le \mu_1^2 r$. Fix $\lambda = (mn/4r\epsilon)^{1/2}$ so that $|\lambda a_{i,j}| \le 1/2$, whence $e^{\lambda a_{ij}} - 1 \le \lambda a_{ij} + 2(\lambda a_{ij})^2$. It then follows that

$$\begin{aligned} \mathbb{E}[e^{\lambda Z}] &= \exp\Big\{ \frac{\epsilon}{\sqrt{mn}} \Big(\sum_{i,j} \lambda a_{i,j} + 2 \sum_{i,j} (\lambda a_{i,j})^2 \Big) - \frac{\lambda\,\epsilon}{\sqrt{mn}} x^T M y \Big\} \\ &\le \exp\left\{ \lambda \mathbb{E}[Z] + 2r\epsilon\mu_1^2 \frac{\lambda^2}{n\alpha^{1/2}} \right\}. \end{aligned}$$

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

Note that $\mathbb{P}\left(-Z > L\sqrt{r\epsilon}\right)$ can also be bounded analogously. We can now finish the upper bound on the light couples contribution. Consider the error event Eq. (16). Using Remark 4.2, we can apply union bound over T_m and T_n to Eq. (17) to obtain

$$\mathbb{P}\{\mathcal{H}(E, \mathcal{A})\} \le 2 \left(\frac{20}{\Delta} \right)^{n+m} e^{-(C_1 - 2\mu_1^2 - \mu_1)\alpha^{1/2} n/2} + \frac{1}{n^3},$$

If C_1 is a large enough constant, the first term is of order $e^{-\Theta(n)}$ (for, say, $\epsilon \ge r$) thus finishing the proof.

4.2 Bounding the contribution of heavy couples

Let Q be an $m \times n$ matrix with $Q_{ij} = 1$ if $(i,j) \in E$ and $i \notin \mathcal{A}_r$, $j \notin \mathcal{A}_l$ (i.e. entry (i,j) is not trimmed by our algorithm), and $Q_{ij} = 0$ otherwise. Due to the incoherence assumption A2, $|M_{ij}| \le \mu_1 r^{1/2}$ and therefore the heavy couples satisfy $|x_i y_j| \ge \sqrt{\epsilon/(\mu_1^2 mn)} = C\sqrt{\epsilon}/n$. We then have

$$\begin{aligned} \left| \sum_{(i,j)\in\overline{L}} x_i \widetilde{M}^E_{ij} y_j \right| &\le \mu_1 \sqrt{r} \sum_{(i,j)\in\overline{L}} Q_{ij} |x_i y_j| \\ &\le \mu_1 \sqrt{r} \sum_{\substack{(i,j)\in E: \\ |x_i y_j| \ge C\sqrt{\epsilon}/n}} Q_{ij} |x_i y_j|. \end{aligned}$$

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 [FO05].

Remark 4.6. *Given vectors x, y, let $\overline{L}' = \{(i,j) \in E : |x_i y_j| \ge C\sqrt{\epsilon}/n\}$. Then there exist a constant C' such that, w.h.p., $\sum_{(i,j)\in\overline{L}'} Q_{ij} |x_i y_j| \le C'\sqrt{\epsilon}$, for all $x \in T_m$, $y \in T_n$.*

For the reader's convenience, a proof of this fact is proposed in Appendix C. The analogous result in [FO05] (for the adjacency matrix of a non-bipartite 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 for large enough C the contribution of heavy couples is, w.h.p., bounded by $C_2\sqrt{r\epsilon}$ for some $C_2 < \infty$.

5 Minimization on Grassmann manifolds and proof of Theorem 1.2

The function $F(X, Y)$ defined in Eq. (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 [EAS99] 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 section Sections 6, 7, 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 ∞. Constants (denoted by $C, C', \dots$) depend implicitly on $\Sigma_{\min}$, $\Sigma_{\max}$. Finally, throughout this section, we use the notation $X^{(i)} \in \mathbb{R}^r$ to refer to the i-th row of the matrix $X \in \mathbb{R}^{m\times n}$ or $X \in \mathbb{R}^{n\times r}$.

5.1 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 in the manifold is the equivalence class of an $n \times r$ orthogonal matrix A

$$[A] = \{AQ : Q \in \mathsf{O}(r)\}\,. \tag{18}$$

For consistency with the rest of the paper, we will assume the normalization $A^TA = n\,\mathbf{1}$. 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 $\mathsf{M}(m, n) \equiv \mathsf{G}(m, r) \times \mathsf{G}(n, r)$:

$$F : \mathsf{M}(m, n) \to \mathbb{R}\,, \tag{19}$$

$$([X], [Y]) \mapsto F(X, Y)\,. \tag{20}$$

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 $W^TX = Z^TY = 0$. The 'canonical' Riemann metric on the Grassmann manifold corresponds to the usual scalar product $\langle W, W'\rangle \equiv \mathrm{Tr}(W^TW')$. 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

$$d(\mathbf{x}_1, \mathbf{x}_2) \equiv \sqrt{d(X_1, X_2)^2 + d(Y_1, Y_2)^2} \tag{21}$$

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 $X_1^T X_2/n$. Then

$$d(X_1, X_2) = ||\theta||_2 \,. \tag{22}$$

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:

$$d_{\rm c}(X_1, X_2) = \frac{1}{\sqrt{n}} \min_{Q_1, Q_2 \in \mathsf{O}(r)} ||X_1 Q_1 - X_2 Q_2||_F \qquad \text{(chordal distance)}, \tag{23}$$

$$d_{\rm p}(X_1, X_2) = \frac{1}{\sqrt{2}n} ||X_1 X_1^T - X_2 X_2^T||_F \qquad \text{(projection distance)}. \tag{24}$$

Notice that $d_{\rm c}$ and $d_{\rm p}$ 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. $d_{\rm c}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{d_{\rm c}(X_1, X_2)^2 + d_{\rm c}(Y_1, Y_2)^2}$.

Remark 5.1. *The geodesic, chordal and projection distance are equivalent, namely*

$$\frac{2}{\pi} d(X_1, X_2) \le \frac{1}{\sqrt{2}} d_{\rm c}(X_1, X_2) \le d_{\rm p}(X_1, X_2) \le d_{\rm c}(X_1, X_2) \le d(X_1, X_2) \,. \tag{25}$$

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

An important remark is that geodesics with respect to the canonical Riemann metric admit an explicit and efficiently computable form. Given $\mathbf{u} \in \mathsf{M}(m, n)$, $\mathbf{w} \in \mathsf{T}_{\mathbf{x}}$ 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 $W = L\Theta R^T$ [EAS99]:

$$X(t) = UR\cos(\Theta t)R^T + L\sin(\Theta t)R^T \,, \tag{26}$$

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

5.2 Gradient and incoherence

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

$$F(\mathbf{x}(t)) = F(\mathbf{x}) + \langle \operatorname{grad} F(\mathbf{x}), \mathbf{w} \rangle \, t + O(t^2) \,. \tag{27}$$

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

$$\mathcal{P}_E(M)_{ij} = \begin{cases} M_{ij} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

The two components of the gradient are then

$$\operatorname{grad} F(\mathbf{x})_X = \mathcal{P}_E(XSY^T - M)YS^T - XQ_X \,, \tag{29}$$

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

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

$$Q_X = \frac{1}{m} X^T \mathcal{P}_E(M - XSY^T)YS^T \,, \tag{31}$$

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

5.3 Algorithm

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

$$\begin{aligned} \widetilde{F}(X,Y) &= F(X,Y) + \rho\, G(X,Y) && (33)\\ &\equiv F(X,Y) + \rho \sum_{i=1}^{m} G_1\left(\frac{||X^{(i)}||^2}{2\mu_0 r}\right) + \rho \sum_{j=1}^{n} G_1\left(\frac{||Y^{(j)}||^2}{2\mu_0 r}\right), && (34)\end{aligned}$$

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

$$G_1(z) = \begin{cases} 0 & \text{if } z \le 1, \\ e^{(z-1)^2} - 1 & \text{if } z \ge 1. \end{cases} \tag{35}$$

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

$$\mathcal{K}(\mu') \equiv \left\{ (X,Y) \text{ such that } ||X^{(i)}||^2 \le \mu' r,\ ||Y^{(j)}||^2 \le \mu' r \right\}. \tag{36}$$

We have $G(X,Y) = 0$ on $\mathcal{K}(2\mu_0)$. Notice that $\mathbf{u} \in \mathcal{K}(\mu_0)$ by the incoherence property. Without loss of generality we can assume $\mathbf{x}_0 \in \mathcal{K}(2\mu_0)$, because otherwise we can rescale all lines of X_0, Y_0 that violate the constraint.

GRADIENT DESCENT(matrix M^E, factors $\mathbf{x}_0$)
For $k = 0, 1, \dots$ do:
Compute $\mathbf{w}_k = \operatorname{grad} \widetilde{F}(\mathbf{x}_k)$;
Let $t \mapsto \mathbf{x}_k(t)$ be the geodesic with $\mathbf{x}_k(t) = \mathbf{x}_k + \mathbf{w}_k t + O(t^2)$;
Minimize $t \mapsto \widetilde{F}(\mathbf{x}_k(t))$ for $t \ge 0$, subject to $d(\mathbf{x}_k(t), \mathbf{x}_0) \le \gamma$;
Set $\mathbf{x}_{k+1} = \mathbf{x}_k(t_k)$ where t_k is the minimum location;
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 5.2. *Let $U, X \in \mathbb{R}^{n\times r}$ with $U^T U = X^T X = m\mathbf{1}$, $V, Y \in \mathbb{R}^{m\times r}$ with $V^T V = Y^T Y = n\mathbf{1}$, and $M = U\Sigma V^T$, $\widehat{M} = XSY^T$ for $\Sigma = \operatorname{diag}(\Sigma_1, \dots, \Sigma_r)$ and $S \in \mathbb{R}^{r\times r}$. If $\Sigma_1, \dots, \Sigma_r \ge \Sigma_{\min}$, then*

$$d_{\rm p}(U,X) \le \frac{1}{m\Sigma_{\min}} ||M - \widehat{M}||_F \quad , \quad d_{\rm p}(V,Y) \le \frac{1}{n\Sigma_{\min}} ||M - \widehat{M}||_F \tag{37}$$

As a consequence of this remark and Theorem 1.1, we can assume that $d(\mathbf{u}, \mathbf{x}_0) \le C\, r/\sqrt{\epsilon}$. We shall then set $\gamma = C' r/\sqrt{\epsilon}$ (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.

(*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 5 (which might appear complex to implement) can be replaced by a standard step selection procedure, such as the one in [Arm66].

(*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).

5.4 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 5.3. *Assume $\epsilon \ge A \max\{r \log n, r^2\}$ with A large enough. Then there exists constants $C_1, C_2, \delta > 0$ (independent of m, n, ϵ and r) such that, with high probability*

$$C_1\, n\epsilon \left(d(\mathbf{x},\mathbf{u})^2 + ||S - \Sigma||_F^2\right) \le F(\mathbf{x}) \le C_2\, n\epsilon\, d(\mathbf{x},\mathbf{u})^2 \tag{38}$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(3\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \le \delta$ (where $S \in \mathbb{R}^{r\times r}$ is the matrix realizing the minimum in Eq. (4)).

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

Lemma 5.4. *Assume $\epsilon \ge A \max\{r \log n, r^2\}$ with A large enough. Then there exists constants $C, \delta > 0$ (independent of m, n, ϵ and r) such that, with high probability*

$$||\mathrm{grad}\, \widetilde{F}(\mathbf{x})||^2 \ge C\, n\epsilon^2\, d(\mathbf{x},\mathbf{u})^2 \tag{39}$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(3\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \le \delta$.

We can now prove Theorem 1.2.

Proof. (Theorem 1.2) Let $\delta > 0$ be such that Lemma 5.3 and Lemma 5.4 are verified, and C_1, C_2 be defined as in Lemma 5.3. We further assume $\delta \le \sqrt{(e^{1/4}-1)/C_1}$. Take ϵ large enough that $d(\mathbf{u},\mathbf{x}_0) \le C\, r/\sqrt{\epsilon} \le \min(1, (C_1/2C_2)^{1/2})\delta/10$. Further, set the algorithm parameter to $\gamma = \delta/4$.

We make the following claims:

1. $\mathbf{x}_k \in \mathcal{K}(3\mu_0)$ for all k.

   Indeed $\mathbf{x}_0 \in \mathcal{K}(2\mu_0)$ whence $\widetilde{F}(\mathbf{x}_0) = F(\mathbf{x}_0) \le C_2 n\epsilon\, \delta^2$. The claim follows because $\widetilde{F}(\mathbf{x}_k)$ is non-increasing and $\widetilde{F}(\mathbf{x}) \ge \rho\, G(X,Y) \ge n\epsilon(e^{1/4}-1)$ for $\mathbf{x} \notin \mathcal{K}(3\mu_0)$.

2. $d(\mathbf{x}_k,\mathbf{u}) \le \delta/10$ for all k.

   Indeed by triangular inequality we can assume to have $d(\mathbf{x}_k,\mathbf{u}) \le \delta/2$. Since $d(\mathbf{x}_0,\mathbf{u}) \le (C_1/2C_2)^{1/2}\delta/10$, 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 non-increasing 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 discent with exact line search. Therefore [Arm66] this must converge to the unique stationary point of $\widetilde{F}$ in $\mathcal{K}(3\mu_0) \cap \{\mathbf{x} :\, d(\mathbf{x},\mathbf{u}) \le \delta/10$, which, by Lemma 5.4, is $\mathbf{u}$. □

6 Proof of Lemma 5.3

We will make use of the following Lemma.

Lemma 6.1. *Assume $\epsilon = A \log n$ with A large enough. Then there exists $C > 0$ such that with high probability*

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

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

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

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

$$x_0 \max_j \deg(j) ||y||_1 \le \max_j \deg(j) ||x||_1 ||y||_1/n \,. \tag{42}$$

For $\epsilon = A \log n$, the maximum degree is with high probability of the same order as the average one, and therefore this term is at most $C\epsilon||x||_1||y||_1/n$.

The second term is upper bounded by $C\sqrt{\epsilon}||x'||_2||y||_2$ using Theorem 1.1 in [FO05] or, equivalently, Theorem 3.3 in the case $r = 1$. The thesis follows because $||x'||_2 \le ||x||_2$. □

Proof. (Lemma 5.3) Throughout the proof we assume $m = n$ to simplify notations.

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 $L^T L = n\mathbf{1}$ and $R^T R = \mathbf{1}$), the explicit expression for geodesics in Eq. (26) yields

$$X = U + \overline{W} \,, \qquad \overline{W} = UR(\cos\Theta - \mathbf{1})R^T + L \sin\Theta R^T \,. \tag{43}$$

An analogous expression can obviously be written for $Y = V + \overline{Z}$. Notice that, if $\mathbf{u}, \mathbf{x} \in \mathcal{K}(3\mu_0)$, then $(\overline{W}, \overline{Z}) \in \mathcal{K}(12\mu_0)$ and $\mathbf{w} \in \mathcal{K}(48\mu_0/\pi^2)$. In the first case this follows from $||\overline{W}^{(i)}||^2 \le 2||X^{(i)}||^2 + 2||U^{(i)}||^2$. In order to prove $\mathbf{w} \in \mathcal{K}(48\mu_0/\pi^2)$, we notice that

$$\begin{aligned} ||W^{(i)}||^2 &= ||\Theta L^{(i)}||^2 \le \frac{4}{\pi^2} ||\sin\Theta L^{(i)}||^2 \\ &\le \frac{4}{\pi^2} ||X^{(i)} - R\cos\Theta R^T U^{(i)}||^2 \le \frac{8}{\pi^2}\Big(||X^{(i)}||^2 + ||U^{(i)}||^2\Big) \,. \end{aligned}$$

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

Denote by $S \in \mathbb{R}^{r\times r}$ the matrix realizing the minimum in Eq. (4). We will start by proving the lower bound in Eq. (38):

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

where in we used Cauchy-Schwarz inequality to argue that $(1/2)(A+B)^2 \geq (A^2/4) - B^2$ and defined

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

We will show that, with high probability $A^2 \geq Cn\epsilon||S-\Sigma||_F^2 + Cn\epsilon d(\mathbf{x},\mathbf{u})^2$ and $B^2 \leq (C/100)n\epsilon\,(1+||S-\Sigma||_F^2)d(\mathbf{x},\mathbf{u})^2$ whence the lower bound in Eq. (38) follows.

Lower bound on A. By Theorem 4.1 in [CR08], we have $A^2 \geq (1-\xi)\mathbb{E}\{A^2\}$ with high probability for each $\xi > 0$. Further

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

The first term is equal to $n\epsilon||S-\Sigma||_F^2$. The second and third terms are lower bounded by

$$\epsilon\sigma_{\min}(S)^2(||\overline{Z}||_F^2 + ||\overline{W}||_F^2) \geq C\sigma_{\min}(S)^2 n\epsilon d_{\rm c}(\mathbf{x},\mathbf{u})^2 \geq C'\sigma_{\min}(S)^2 n\epsilon d(\mathbf{x},\mathbf{u})^2 \,.$$

The absolute value of the fourth term can be written as

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

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

$$||U^T\overline{W}||_F = n||\cos\theta - 1|| = \frac{n}{2}\,||4\sin^2(\theta/2)|| \leq \frac{n}{2}\,||2\sin(\theta/2)||^2 \tag{44}$$

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

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

Proceeding analogously for the other terms in the expression of $\mathbb{E}\{A^2\}$, we proved that with high probability

$$\begin{aligned} A^2 &\geq n\epsilon||S-\Sigma||_F^2 + n\epsilon\big(C\sigma_{\min}(S)^2 - C'\delta^2\,\sigma_{\max}(S)^2\big)\, d(\mathbf{x},\mathbf{u})^2 &(45)\\ &\geq n\epsilon(1 - Cd(\mathbf{x},\mathbf{u})^2)||S-\Sigma||_F^2 + n\epsilon\big(C\Sigma_{\min}^2 - C\delta^2\,\Sigma_{\max}^2\big)\, d(\mathbf{x},\mathbf{u})^2 \,, &(46)\end{aligned}$$

where we used the bounds $\sigma_{\min}(S)^2 \geq \Sigma_{\min}^2/2 - ||S-\Sigma||_F^2$ and $\sigma_{\max}(S)^2 \leq 2\Sigma_{\max}^2 + 2\,||S-\Sigma||_F^2$. The above inequality implies the desired claim if we take $d(\mathbf{x},\mathbf{u}) \leq \delta$ small enough.

Upper bound on B. By Lemma 6.1 we have

$$
\begin{aligned}
B^2 &\le \sigma_{\max}(S)^2 \sum_{a,b} \sum_{(i,j)\in E} \overline{W}_{ia}^2 \overline{Z}_{jb}^2 && (47)\\
&\le \frac{C\epsilon}{n}\sigma_{\max}(S)^2 \sum_{i,j} ||\overline{W}^{(i)}||^2 ||\overline{Z}^{(j)}||^2 + C\sigma_{\max}(S)^2\sqrt{\epsilon}\sum_{a,b}\left(\sum_i ||\overline{W}^{(i)}||^4\right)^{1/2}\left(\sum_j ||\overline{Z}^{(j)}||^4\right)^{1/2} && (48)\\
&\le \frac{C\epsilon}{n}\sigma_{\max}(S)^2 \sum_{i,j} ||\overline{W}^{(i)}||^2 ||\overline{Z}^{(j)}||^2 + C'\sigma_{\max}(S)^2\sqrt{\epsilon}r\left(\sum_i ||\overline{W}^{(i)}||^2\right)^{1/2}\left(\sum_j ||\overline{Z}^{(j)}||^2\right)^{1/2} && (49)
\end{aligned}
$$

where in the last step we used the incoherence condition. We conclude therefore that

$$
\begin{aligned}
B^2 &\le n\sigma_{\max}(S)^2\left(C\epsilon\,\delta^2 + C'\sqrt{\epsilon}r\right) d(\mathbf{x},\mathbf{u})^2 && (50)\\
&\le Cn\epsilon(\Sigma_{\max}^2 + ||S-\Sigma||_F^2)\max(r/\sqrt{\epsilon},\delta^2) d(\mathbf{x},\mathbf{u})^2 && (51)
\end{aligned}
$$

which implies the thesis for $r/\sqrt{\epsilon}, \delta$ small enough.

In order to prove the upper bound in Eq. (38) we can set $\Sigma = S$, thus obtaining

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

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 [CR08], we have $\widehat{A}^2 \le (1-\xi)\mathbb{E}\{\widehat{A}^2\}$ and

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

Further by setting $S = \Sigma$ in the derivation for B we get $\widehat{B}^2 \le (C/100)n\epsilon\, d(\mathbf{x},\mathbf{u})^2$. □

7 Proof of Lemma 5.4

Throughout this proof we assume $m = n$ to lighten the notation.

As in the proof of Lemma 5.3, 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 $W = L\Theta R^T$, we obtain

$$\widehat{W} = -UR\Theta \sin\Theta\, R^T + L\Theta\cos\Theta\, R^T\,. \tag{52}$$

An analogous expression holds for $\widehat{Z}$. Since $L^TU = 0$, we have $||\widehat{W}||_F^2 = n||\Theta\sin\Theta||_F^2 + n||\Theta\cos\Theta||_F^2 = n||\theta||^2$. Hence $||\widehat{\mathbf{w}}||^2 = nd(\mathbf{x},\mathbf{u})^2$. In order to prove the thesis, it is therefore sufficient to show that $\langle \mathrm{grad}\,\widetilde{F}(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq Cn\epsilon\, d(\mathbf{x},\mathbf{u})^2$. In the following we will indeed show that $\langle \mathrm{grad}\, F(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq Cn\epsilon\, d(\mathbf{x},\mathbf{u})^2$, and $\langle \mathrm{grad}\, G(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq 0$.

As a preliminary remark, we notice that $\widehat{\mathbf{w}} \in \mathcal{K}((3\pi^2/2 + 96/\pi^2)\mu_0)$. Indeed

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

By assumption we have $||U^{(i)}||^2 \leq 3\mu_0 r$ and we proved $||W^{(i)}||^2 \leq 48\mu_0 r/\pi^2$ in the previous section.

7.1 Lower bound on $\mathrm{grad}\, F(\mathbf{x})$

Recalling that $\mathcal{P}_E$ is the projector defined in Eq. (28), and using the expression (29), (30), for the gradient, we have

$$\begin{aligned}
\langle \mathrm{grad}\, F(\mathbf{x}), \widehat{\mathbf{w}}\rangle &= \langle \mathcal{P}_E(XSY^T - M), (XS\widehat{Z}^T + \widehat{W}SY^T)\rangle \\
&= \langle \mathcal{P}_E(U(S-\Sigma)V^T + US\overline{Z}^T + \overline{W}SV^T + \overline{W}S\overline{Z}^T), (US\widehat{Z}^T + \widehat{W}SV^T + \overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle \\
&\geq A - B_1 - B_2 - B_3
\end{aligned} \tag{54}$$

where we defined

$$\begin{aligned}
A &= \langle \mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T), (US\widehat{Z}^T + \widehat{W}SV^T)\rangle\,, &(55)\\
B_1 &= |\langle \mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T), (\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle|\,, &(56)\\
B_2 &= |\langle \mathcal{P}_E(U(S-\Sigma)V^T + \overline{W}S\overline{Z}^T), (US\widehat{Z}^T + \widehat{W}SV^T)\rangle|\,, &(57)\\
B_3 &= |\langle \mathcal{P}_E(U(S-\Sigma)V^T + \overline{W}S\overline{Z}^T), (\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)\rangle|\,. &(58)
\end{aligned}$$

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. 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 $\overline{Z}$. Before proceeding, it is worth deriving a few estimates of this type. In particular, using Eqs. (43) and (52) we get

$$\begin{aligned}
||\widehat{W}||_F^2 &= ||W||_F^2 = n||\theta||^2\,, &(59)\\
||\overline{W}||_F^2 &= n||2\sin\theta/2||^2\,, &(60)\\
\langle \widehat{W}, \overline{W}\rangle &= n\sum_{a=1}^r \theta_a \sin\theta_a\,, &(61)\\
\langle \widehat{W}, W\rangle &= n\sum_{a=1}^r \theta_a^2\cos\theta_a\,, &(62)
\end{aligned}$$

and therefore

$$\begin{aligned}
||\widehat{W} - \overline{W}||_F^2 &= n\sum_{i=a}^r [(2\sin\theta_a/2)^2 + \theta_a^2 - 2\theta_a\sin\theta_a] &(63)\\
&\leq n\sum_{i=a}^r (\theta_a - 2\sin\theta_a/2)^2 \leq \frac{n}{24^2}||\theta||^4 \leq \frac{n}{24^2} d(\mathbf{u},\mathbf{x})^4\,. &(64)
\end{aligned}$$

Analogously

$$||\widehat{W} - W||_F^2 \;=\; n\sum_{i=a}^{r}[2\theta_a^2 - 2\theta_a^2\cos\theta_a] \le n\,||\theta||^4 \le n\,d(\mathbf{u},\mathbf{x})^4 \tag{65}$$

The last inequality implies in particular

$$||U^T\widehat{W}||_F = ||U^T(W-\widehat{W})||_F \le nd(\mathbf{u},\mathbf{x})^2\,. \tag{66}$$

Similar bounds hold of course for $Z, \widehat{Z}, \overline{Z}$ (for instance we have $||V^T\widehat{Z}||_F \le nd(\mathbf{u},\mathbf{x})^2$). Finally, we shall use repeatedly the fact that $||S-\Sigma||_F^2 \le Cd(\mathbf{x},\mathbf{u})^2$, which follows from Lemma 5.3. This in turns implies

$$\sigma_{\max}(S) \;\le\; \Sigma_{\max} + C\,d(\mathbf{x},\mathbf{u})^2\,, \tag{67}$$

$$\sigma_{\min}(S) \;\ge\; \Sigma_{\min} - C\,d(\mathbf{x},\mathbf{u})^2\,. \tag{68}$$

We can now bound the various terms on Eq. (54).

Lower bound on A. Using Theorem 4.1 in [CR08] we obtain, with high probability for any $\xi > 0$:

$$A \;\ge\; \frac{\epsilon}{n}\langle(US\overline{Z}^T + \overline{W}SV^T), (US\widehat{Z}^T + \widehat{W}SV^T)\rangle \tag{69}$$

$$-\frac{\xi\,\epsilon}{n}||US\overline{Z}^T + \overline{W}SV^T||_F||US\widehat{Z}^T + \widehat{W}SV^T||_F \ge (1-\xi)A_0 - (1+\xi)B_0 \tag{70}$$

where

$$A_0 \;=\; \frac{\epsilon}{n}||US\overline{Z}^T + \overline{W}SV^T||_F^2 \tag{71}$$

$$B_0 \;=\; \frac{\epsilon}{n}||US\overline{Z}^T + \overline{W}SV^T||_F||US(\overline{Z}-\widehat{Z})^T + (\overline{W}-\widehat{W})SV^T||_F\,. \tag{72}$$

The term A_0 is lower bounded analogously to $\mathbb{E}\{A^2\}$ in the proof of Lemma 5.3 (see Eq. (44) and below). Using the eigenvalue bounds Eq. (67) and (68), we obtain $A_0 \ge Cn\epsilon d(\mathbf{x},\mathbf{u})^2$. As for the second term we notice that

$$B_0 \;\le\; 2\epsilon\mathrm{Tr}(SS^T(\overline{Z}-\widehat{Z})(\overline{Z}-\widehat{Z})^T) + 2\epsilon\mathrm{Tr}(S^TS(\overline{W}-\widehat{W})(\overline{W}-\widehat{W})^T) \tag{73}$$

$$\le\; 2\epsilon\sigma_{\max}(S)(||\overline{Z}-\widehat{Z}||_F^2 + ||\overline{W}-\widehat{W}||_F^2) \le Cn\epsilon d(\mathbf{x},\mathbf{u})^4\,. \tag{74}$$

Therefore for $\delta \ge d(\mathbf{x},\mathbf{u})$ small enough $A_0 > 2B_0$, whence $A_0 \ge Cn\epsilon d(\mathbf{x},\mathbf{u})^2/4$.

Upper bound on B_1. We begin by using Cauchy-Schwarz inequality:

$$B_1 \le ||\mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T)||_F||\mathcal{P}_E(\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)||_F\,, \tag{75}$$

Proceeding as above we obtain (with high probability)

$$||\mathcal{P}_E(US\overline{Z}^T + \overline{W}SV^T)||_F^2 \;\le\; (1+\xi)\frac{\epsilon}{n}||US\overline{Z}^T + \overline{W}SV^T||_F^2 \tag{76}$$

$$\le\; (1+\xi)\frac{2\epsilon}{n}(||US\overline{Z}^T||_F^2 + ||\overline{W}SV^T||_F^2) \tag{77}$$

$$\le\; (1+\xi)2\epsilon\Big(\mathrm{Tr}(S^TS\overline{Z}^T\overline{Z}) + \mathrm{Tr}(SS^T\overline{W}^T\overline{W})\Big) \tag{78}$$

$$\le\; Cn\epsilon d(\mathbf{x},\mathbf{u})^2\,. \tag{79}$$

In order to estimate the second factor in Eq. (75), we first notice that $||\mathcal{P}_E(\overline{W}S\widehat{Z}^T + \widehat{W}S\overline{Z}^T)||_F^2 \leq 2||\mathcal{P}_E(\overline{W}S\widehat{Z}^T)||_F^2 + 2||\mathcal{P}_E(\widehat{W}S\overline{Z}^T)||_F^2$ and then bound each of the two terms in the same way. Consider, to be definite, the first one:

$$||\mathcal{P}_E(\overline{W}S\widehat{Z}^T)||_F^2 \;=\; \sum_{(ij)\in E}(\overline{W}S\widehat{Z}^T)_{ij}^2 \leq \sigma_{\max}(S)^2 \sum_{a,b=1}^{r}\sum_{(ij)\in E}\overline{W}_{ia}^2\widehat{Z}_{jb}^2\,.$$

At this point we can apply Lemma the same argument as after Eq. (47). Bounding $\sigma_{\max}(S)^2$ and using the fact that both $\overline{W}, \widehat{Z} \in \mathcal{K}(C\mu_0)$ we get

$$||\mathcal{P}_E(\overline{W}S\widehat{Z}^T)||_F^2 = Cn\epsilon\max(r/\sqrt{\epsilon}, \delta^2)d(\mathbf{x},\mathbf{u})^2\,.$$

Putting together Eqs. (79) and (80) we finally get

$$B_1 = Cn\epsilon\max(r^{1/2}/\epsilon^{1/4}, \delta)d(\mathbf{x},\mathbf{u})^2\,,$$

which is smaller than $A/100$ for δ, $r/\sqrt{\epsilon}$ small enough.

Upper bound on B_2. We have

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

The upper bound on B_2' is obtained similarly to the the one on B_1. Indeed proceeding as above we obtain

$$||\mathcal{P}_E(US\widehat{Z}^T + \widehat{W}SV^T)||_F^2 \;\leq\; Cn\epsilon d(\mathbf{x},\mathbf{u})^2\,, \tag{80}$$

$$||\overline{W}S\overline{Z}^T||_F^2 \;\leq\; Cn\epsilon\max(r/\sqrt{\epsilon}, \delta)d(\mathbf{x},\mathbf{u})^2\,, \tag{81}$$

whence $B_2' \leq Cn\epsilon d(\mathbf{x},\mathbf{u})^2/100$ for δ, $r/\sqrt{\epsilon}$ small enough.

Consider now B_2''. By Theorem 4.1 in [CR08], we have

$$B_2'' \;\leq\; \frac{\epsilon}{n}|\langle US\widehat{Z}^T, U(S-\Sigma)V^T\rangle| + \frac{\xi\epsilon}{n}||US\widehat{Z}^T||_F||U(S-\Sigma)V^T||_F \tag{82}$$

$$\leq\; \sigma_{\max}(S)\epsilon||S-\Sigma||_F||\widehat{Z}^TV||_F + \xi\epsilon||US\widehat{Z}^T||_F||S-\Sigma||_F\,. \tag{83}$$

The first term is bounded using (the analogous of) Eq. (66) and $||S-\Sigma||_F \leq d(\mathbf{x},\mathbf{u})$. For the second term we use $||\widehat{Z}||_F^2 \leq Cnd(\mathbf{x},\mathbf{u})^2$, thus getting

$$B_2'' \;\leq\; Cn\epsilon d(\mathbf{x},\mathbf{u})^3 + Cn\epsilon\xi d(\mathbf{x},\mathbf{u})^2 \leq Cn\epsilon(\delta+\xi)d(\mathbf{x},\mathbf{u})^2 \tag{84}$$

whence $B_2'' \leq Cn\epsilon d(\mathbf{x},\mathbf{u})^2/100$ for δ small enough. The same argument applies to B_2''' thus proving the desired bound.

Upper bound on B_3. Finally for the last term it is sufficient to use a crude bound

$$B_3 \leq 4\Big(||\mathcal{P}_E(\overline{W}S\widehat{Z}^T)||_F + ||\mathcal{P}_E(\widehat{W}S\overline{Z}^T)||_F\Big)\Big(||\mathcal{P}_E(U(S-\Sigma)V^T)||_F + ||\mathcal{P}_E(\overline{W}S\overline{Z}^T)||_F\Big)\,, \tag{85}$$

and all of the factors have been estimated above.

7.2 Lower bound on $\operatorname{grad} G(\mathbf{x})$

By the definition of G in Eq. (34), we have

$$\langle \operatorname{grad} G(\mathbf{x}), \widehat{\mathbf{w}} \rangle = \frac{1}{\mu_0 r} \sum_{i=1}^{m} G_1' \left(\frac{||X^{(i)}||^2}{2\mu_0 r} \right) \langle X^{(i)}, \widehat{W}^{(i)} \rangle + \frac{1}{\mu_0 r} \sum_{j=1}^{n} G_1' \left(\frac{||Y^{(i)}||^2}{2\mu_0 r} \right) \langle Y^{(i)}, \widehat{Z}^{(i)} \rangle \,. \quad (86)$$

It is therefore sufficient to show that if $||X^{(i)}||^2 > 2\mu_0 r$, then $\langle X^{(i)}, \widehat{W}^{(i)} \rangle > 0$, and if $||Y^{(j)}||^2 > 2\mu_0 r$, 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 (52) we get

$$X^{(i)} = R \left\{ \cos\Theta \, R^T U^{(i)} + \sin\Theta \, L^{(i)} \right\} , \quad (87)$$

$$\widehat{W}^{(i)} = R \left\{ \Theta \cos\Theta L^{(i)} - \Theta \sin\Theta R^T U^{(i)} \right\} . \quad (88)$$

From the first expression it follows that

$$||\sin\Theta \, L^{(i)}||^2 \leq ||X^{(i)}||^2 + ||\cos\Theta \, R^T U^{(i)}||^2 \leq 3 \, \mu_0 r \,. \quad (89)$$

On the other hand, by taking the difference of Eqs. (87) and (88) we have

$$||X^{(i)} - \widehat{W}^{(i)}|| \leq ||(\sin\Theta - \Theta\cos\Theta) L^{(i)}|| + ||(\cos\Theta + \Theta\sin\Theta) R^T U^{(i)}|| \quad (90)$$

$$\leq \max_i(\theta_i^2) ||\sin\Theta L^{(i)}|| + ||U^{(i)}|| \leq \delta\sqrt{3\mu_0 r} + \sqrt{\mu_0 r} \,. \quad (91)$$

where we used the inequality $(\sin\omega - \omega\cos\omega) \leq \omega^2 \sin\omega$ valid for $\omega \in [0, \pi/2]$. For δ small enough we have therefore $||X^{(i)} - \widehat{W}^{(i)}|| \leq (9/10)\sqrt{2\mu_0 r}$. To conclude, for $||X^{(i)}|| \geq 2\mu_0 r$

$$\langle X^{(i)}, \widehat{W}^{(i)} \rangle \geq ||X^{(i)}||^2 - ||X^{(i)}|| \, ||X^{(i)} - \widehat{W}^{(i)}|| \geq ||X^{(i)}|| (\sqrt{2\mu_0 r} - (9/10)\sqrt{2\mu_0 r}) \geq 0 \,. \quad (92)$$

Acknowledgements

We thank Emmanuel Candés and Benjamin Recht for stimulating discussions on the subject of this paper. This work was partially supported by a Terman fellowship and an NSF CAREER award (CCF-0743978).

A Proof of Remark 4.3

The proof consists in showing that $|\overline{\mathcal{A}}_l| > \max\{e^{-C_1\epsilon} m, C_2\alpha\}$ with probability less than $1/2n^3$, using Chernoff bound. In the case of large ϵ, when $\epsilon > 2\sqrt{\alpha}\log(n)$, we have $\mathbb{P}\left\{|\overline{\mathcal{A}}_l| > C_2\alpha\right\} \leq 1/2n^3$, for $C_2 > 4/\alpha$. In the case of small ϵ, when $\epsilon \leq 2\sqrt{\alpha}\log(n)$, $\mathbb{P}\left\{|\overline{\mathcal{A}}_l| > e^{-C_1\epsilon} m\right\} \leq 1/2n^3$, for $C_1 < 1/4\sqrt{\alpha}$, which proves the thesis.

Analogously, we can prove that $\mathbb{P}\left\{|\overline{\mathcal{A}}_r| > \max\{e^{-C_1\epsilon} n, C_2\}\right\} \leq 1/2n^3$, which finishes the proof of Remark 4.3.

B Proof of Remark 4.5

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 $\sum_{(i,j)\in L} x_i M^A_{ij} y_j - x^T M y$, we write,

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

The first term can be bounded by noting that $|(M^A - M)_{ij}|$ is positive only if $i \notin A_l$ or $j \notin A_r$ in which case $|(M^A - M)_{ij}| \le \mu_1 \sqrt{r}$ by the incoherence condition A2. Also, by Remark 4.3, there exists $\delta \le \max\{e^{-C_1\epsilon}, C_2/n\}$ 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

$$\begin{aligned}
\left| x^T \Big(M^A - M \Big) y \right| &\le \sum_{ij} |x_i| |y_j| \Big(\mathbb{I}(i \notin A_l) + \mathbb{I}(j \notin A_r) \Big) \mu_1 \sqrt{r} \\
&= \left(\sum_i |x_i| \mathbb{I}(i \notin A_l) \sum_j |y_j| + \sum_j |y_j| \mathbb{I}(j \notin A_r) \sum_i |x_i| \right) \mu_1 \sqrt{r} \\
&\le \left(\sqrt{\delta m}\sqrt{n} + \sqrt{\delta n}\sqrt{m} \right) \mu_1 \sqrt{r} \\
&\le \frac{\mu_1 \sqrt{mnr}}{\sqrt{\epsilon}} \,.
\end{aligned}$$

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

$$\begin{aligned}
\left| \sum_{(i,j)\in \overline{L}} x_i M^A_{ij} y_j \right| &\le \sum_{(i,j)\in \overline{L}} \frac{\left| x_i M^A_{ij} y_j \right|^2}{\left| x_i M^A_{ij} y_j \right|} \\
&\le \sqrt{\frac{mn}{r\epsilon}} \sum_{(i,j)\in \overline{L}} \left| x_i M^A_{ij} y_j \right|^2 \\
&\le \sqrt{\frac{mn}{r\epsilon}} \sum_{i\in[m], j\in[n]} \left| x_i M^A_{ij} y_j \right|^2 \\
&\le \frac{\mu_1^2 \sqrt{mnr}}{\sqrt{\epsilon}} \,,
\end{aligned}$$

where the second inequality follows from the definition of heavy couples and the last inequality is due to incoherence condition A2.

Hence summing two contributions, we get

$$|\mathbb{E}[Z]| \leq \left(\mu_1 + \mu_1^2\right)\sqrt{r\epsilon}\,.$$

C Proof of Remark 4.6

We can associate to the matrix Q a bipartite graph $\mathcal{G} = ([m], [n], \mathcal{E})$. The proof is similar to the one in [FKS89, FO05] and is based on two properties of the graph cG:

1. The graph $\mathcal{G}$ has maximum degree *bounded* by a constant times the average degree:

$$\deg(i) \leq \frac{2\epsilon}{\sqrt{\alpha}}\,, \tag{93}$$

$$\deg(j) \leq 2\epsilon\sqrt{\alpha}\,, \tag{94}$$

   for all $i \in [m]$ and $j \in [n]$.

2. *Discrepancy.* We will 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:

   1. $$\frac{e(A,B)}{\mu(A,B)} \leq \xi_1\,, \tag{95}$$

   2. $$e(A,B)\ln\left(\frac{e(A,B)}{\mu(A,B)}\right) \leq \xi_2 \max\{|A|, \alpha|B|\}\ln\left(\frac{\sqrt{mn}}{\max\{|A|,\alpha|B|\}}\right)\,. \tag{96}$$

   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.

We will prove that the discrepancy property holds with high probability later in this section, see Lemma C.1.

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

$$A_i = \{u \in [m] \,:\, \frac{\Delta}{\sqrt{m}}2^{i-1} \leq |x_u| < \frac{\Delta}{\sqrt{m}}2^i\}\,,$$

$$B_j = \{v \in [n] \,:\, \frac{\Delta}{\sqrt{n}}2^{j-1} \leq |y_v| < \frac{\Delta}{\sqrt{n}}2^j\}\,,$$

for $i \in \{1, 2, \ldots, \lceil \log(\sqrt{m}/\Delta)/\log 2\rceil\}$, and $j \in \{1, 2, \ldots, \lceil \log(\sqrt{n}/\Delta)/\log 2\rceil\}$, 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 non-zero element of $x \in T_m$ in absolute value is $\Delta/\sqrt{m}$. 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

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

Note that, by definition, we have

$$\sum_i \alpha_i \leq 4||x||^2/\Delta^2 \,, \tag{97}$$

$$\sum_i \beta_i \leq 4||y||^2/\Delta^2 \,. \tag{98}$$

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,

$$\mathcal{C}_1 \equiv \left\{(i,j) \,:\, 2^{i+j} \geq \frac{4C\sqrt{\alpha\epsilon}}{\Delta^2} \text{ and } (A_i, B_j) \text{ satisfies (95)}\right\}, \tag{99}$$

$$\mathcal{C}_2 \equiv \left\{(i,j) \,:\, 2^{i+j} \geq \frac{4C\sqrt{\alpha\epsilon}}{\Delta^2} \text{ and } (A_i, B_j) \text{ satisfies (96)}\right\} \setminus \mathcal{C}_1 \,. \tag{100}$$

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 $2^{i+j} \geq \frac{4C\sqrt{\alpha\epsilon}}{\Delta^2}$, it follows that $\sigma_{i,j} \leq \xi_1\Delta^2/4C\sqrt{\alpha}$. By Eqs. (97) and (98), we have $\sum_{\mathcal{C}_1} \alpha_i\beta_j\sigma_{i,j} \leq (\xi_1\Delta^2/4C\sqrt{\alpha})(2/\Delta)^4 = O(1)$.

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

$$e_{i,j} \log\left(\frac{e_{i,j}}{\mu_{i,j}}\right) \leq \xi_2 \max\{a_i, \alpha b_j\} \log\left(\frac{\sqrt{mn}}{\max\{a_i, \alpha b_j\}}\right) \,. \tag{101}$$

We start by changing variables on both sides of Eq. (101).

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

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

$$\sigma_{i,j}\alpha_i \log\left(\frac{e_{i,j}}{\mu_{i,j}}\right) \leq \frac{\xi_2 2^i}{\sqrt{\epsilon}2^j} \left[\log(2^{2j}) - \log(\beta_j\sqrt{\alpha})\right] \,. \tag{102}$$

To achieve the desired bound, we partition the analysis into 5 cases:

1. $\sigma_{i,j} \leq 1$: By Eqs. (97) and (98), we have $\sum \alpha_i \beta_j \sigma_{i,j} \leq (2/\Delta)^4 = O(1)$.

2. $2^i > \sqrt{\epsilon} 2^j$: By the bounded degree property in Eq. (94), we have $e_{i,j} \leq a_i 2\epsilon/\sqrt{\alpha}$, which implies that $e_{i,j}/\mu_{i,j} \leq 2n/b_j$. For a fixed i we have, $\sum_j \beta_j \sigma_{i,j} \mathbb{I}(2^i > \sqrt{\epsilon} 2^j) \leq 2\sqrt{\epsilon} \sum_j 2^{j-i} \mathbb{I}(2^i > \sqrt{\epsilon} 2^j) \leq 4$. Then, $\sum \alpha_i \beta_j \sigma_{i,j} \leq 16/\Delta^2 = O(1)$.

3. $\log(e_{i,j}/\mu_{i,j}) > \frac{1}{4}\left[\log(2^{2j}) - \log(\beta_j \sqrt{\alpha})\right]$: From Eq.(102), it immediately follows that $\sigma_{i,j}\alpha_i \leq \frac{4\xi_2 2^i}{\sqrt{\epsilon} 2^j}$. Because of case 2, we can assume $2^i \leq \sqrt{\epsilon} 2^j$, which implies that for a fixed j we have the following inequality : $\sum_i \sigma_{i,j}\alpha_i \leq 4\xi_2 \sum_i \frac{2^i}{\sqrt{\epsilon} 2^j} \mathbb{I}(2^i \leq \sqrt{\epsilon} 2^j) \leq 8\xi_2$. Then it follows by Eq. (98) that $\sum \alpha_i \beta_j \sigma_{i,j} \leq 32\xi_2/\Delta^2 = O(1)$.

4. $\log(2^{2j}) \geq -\log(\beta_j \sqrt{\alpha})$: Because of case 3, we can assume $\log(e_{i,j}/\mu_{i,j}) \leq \frac{1}{4}\left[\log(2^{2j}) - \log(\beta_j \sqrt{\alpha})\right]$, which implies that $\log(e_{i,j}/\mu_{i,j}) \leq \log(2^j)$. Further, because of case 1, we assume $1 < \sigma_{i,j} = e_{i,j}\sqrt{\epsilon}/\mu_{i,j} 2^{i+j}$. Combining those two inequalities, we get $2^i \leq \sqrt{\epsilon}$.

   Since in defining $\mathcal{C}_2$ we excluded $\mathcal{C}_1$, if $(i,j) \in \mathcal{C}_2$ then $\log(e_{i,j}/\mu_{i,j}) \geq 1$. Applying Eq. (102) we get $\sigma_{i,j}\alpha_i \leq \sigma_{i,j}\alpha_i \log(e_{i,j}/\mu_{i,j}) \leq (\xi_2 2^{i-j}/\sqrt{\epsilon})\left[\log(2^{2j}) - \log(\beta_j \sqrt{\alpha})\right] \leq 4\xi_2 2^i/\sqrt{\epsilon}$.

   Combining above two results, it follows that $\sum_i \sigma_{i,j}\alpha_i \leq 4\xi_2 \sum_i \frac{2^i}{\sqrt{\epsilon}} \mathbb{I}(2^i \leq \sqrt{\epsilon}) \leq 8\xi_2$. Then, we have the desired bound : $\sum \alpha_i \beta_j \sigma_{i,j} \leq \frac{32\xi_2}{\Delta^2} = O(1)$.

5. $\log(2^{2j}) < -\log(\beta_j \sqrt{\alpha})$: Because of case 4, we assume $\log(2^{2j}) \leq -\log(\beta_j \sqrt{\alpha})$. Then it follows, since we're not in case 3, that $\log(e_{i,j}/\mu_{i,j}) \leq \frac{1}{4}\left[\log(2^{2j}) - \log(\beta_j \sqrt{\alpha})\right] \leq -\log(\beta_j \sqrt{\alpha})$. Hence, $e_{i,j}/\mu_{i,j} \leq 1/\beta_j \sqrt{\alpha}$. This implies that $\sigma_{i,j} = e_{i,j}\sqrt{\epsilon}/\mu_{i,j} 2^{i+j} \leq \sqrt{\epsilon}/\beta_j \sqrt{\alpha} 2^{i+j}$. Since the summation is over pairs of indices (i,j) such that $2^{i+j} \geq 4C\sqrt{\alpha\epsilon}/\Delta^2$, we have $\sum_j \sigma_{i,j}\beta_j \leq \frac{\Delta^2}{2\alpha C}$. Then it follows that $\sum \alpha_i \beta_j \sigma_{i,j} \leq \frac{2}{\alpha C} = O(1)$.

Summing up the results, we get that there exists a constant $C' \leq \frac{16}{\Delta^4} + \frac{4\xi_1}{C\Delta^2\sqrt{\alpha}} + \frac{16}{\Delta^2} + \frac{32\xi_2}{\Delta^2} + \frac{2}{\alpha C}$, such that

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

This finishes the proof of Remark 4.6.

Lemma C.1. *The adjacency matrix Q has discrepancy property with probability at least $1 - 1/n$.*

Proof. The proof is a generalization of analogous result in [FKS89, FO05] 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 non-bipartite graph in Section 2.2.5 of [FO05], and analogous analysis for bipartite graph shows that for all subsets $A \subseteq [m]$ and $B \subseteq [n]$, with probability at least $1 - 1/n$, the discrepancy condition holds with $\xi_1 = \max\{4, 2e\}$ and $\xi_2 = \max\{12 + \frac{15}{\alpha}, 15 + \frac{12}{\alpha}\}$.

□

D Proof of remarks 5.1 and 5.2

Proof. (Remark 5.1.) Let $\theta = (\theta_1, \dots, \theta_p)$, $\theta_i \in [-\pi/2, \pi/2]$ be the principal angles between the planes spanned by the columns of X_1 and X_2. It is known that $d_{\rm c}(X_1, X_2) = ||2\sin(\theta/2)||_2$ and $d_{\rm p}(X_1, X_2) = ||\sin\theta||_2$. The thesis follows from the elementary inequalities

$$\frac{1}{\pi}\alpha \le \sqrt{2}\,\sin(\alpha/2) \le \sin\alpha \le 2\sin(\alpha/2) \tag{103}$$

valid for $\alpha \in [0, \pi/2]$. □

Proof. (Remark 5.2.) We start by observing that

$$d_{\rm p}(V, Y) = \frac{1}{\sqrt{n}} \min_{A\in\mathbb{R}^{r\times r}} ||V - YA||_F\,. \tag{104}$$

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

$$\begin{aligned}
||V - YA||_F &= \sup_{B,||B||_F\le 1} \langle B, (V - YA)\rangle && (105)\\
&= \sup_{B,||B||_F\le 1} \langle B^T, \frac{1}{n}\Sigma^{-1}U^T(U\Sigma V^T - XSY^T)\rangle && (106)\\
&= \frac{1}{n}\sup_{B,||B||_F\le 1} \langle U\Sigma^{-1}B^T, (M - \widehat{M})\rangle && (107)\\
&\le \frac{1}{n}\sup_{B,||B||_F\le 1} ||U\Sigma^{-1}B^T||_F\,||M - \widehat{M}||_F\,. && (108)
\end{aligned}$$

On the other hand

$$||U\Sigma^{-1}B^T||_F^2 = {\rm Tr}(B\Sigma^{-1}U^TU\Sigma^{-1}B^T) = n{\rm Tr}(B^TB\Sigma^{-2}) \le n\Sigma_{\rm min}^{-2}||B||_F^2\,,$$

whereby the last inequality follows from the fact that Σ is diagonal. Together (104) and (108), this implies the thesis. □

References

[AFK+01] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia, *Spectral analysis of data*, Proceedings of the thirty-third annual ACM symposium on Theory of computing (New York, NY, USA), ACM, 2001, pp. 619–626.

[AM07] D. Achlioptas and F. McSherry, *Fast computation of low-rank matrix approximations*, J. ACM **54** (2007), no. 2, 9.

[Arm66] L. Armijo, *Minimization of functions having lipschitz continuous first partial derivatives*, Pacific J. Math. **16** (1966), no. 1, 1–3.

[BDJ99] M. W. Berry, Z. Drmać, and E. R. Jessup, *Matrices, vector spaces, and information retrieval*, SIAM Review **41** (1999), no. 2, 335–362.

[Ber92] M. W. Berry, *Large scale sparse singular value computations*, International Journal of Supercomputer Applications **6** (1992), 13–49.

[CCS08] J-F Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, `arXiv:0810.3286`, 2008.

[CR08] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, `arxiv:0805.4471`, 2008.

[CRT06] E. J. Candes, J. K. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. on Inform. Theory **52** (2006), 489– 509.

[CT09] E. J. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, `arXiv:0903.1476`, 2009.

[Don06] D. L. Donoho, *Compressed Sensing*, IEEE Trans. on Inform. Theory **52** (2006), 1289–1306.

[EAS99] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matr. Anal. Appl. **20** (1999), 303–353.

[Faz02] M. Fazel, *Matrix rank minimization with applications*, Ph.D. thesis, Stanford University, 2002.

[FKS89] J. Friedman, J. Kahn, and E. Szemerédi, *On the second eigenvalue in random regular graphs*, Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing (Seattle, Washington, USA), ACM, may 1989, pp. 587–598.

[FKV04] A. Frieze, R. Kannan, and S. Vempala, *Fast monte-carlo algorithms for finding low-rank approximations*, J. ACM **51** (2004), no. 6, 1025–1041.

[FO05] U. Feige and E. Ofek, *Spectral techniques applied to sparse random graphs*, Random Struct. Algorithms **27** (2005), no. 2, 251–275.

[KMO08] R. H. Keshavan, A. Montanari, and S. Oh, *Learning low rank matrices from $O(n)$ entries*, Proc. of the Allerton Conf. on Commun., Control and Computing, September 2008.

[KOM09] R. H. Keshavan, S. Oh, and A. Montanari, *Matrix completion from a few entries*, arXiv:0901.3150, January 2009.

[Net] *Netflix prize*.

[RFP07] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization*, `arxiv:0706.4138`, 2007.

[SC09] A. Singer and M. Cucuringu, *Uniqueness of low-rank matrix completion by rigidity theory*, `arXiv:0902.3846`, January 2009.