Matrix Completion from a Few Entries

Raghunandan H. Keshavan and Sewoong Oh
EE Department
Stanford University, Stanford, CA 94304

Andrea Montanari
EE and Statistics Departments
Stanford University, Stanford, CA 94304

Abstract—Let \( M \) be an \( n \times n \) 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 \( |E| = O(rn) \) observed entries with relative root mean square error \( \text{RMSE} \leq C(\alpha) \left( nr/|E| \right)^{1/2} \). Further, if \( r = O(1) \) and \( M \) is sufficiently unstructured, then it can be reconstructed exactly from \( |E| = O(n \log n) \) entries.

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-Szemeredi and Feige-Ofek on the spectrum of sparse random matrices.

I. INTRODUCTION

Imagine that each one 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 missing pairs in such a way to provide targeted suggestions. As an example, in 2006, Netflix made public such a dataset with \( m \approx 5 \times 10^5 \), \( n \approx 2 \times 10^4 \) and \( |E| \approx 10^8 \) and challenged the research community to predict the missing ratings with root mean square error below 0.8563 [1].

The general question we address here is: under which conditions do the known ratings provide sufficient information to efficiently infer the unknown ones?

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 \( n \times r \), and a diagonal matrix \( \Sigma \), of dimensions \( r \times r \) such that

\[
M = U \Sigma V^T.
\]

For justification of these assumptions and background on the use of low rank matrices in information retrieval, we refer to [2]. Motivated by the massive size of actual datasets, we shall focus on the limit of large \( m, n \) with \( m/n = \alpha \) of order 1.

We further assume that the factors \( U, V \) are unstructured. This notion is formalized by the incoherence condition [3] as defined in Section II. In particular the incoherence condition is satisfied with high probability if \( M = U^V V^T \) with \( U' \) and \( V' \) uniformly random orthogonal matrices.

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

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

The set \( E \) will be uniformly random given its size \( |E| \).

B. Algorithm and guarantees

A naive algorithm consists of the following operation.

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

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

And return the matrix \( T_r(M^E) = (mn/|E|) \sum_{i=1}^r \sigma_i x_i y_i^T \) obtained by setting to 0 all but the \( r \) largest singular values. Notice that, apart from the rescaling factor \( (mn/|E|) \), \( T_r(M^E) \) 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 \( M \).

This algorithm fails if \( |E| = \Theta(n) \). The reason is that, in this regime, the matrix \( M^E \) contains columns and rows with \( \Omega(\log n/\log \log n) \) non-zero (revealed) entries. The largest singular values of \( M^E \) are an artifact of these 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 than \( 2|E|/n \). Set to 0 all rows with degree larger than \( 2|E|/m \).

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

Spectral Matrix Completion (matrix \( M^E \))

1: Trim \( M^E \), and let \( \tilde{M}^E \) be the output;
2: Project \( \tilde{M}^E \) to \( T_r(\tilde{M}^E) \);
3: Clean residual errors by minimizing \( F(X,Y) \).

\footnotetext{1}{Throughout this paper we denote by \([N] = \{1, 2, \ldots, N\}\) the set of first \( N \) integers.}
Various implementations are possible, but we found the following one particularly appealing. Given $X \in \mathbb{R}^{n \times r}$, $Y \in \mathbb{R}^{n \times r}$ with $X^T X = n \mathbf{1}$ and $Y^T Y = n \mathbf{1}$, we define

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

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

The cleaning step consists in writing $T_r(\hat{M}^E) = X_0 S_0 Y_0^T$ 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 the journal version of this paper for background and references). Optimization over Grassmann manifolds is a well understood topic [4] 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 $F(X,Y)$.

Our main result establishes that this simple procedure achieves arbitrarily small root mean square error $||M - T_r(\hat{M}^E)||_F/\sqrt{mnr}$ with $O(nr)$ revealed entries.

**Theorem I.1.** Assume $M$ to be a rank $r \leq n^{1/2}$ matrix with $|M_{ij}| \leq M_{\text{max}}$ for all $i,j$. Then with high probability

$$\frac{1}{\min n M_{\text{max}}} ||M - T_r(\hat{M}^E)||_F^2 \leq C(\alpha) \frac{nr}{|E|}. \quad (6)$$

The proof is provided in Section IV (the proofs of several technical remarks can be found in the journal version [5]).

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

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

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

The proof will appear in the journal version of this paper [5]. The basic intuition is that, for $|E| \geq C'(\alpha)nr \max\{\log n, r\}$, $T_r(\hat{M}^E)$ is so close to $M$ that the cost function is well approximated by a quadratic function.

**Theorem I.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 I.2 is due to a coupon-collector effect [3], [6], [5]: 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 rank $r$ bounded, Theorem I.2 is optimal. It is suboptimal by a polylogarithmic factor for $r = O(\log n)$.

**C. Related work**

Beyond collaborative filtering, low rank models are used for clustering, information retrieval, machine learning, and image processing. In [7], 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 [8], [9]. The connection with compressed sensing was emphasized in [10], 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 [3] proved that, if $E$ is random, the convex relaxation correctly reconstructs $M$ as long as $|E| \geq C r n^{b/5} \log n$. On the other hand, from a purely information theoretic point of view (i.e. disregarding algorithmic considerations), it is clear that $|E| = O(nr)$ observations should allow to reconstruct $M$ with arbitrary precision. Indeed this point was raised in [3] and proved in [6], through a counting argument.

The present paper fills this gap. We describe an efficient algorithm that reconstructs a rank-$r$ matrix from $O(nr)$ random observations. The most complex component of our algorithm is the SVD in step 2. Generic routines accomplish this task with $O(n^3)$ operations. Thanks to the sparsity of $\hat{M}^E$, this step can be implemented using the Lanczos procedure with $O(|E|r \log n)$ complexity. We were able to treat realistic data sets with $n \approx 10^5$. This must be compared with the $O(n^4)$ complexity of [3] (but see [11] for an iterative implementation of the latter).

After this paper was submitted to ISIT, Candès and Tao [12] proved a guarantee for the convex relaxation algorithm, that is comparable with Theorem I.2. A longer version of the present paper was submitted to IEEE Transactions on Information Theory [5].

**II. 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 $||u_i|| = \sqrt{m_i}$, $||v_i|| = \sqrt{n}$ 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 $\Sigma$). We shall further write $\Sigma = \text{diag}(\Sigma_1, \ldots, \Sigma_r)$ with $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_r \geq 0$.

The matrices $U$, $V$ and $\Sigma$ will be said to be $(\mu_0, \mu_1)$-incoherent if they satisfy the following properties:
A1. For all \( i \in [m], \) \( j \in [n] \), we have \( \sum_{k=1}^{r} U_{i,k}^2 \leq \mu_0 r \), \( \sum_{k=1}^{r} V_{j,k}^2 \leq \mu_0 r \).

A2. For all \( i \in [m], j \in [n] \), we have \( |\sum_{k=1}^{r} U_{i,k} \Sigma_k V_{j,k}| \leq \mu_1 r^{1/2} \).

The first one coincides with one of the incoherence assumptions in [3]. The second one is easier to verify than the analogous one in [3], in that it concerns the matrix elements themselves.

Notice that assumption A2 implies the bounded entry condition in Theorem I.1 with \( M_{\max} = \mu_1 r^{1/2} \). In the following, whenever we write that a property \( A \) holds with high probability (w.h.p.), we mean that there exists a function \( f(n) = f(n; \alpha) \) such that \( \Pr(A) \geq 1 - f(n) \) and \( f(n) \to 0 \).

Define a constant \( \epsilon \equiv |E|/\sqrt{mn} \). Then it is convenient to work with a model in which each entry is revealed independently with probability \( \epsilon/\sqrt{mn} \). Since, w.h.p., \( |E| = \epsilon/\sqrt{n} + A\sqrt{n \log n} \), it will be sufficient to prove that our algorithm is successful for \( \epsilon \geq Cr \). Finally, we will use \( C, C' \) etc. to denote generic constants that depend uniquely on \( \alpha, \Sigma_{\min}, \Sigma_{\max}, \mu_0, \mu_1 \).

Given a vector \( x \in \mathbb{R}^n, \) \( ||x|| \) will denote its Euclidean norm. For a matrix \( X \in \mathbb{R}^{m \times n'}, \) \( ||X|| \) is its Frobenius norm, and \( ||X||_2 \) its operator norm.

### III. ALGORITHM IMPLEMENTATION AND SIMULATIONS

A MATLAB implementation of our algorithm is available from http://www.stanford.edu/~raghuram. In Fig. 1, we plot the probability that \( \text{SPECTRAL} \) matrix completion exactly reconstructs \( M \) as a function of the number of revealed entries \( |E| \). The algorithm is evaluated on random matrices of rank \( r = 4 \). As predicted by Theorem I.2, the success probability presents a sharp threshold for \( |E| = C n \log n \).

The location of the threshold is surprisingly close to the lower bound proved in [13], below which the problem admits more than one solution.

In Fig. 2 we apply our algorithm to ‘approximately’ low-rank matrices as defined in [14]. The resulting root mean square error is smaller by roughly 50% with respect to the one obtained with the convex relaxation of [3], [14].

### IV. PROOF OF THEOREM I.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

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

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 \( M \).

**Lemma IV.1.** There exists a constant \( C > 0 \) such that, with high probability

\[
\frac{\sigma_q}{\epsilon} - \Sigma_q \leq C M_{\max} \sqrt{\epsilon},
\]

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

This result generalizes a celebrated bound on the second eigenvalue of random graphs [15], [16] and is illustrated in Fig. 3: the spectrum of \( \widetilde{M}^E \) clearly reveals the rank-4 structure of \( M \).

As shown in Section VI, Lemma IV.1 is a direct consequence of the following estimate.

**Lemma IV.2.** There exists a constant \( C > 0 \) such that, with high probability

\[
\frac{\epsilon}{\sqrt{mn}} \frac{1}{2} \leq C M_{\max} \sqrt{\epsilon}.
\]

The proof of this lemma is given in Section V.

We will now prove Theorem I.1.
Apply union bound to these sets, with a large deviation (1)

Reduce to

for any rank 4 matrix $T$ of rank at most $A$ Lemma IV.1 for the last inequality. Now, for any matrix $x$ probability of $S$ of rank 4 matrix $T$ belonging to discrete sets $\bar{M}_{E}$ with $\epsilon = 50$ and $\Sigma = \text{diag}(1.3, 1.2, 1.1, 1.1)$.

\[ \sqrt{\frac{\epsilon}{\sqrt{mn}}} \left\| M - T_{r}(\bar{M}_{E}) \right\|_{2} \leq \sqrt{\frac{\epsilon}{\sqrt{mn}}} \left\| M - T_{r}(\bar{M}_{E}) \right\|_{2} \leq \sqrt{\frac{\epsilon}{\sqrt{mn}}} \left\| M - T_{r}(\bar{M}_{E}) \right\|_{2} \leq 2C_{\text{max}} \sqrt{\frac{\epsilon}{\sqrt{mn}}} \],

where we used Lemma IV.2 for the second inequality and Lemma IV.1 for the last inequality. Now, for any matrix $A$ of rank at most $2r$, $\| A \|_{F} \leq \sqrt{2r} \| A \|_{2}$, whence

\[ \frac{1}{\sqrt{\sqrt{mn}}} \left\| M - T_{r}(\bar{M}_{E}) \right\|_{F} \leq \frac{\sqrt{\epsilon}}{\sqrt{\sqrt{mn}}} \left\| M - T_{r}(\bar{M}_{E}) \right\|_{2} \leq C' \text{max} \sqrt{\frac{\epsilon}{\sqrt{mn}}}. \]

\section{V. Proof of Lemma IV.2}

We want to show that $|x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y| \leq C M_{\text{max}} \sqrt{\epsilon}$ for any $x \in \mathbb{R}^{m}$, $y \in \mathbb{R}^{n}$ such that $\|x\| = \|y\| = 1$. Our basic strategy (inspired by [15]) will be the following:

1. Reduce to $x$, $y$ belonging to discrete sets $T_{m}$, $T_{n}$;
2. Apply union bound to these sets, with large deviation estimate on the random variable $x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y$.

The technical challenge is that a worst-case bound on the tail probability of $x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y$ is not good enough, and we must keep track of its dependence on $x$ and $y$.

\subsection{A. Discretization}

We define

\[ T_{n} = \left\{ x \in \left\{ \frac{A}{\sqrt{n}} \mathbb{Z} \right\}^{n} : \| x \| \leq 1 \right\}, \]

Notice that $T_{n} \subseteq S_{n} \equiv \{ x \in \mathbb{R}^{n} : \| x \| \leq 1 \}$. The next two remarks are proved in [15], [16], and relate the original problem to the discretized one.

\[ \exists x, y : \sum_{L} x_{i} M_{i,j}^{E,A} y_{j} - \frac{\epsilon}{\sqrt{mn}} x^{T} M y > C_{1} M_{\text{max}} \sqrt{\epsilon}, \]

\[ L = \left\{ (i, j) : \| x_{i} M_{i,j} y_{j} \| \leq M_{\text{max}} (\epsilon / mn)^{1/2} \right\}. \]

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

\[ \left| x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y \right| \leq \left| \sum_{(i,j) \in L} x_{i} M_{i,j} y_{j} - \frac{\epsilon}{\sqrt{mn}} x^{T} M y \right| + \left| \sum_{(i,j) \in \overline{L}} x_{i} M_{i,j} y_{j} \right|. \]

In the next subsection, we will prove that the first contribution is upper bounded by $C_{1} M_{\text{max}} \sqrt{\epsilon}$ for all $x \in T_{m}, y \in T_{n}$. The analogous proof for heavy couples can be found in the journal version [5]. Applying Remark V.1 to $|x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y|$, this proves the thesis.

\subsection{B. Bounding the contribution of light couples}

Let us define the subset of row and column indices which have not been trimmed as $A_{l}$ and $A_{r}$:

\[ A_{l} = \{ i \in [m] : \text{deg}(i) \leq 2 \epsilon / \sqrt{\alpha} \}, \]

\[ A_{r} = \{ j \in [n] : \text{deg}(j) \leq 2 \epsilon \sqrt{\alpha} \}, \]

where $\text{deg}()$ denotes the degree (number of revealed entries) of a row or a column. Notice that $A = (A_{l}, A_{r})$ is a function of the random set $E$. It is easy to get a rough estimate of the sizes of $A_{l}, A_{r}$.

\[ \text{Remark V.2. There exists } C_{1} \text{ and } C_{2} \text{ depending only on } \alpha \text{ such that, with probability larger than } 1 - 1/n^{3}, |A_{l}| \geq m - \text{max} \{ e^{-C_{1} \epsilon m}, C_{2} \alpha \}, |A_{r}| \geq n - \text{max} \{ e^{-C_{1} \epsilon n}, C_{2} \}. \]

For any $E \subseteq [m] \times [n]$ and $A = (A_{l}, A_{r})$ with $A_{l} \subseteq [m], A_{r} \subseteq [n]$, we define $\hat{M}^{E,A}$ 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 $H(E, A) = \{ (x, y) : \sum_{L} x_{i} M_{i,j}^{E,A} y_{j} - \frac{\epsilon}{\sqrt{mn}} x^{T} M y \geq C_{1} M_{\text{max}} \sqrt{\epsilon} \}.$

\[ \text{Remark V.1. Let } R \in \mathbb{R}^{m \times n} \text{ be a matrix. If } |x^{T} R y| \leq B \text{ for all } x \in T_{m} \text{ and } y \in T_{n}, \text{ then } |x^{T} R y| \leq (1 - \Delta)^{-2} B \text{ for all } x' \in S_{m} \text{ and } y' \in S_{n}. \]

Hence it is enough to show that, with high probability, $|x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y| \leq C M_{\text{max}} \sqrt{\epsilon}$ 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 $x^{T}(\overline{M}^{E} - \sqrt{\frac{\epsilon}{\sqrt{mn}}} M)y$. 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 

\[ L = \{ (i, j) : \| x_{i} M_{i,j} y_{j} \| \leq M_{\text{max}} (\epsilon / mn)^{1/2} \}. \]
where it is understood that $x$ and $y$ belong, respectively, to $T_m$ and $T_n$. Note that $\overline{M^E} = M^E,A$, and hence we want to bound $\mathbb{P}\{\mathcal{H}(E,A)\}$. We proceed as follows

$$\mathbb{P}\{\mathcal{H}(E,A)\} = \sum_A \mathbb{P}\{\mathcal{H}(E,A), A = A\} \leq \sum_{|A_t| \geq m(1-\delta), |A_r| \geq n(1-\delta)} \mathbb{P}\{\mathcal{H}(E,A), A = A\} + \frac{1}{n^3} \leq 2^{(n+m)H(\delta)} \max_{|A_t| \geq m(1-\delta), |A_r| \geq n(1-\delta)} \mathbb{P}\{\mathcal{H}(E,A), A = A\} + \frac{1}{n^3}, \quad (12)$$

with $\delta \equiv \max\{e^{-C_1^2}, C_2/n\}$ and $H(x)$ the 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 Eq. (11). The key step consists in proving the following tail estimate

**Lemma V.3.** Let $x \in S_m$, $y \in S_n$, $Z = \sum_{i,j \in E} x_j M_{ij}^E y_j - \frac{\delta}{\sqrt{mn}} x^T M y$, and assume $|A_t| \geq m(1-\delta), |A_r| \geq n(1-\delta)$ with $\delta$ small enough. Then

$$\mathbb{P}\{Z > \lambda M_{\text{max}} \sqrt{\delta}\} \leq \exp\left\{ - \frac{n\lambda^2/2(L-3)}{2} \right\}.$$  

**Proof:** It is shown in [5] that $\mathbb{E}[Z] \leq 2M_{\text{max}} \sqrt{\varepsilon}$. For $A = (A_l,A_r)$, let $\overline{M^A}$ 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

$$a_{ij} = \begin{cases} x_j M_{ij}^A y_j & \text{if } |x_j M_{ij}^A y_j| \leq M_{\text{max}} \langle \varepsilon/\sqrt{mn} \rangle^{1/2}, \\ 0 & \text{otherwise}, \end{cases}$$

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

Let $Z_1 = \sum_{i,j} Z_{ij}$ so that $Z = Z_1 - \frac{\varepsilon}{\sqrt{mn}} x^T M y$. Note that $\sum_{i,j} a_{ij} \leq \sum_{i,j} (x_j M_{ij}^A y_j)^2 \leq M_{\text{max}}^2$. Fix $\lambda = \sqrt{mn}/2M_{\text{max}} \sqrt{C}$ so that $|\alpha_{ai,j}| \leq 1/2$, whence $e^{\lambda a_{ij}} - 1 \leq \lambda a_{ij} + 2(\lambda a_{ij})^2$. It then follows that

$$\mathbb{E}[e^{\lambda Z}] = \exp\left\{ \varepsilon \mathbb{E}\left[ \sum_{i,j} a_{ij} + 2 \sum_{i,j} (\lambda a_{ij})^2 \right] - \frac{\lambda \varepsilon}{\sqrt{mn}} x^T M y \right\} \leq \exp\left\{ \lambda \mathbb{E}[Z] + \frac{\lambda^2}{\sqrt{mn}/2} \right\}.$$  

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

Note that $\mathbb{P}(Z > \lambda M_{\text{max}} \sqrt{\varepsilon})$ can also be bounded analogously. We can now finish the upper bound on the light couples contribution. Consider the error event Eq. (11). A simple volume calculation shows that $|T_m| \leq (10/\Delta)^m$. We can therefore apply union bound over $T_m$ and $T_n$ to Eq. (12) to obtain

$$\mathbb{P}\{\mathcal{H}(E,A)\} \leq 2 \left( \frac{20}{\Delta} \right)^{n+m} 2^{(n+m)H(\delta)} e^{-\frac{(C_1+2)^2 \delta \varepsilon\gamma_{mn}}{2}} + \frac{1}{n^3},$$

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

**VI. PROOF OF LEMMA IV.1**

Recall the variational principle for the singular values.

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

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

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

Using Eq. (13) with $H$ the orthogonal complement of $\text{span}(v_1, \ldots, v_{q-1})$, we have, by Lemma IV.2,

$$\sigma_q \leq \max_{y \in H, ||y|| = 1} ||M^E y|| \leq \frac{\epsilon}{\sqrt{\min}} \left( \max_{y \in H, ||y|| = 1} ||M y|| \right) + \epsilon \sum_q + CM_{\text{max}} \sqrt{\varepsilon}$$

The lower bound is proved analogously, by using Eq. (14) with $H = \text{span}(v_1, \ldots, v_q)$.

**References**


