Journal of Machine Learning Research 11 (2010) 2057-2078 Submitted 6/09; Revised 4/10; Published 7/10

Matrix Completion from Noisy Entries

Raghunandan H. Keshavan RAGHURAM@STANFORD.EDU
Andrea Montanari* MONTANARI@STANFORD.EDU
Sewoong Oh SWOH@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94304, USA

Editor: Tommi Jaakkola

Abstract

Given a matrix M of low-rank, we consider the problem of reconstructing it from noisy observations of a small, random subset of its entries. The problem arises in a variety of applications, from collaborative filtering (the 'Netflix problem') to structure-from-motion and positioning. We study a low complexity algorithm introduced by Keshavan, Montanari, and Oh (2010), based on a combination of spectral techniques and manifold optimization, that we call here OPTSPACE. We prove performance guarantees that are order-optimal in a number of circumstances.

Keywords: matrix completion, low-rank matrices, spectral methods, manifold optimization

1. Introduction

Spectral techniques are an authentic workhorse in machine learning, statistics, numerical analysis, and signal processing. Given a matrix M, its largest singular values—and the associated singular vectors—'explain' the most significant correlations in the underlying data source. A low-rank approximation of M can further be used for low-complexity implementations of a number of linear algebra algorithms (Frieze et al., 2004).

In many practical circumstances we have access only to a sparse subset of the entries of an $m \times n$ matrix M. It has recently been discovered that, if the matrix M has rank r, and unless it is too 'structured', a small random subset of its entries allow to reconstruct it *exactly*. This result was first proved by Candès and Recht (2008) by analyzing a convex relaxation introduced by Fazel (2002). A tighter analysis of the same convex relaxation was carried out by Candès and Tao (2009). A number of iterative schemes to solve the convex optimization problem appeared soon thereafter (Cai et al., 2008; Ma et al., 2009; Toh and Yun, 2009).

In an alternative line of work, Keshavan, Montanari, and Oh (2010) attacked the same problem using a combination of spectral techniques and manifold optimization: We will refer to their algorithm as OPTSPACE. OPTSPACE is intrinsically of low complexity, the most complex operation being computing r singular values (and the corresponding singular vectors) of a sparse $m \times n$ matrix. The performance guarantees proved by Keshavan et al. (2010) are comparable with the information theoretic lower bound: roughly $nr \max\{r, \log n\}$ random entries are needed to reconstruct M exactly (here we assume m of order n). A related approach was also developed by Lee and Bresler (2009), although without performance guarantees for matrix completion.

*. Also in Department of Statistics.

The above results crucially rely on the assumption that M is *exactly* a rank r matrix. For many applications of interest, this assumption is unrealistic and it is therefore important to investigate their robustness. Can the above approaches be generalized when the underlying data is 'well approximated' by a rank r matrix? This question was addressed by Candès and Plan (2009) within the convex relaxation approach of Candès and Recht (2008). The present paper proves a similar robustness result for OPTSPACE. Remarkably the guarantees we obtain are order-optimal in a variety of circumstances, and improve over the analogous results of Candès and Plan (2009).

1.1 Model Definition

Let M be an $m \times n$ matrix of rank r, that is

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

where U has dimensions $m \times r$, V has dimensions $n \times r$, and Σ is a diagonal $r \times r$ matrix. We assume that each entry of M is perturbed, thus producing an 'approximately' low-rank matrix N, with

$$N_{ij} = M_{ij} + Z_{ij} ,$$

where the matrix Z will be assumed to be 'small' in an appropriate sense.

Out of the $m \times n$ entries of N, a subset $E \subseteq [m] \times [n]$ is revealed. We let N^E be the $m \times n$ matrix that contains the revealed entries of N, and is filled with 0's in the other positions

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

Analogously, we let M^E and Z^E be the $m \times n$ matrices that contain the entries of M and Z, respectively, in the revealed positions and is filled with 0's in the other positions. The set E will be uniformly random given its size $|E|$.

1.2 Algorithm

For the reader's convenience, we recall the algorithm introduced by Keshavan et al. (2010), which we will analyze here. The basic idea is to minimize the cost function $F(X,Y)$, defined by

$$\begin{aligned} F(X,Y) &\equiv \min_{S \in \mathbb{R}^{r\times r}} \mathcal{F}(X,Y,S) , \\ \mathcal{F}(X,Y,S) &\equiv \frac{1}{2} \sum_{(i,j)\in E} (N_{ij} - (XSY^T)_{ij})^2 . \end{aligned} \tag{2}$$

Here $X \in \mathbb{R}^{n\times r}$, $Y \in \mathbb{R}^{m\times r}$ are orthogonal matrices, normalized by $X^TX = m\mathbf{I}$, $Y^TY = n\mathbf{I}$.

Minimizing $F(X,Y)$ is an *a priori* difficult task, since F is a non-convex function. The key insight is that the singular value decomposition (SVD) of N^E provides an excellent initial guess, and that the minimum can be found with high probability by standard gradient descent after this initialization. Two caveats must be added to this description: (1) In general the matrix N^E must be 'trimmed' to eliminate over-represented rows and columns; (2) For technical reasons, we consider a slightly modified cost function to be denoted by $\widetilde{F}(X,Y)$.

OPTSPACE(matrix N^E)

1: Trim N^E, and let $\widetilde{N}^E$ be the output;
2: Compute the rank-r projection of $\widetilde{N}^E$, $\mathsf{P}_r(\widetilde{N}^E) = X_0 S_0 Y_0^T$;
3: Minimize $\widetilde{F}(X,Y)$ through gradient descent, with initial condition (X_0, Y_0).

We may note here that the rank of the matrix M, if not known, can be reliably estimated from $\widetilde{N}^E$ (Keshavan and Oh, 2009).

The various steps of the above algorithm are defined as follows.

Trimming. We say that a row is 'over-represented' if it contains more than $2|E|/m$ revealed entries (i.e., more than twice the average number of revealed entries per row). Analogously, a column is over-represented if it contains more than $2|E|/n$ revealed entries. The trimmed matrix $\widetilde{N}^E$ is obtained from N^E by setting to 0 over-represented rows and columns.

Rank-r projection. Let

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

be the singular value decomposition of $\widetilde{N}^E$, with singular values $\sigma_1 \geq \sigma_2 \geq \dots$. We then define

$$\mathsf{P}_r(\widetilde{N}^E) = \frac{mn}{|E|} \sum_{i=1}^{r} \sigma_i x_i y_i^T .$$

Apart from an overall normalization, $\mathsf{P}_r(\widetilde{N}^E)$ is the best rank-r approximation to $\widetilde{N}^E$ in Frobenius norm.

Minimization. The modified cost function $\widetilde{F}$ is defined as

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

where $X^{(i)}$ denotes the i-th row of X, and $Y^{(j)}$ the j-th row of Y. The function $G_1 : \mathbb{R}^+ \to \mathbb{R}$ is such that $G_1(z) = 0$ if $z \leq 1$ and $G_1(z) = e^{(z-1)^2} - 1$ otherwise. Further, we can choose $\rho = \Theta(|E|)$.

Let us stress that the regularization term is mainly introduced for our proof technique to work (and a broad family of functions G_1 would work as well). In numerical experiments we did not find any performance loss in setting $\rho = 0$.

One important feature of OPTSPACE is that $F(X,Y)$ and $\widetilde{F}(X,Y)$ are regarded as functions of the r-dimensional subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$ generated (respectively) by the columns of X and Y. This interpretation is justified by the fact that $F(X,Y) = F(XA, YB)$ for any two orthogonal matrices A, $B \in \mathbb{R}^{r\times r}$ (the same property holds for $\widetilde{F}$). The set of r dimensional subspaces of $\mathbb{R}^m$ is a differentiable Riemannian manifold $\mathsf{G}(m,r)$ (the Grassmann manifold). The gradient descent algorithm is applied to the function $\widetilde{F} : \mathsf{M}(m,n) \equiv \mathsf{G}(m,r) \times \mathsf{G}(n,r) \to \mathbb{R}$. For further details on optimization by gradient descent on matrix manifolds we refer to Edelman et al. (1999) and Absil et al. (2008).

1.3 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 absorbed by redefining Σ).

We shall write $\Sigma = \mathrm{diag}(\Sigma_1, \dots, \Sigma_r)$ with $\Sigma_1 \geq \Sigma_2 \geq \cdots \geq \Sigma_r > 0$. The maximum and minimum singular values will also be denoted by $\Sigma_{\max} = \Sigma_1$ and $\Sigma_{\min} = \Sigma_r$. Further, the maximum size of an entry of M is $M_{\max} \equiv \max_{ij} |M_{ij}|$.

Probability is taken with respect to the uniformly random subset $E \subseteq [m] \times [n]$ given $|E|$ and (eventually) the noise matrix Z. Define $\varepsilon \equiv |E|/\sqrt{mn}$. In the case when $m = n$, ε corresponds to the average number of revealed entries per row or column. Then it is convenient to work with a model in which each entry is revealed independently with probability $\varepsilon/\sqrt{mn}$. Since, with high probability $|E| \in [\varepsilon\sqrt{\alpha}n - A\sqrt{n\log n}, \varepsilon\sqrt{\alpha}n + A\sqrt{n\log n}]$, 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. We will use C, C' etc. to denote universal numerical constants.

It is convenient to define the following projection operator $\mathcal{P}_E(\cdot)$ as the sampling operator, which maps an $m \times n$ matrix onto an $|E|$-dimensional subspace in $\mathbb{R}^{m\times n}$

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

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 $\langle X, Y\rangle \equiv \mathrm{Tr}(X^T Y)$, respectively. Finally, we use the standard combinatorics notation $[n] = \{1, 2, \dots, n\}$ to denote the set of first n integers.

1.4 Main Results

Our main result is a performance guarantee for OPTSPACE under appropriate incoherence assumptions, and is presented in Section 1.4.2. Before presenting it, we state a theorem of independent interest that provides an error bound on the simple trimming-plus-SVD approach. The reader interested in the OPTSPACE guarantee can go directly to Section 1.4.2.

Throughout this paper, without loss of generality, we assume $\alpha \equiv m/n \geq 1$.

1.4.1 SIMPLE SVD

Our first result shows that, in great generality, the rank-r projection of $\widetilde{N}^E$ provides a reasonable approximation of M. We define $\widetilde{Z}^E$ to be an $m \times n$ matrix obtained from Z^E, after the trimming step of the pseudocode above, that is, by setting to zero the over-represented rows and columns.

Theorem 1.1 *Let $N = M + Z$, where M has rank r, and assume that the subset of revealed entries $E \subseteq [m] \times [n]$ is uniformly random with size $|E|$. Let $M_{\max} = \max_{(i,j)\in[m]\times[n]} |M_{ij}|$. Then there exists numerical constants C and C' such that*

$$\frac{1}{\sqrt{mn}} \|M - \mathsf{P}_r(\widetilde{N}^E)\|_F \leq C M_{\max} \left(\frac{nr\alpha^{3/2}}{|E|}\right)^{1/2} + C' \frac{n\sqrt{r\alpha}}{|E|} \|\widetilde{Z}^E\|_2\,,$$

with probability larger than $1-1/n^3$.

Projection onto rank-r matrices through SVD is a pretty standard tool, and is used as first analysis method for many practical problems. At a high-level, projection onto rank-r matrices can be interpreted as 'treat missing entries as zeros'. This theorem shows that this approach is reasonably robust if the number of observed entries is as large as the number of degrees of freedom (which is about $(m+n)r$) times a large constant. The error bound is the sum of two contributions: the first one can be interpreted as an undersampling effect (error induced by missing entries) and the second as a noise effect. Let us stress that trimming is crucial for achieving this guarantee.

1.4.2 OPTSPACE

Theorem 1.1 helps to set the stage for the key point of this paper: *a much better approximation is obtained by minimizing the cost* $\widetilde{F}(X,Y)$ *(step 3 in the pseudocode above), provided M satisfies an appropriate incoherence condition.* Let $M = U\Sigma V^T$ be a low rank matrix, and assume, without loss of generality, $U^TU = m\mathbf{I}$ and $V^TV = n\mathbf{I}$. We say that M is (μ_0,μ_1)-*incoherent* if the following conditions hold.

A1. For all $i \in [m]$, $j \in [n]$ we have, $\sum_{k=1}^r U_{ik}^2 \le \mu_0 r$, $\sum_{k=1}^r V_{ik}^2 \le \mu_0 r$.

A2. For all $i \in [m]$, $j \in [n]$ we have, $|\sum_{k=1}^r U_{ik}(\Sigma_k/\Sigma_1)V_{jk}| \le \mu_1 r^{1/2}$.

Theorem 1.2 *Let $N = M + Z$, where M is a (μ_0,μ_1)-incoherent matrix of rank r, and assume that the subset of revealed entries $E \subseteq [m]\times[n]$ is uniformly random with size $|E|$. Further, let $\Sigma_{\min} = \Sigma_r \le \cdots \le \Sigma_1 = \Sigma_{\max}$ with $\Sigma_{\max}/\Sigma_{\min} \equiv \kappa$. Let $\widehat{M}$ be the output of* OPTSPACE *on input N^E. Then there exists numerical constants C and C' such that if*

$$|E| \;\ge\; Cn\sqrt{\alpha}\kappa^2 \max\left\{\mu_0 r\sqrt{\alpha}\log n\,;\, \mu_0^2 r^2\alpha\kappa^4\,;\, \mu_1^2 r^2\alpha\kappa^4\right\}\,,$$

then, with probability at least $1-1/n^3$,

$$\frac{1}{\sqrt{mn}}\|\widehat{M}-M\|_F \le C'\kappa^2\frac{n\sqrt{r\alpha}}{|E|}\|Z^E\|_2\,. \tag{3}$$

provided that the right-hand side is smaller than $\Sigma_{\min}$.

As discussed in the next section, this theorem captures rather sharply the effect of important classes of noise on the performance of OPTSPACE.

1.5 Noise Models

In order to make sense of the above results, it is convenient to consider a couple of simple models for the noise matrix Z:

Independent entries model. We assume that Z's entries are i.i.d. random variables, with zero mean $\mathbb{E}\{Z_{ij}\} = 0$ and sub-Gaussian tails. The latter means that

$$\mathbb{P}\{|Z_{ij}| \ge x\} \le 2e^{-\frac{x^2}{2\sigma^2}}\,,$$

for some constant σ^2 uniformly bounded in n.

Worst case model. In this model Z is arbitrary, but we have an uniform bound on the size of its entries: $|Z_{ij}| \le Z_{\max}$.

The basic parameter entering our main results is the operator norm of $\widetilde{Z}^E$, which is bounded as follows in these two noise models.

Theorem 1.3 *If Z is a random matrix drawn according to the independent entries model, then for any sample size $|E|$ there is a constant C such that,*

$$\|\widetilde{Z}^E\|_2 \le C\sigma\left(\frac{|E|\log n}{n}\right)^{1/2}, \tag{4}$$

with probability at least $1-1/n^3$. Further there exists a constant C' such that, if the sample size is $|E| \ge n\log n$ (for $n \ge \alpha$), we have

$$\|\widetilde{Z}^E\|_2 \le C'\sigma\left(\frac{|E|}{n}\right)^{1/2}, \tag{5}$$

with probability at least $1-1/n^3$.

If Z is a matrix from the worst case model, then

$$\|\widetilde{Z}^E\|_2 \le \frac{2|E|}{n\sqrt{\alpha}}Z_{\max},$$

for any realization of E.

It is elementary to show that, if $|E| \ge 15\alpha n\log n$, no row or column is over-represented with high probability. It follows that in the regime of $|E|$ for which the conditions of Theorem 1.2 are satisfied, we have $Z^E = \widetilde{Z}^E$ and hence the bound (5) applies to $\|\widetilde{Z}^E\|_2$ as well. Then, among the other things, this result implies that for the independent entries model the right-hand side of our error estimate, Eq. (3), is with high probability smaller than $\Sigma_{\min}$, if $|E| \ge Cr\alpha n\kappa^4(\sigma/\Sigma_{\min})^2$. For the worst case model, the same statement is true if $Z_{\max} \le \Sigma_{\min}/C\sqrt{r}\kappa^2$.

1.6 Comparison with Other Approaches to Matrix Completion

Let us begin by mentioning that a statement analogous to our preliminary Theorem 1.1 was proved by Achlioptas and McSherry (2007). Our result however applies to any number of revealed entries, while the one of Achlioptas and McSherry (2007) requires $|E| \ge (8\log n)^4 n$ (which for $n \le 5\cdot 10^8$ is larger than n^2). We refer to Section 1.8 for further discussion of this point.

As for Theorem 1.2, we will mainly compare our algorithm with the convex relaxation approach recently analyzed by Candès and Plan (2009), and based on semidefinite programming. Our basic setting is indeed the same, while the algorithms are rather different.

Figures 1 and 2 compare the average root mean square error $\|\widehat{M}-M\|_F/\sqrt{mn}$ for the two algorithms as a function of $|E|$ and the rank-r respectively. Here M is a random rank r matrix of dimension $m = n = 600$, generated by letting $M = \widetilde{U}\widetilde{V}^T$ with $\widetilde{U}_{ij}, \widetilde{V}_{ij}$ i.i.d. $N(0, 20/\sqrt{n})$. The noise is distributed according to the independent noise model with $Z_{ij} \sim N(0,1)$. In the first suite of simulations, presented in Figure 1, the rank is fixed to $r = 2$. In the second one (Figure 2), the number of samples is fixed to $|E| = 72000$. These examples are taken from Candès and Plan (2009, Figure

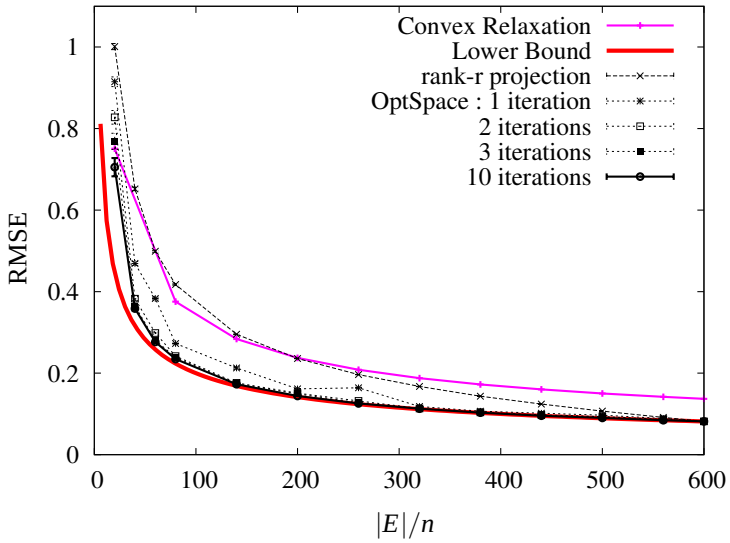

Figure 1: Numerical simulation with random rank-2 600×600 matrices. Root mean square error achieved by OPTSPACE is shown as a function of the number of observed entries $|E|$ and of the number of line minimizations. The performance of nuclear norm minimization and an information theoretic lower bound are also shown.

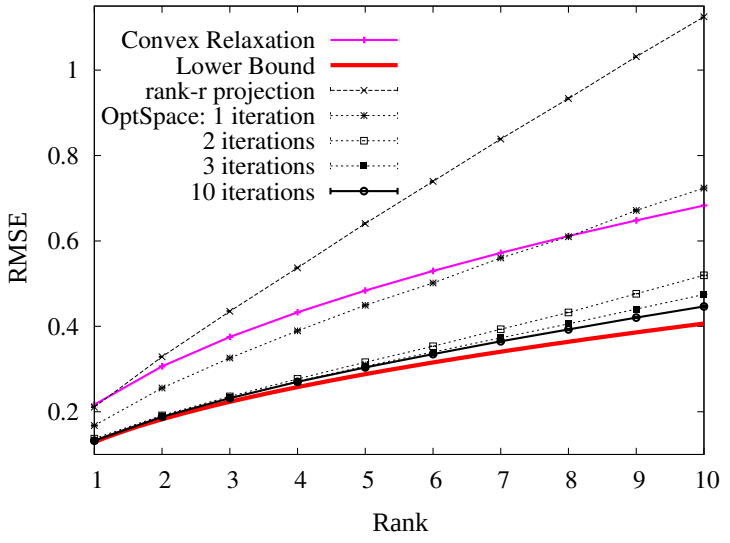

Figure 2: Numerical simulation with random rank-r 600×600 matrices and number of observed entries $|E|/n = 120$. Root mean square error achieved by OPTSPACE is shown as a function of the rank and of the number of line minimizations. The performance of nuclear norm minimization and an information theoretic lower bound are also shown.

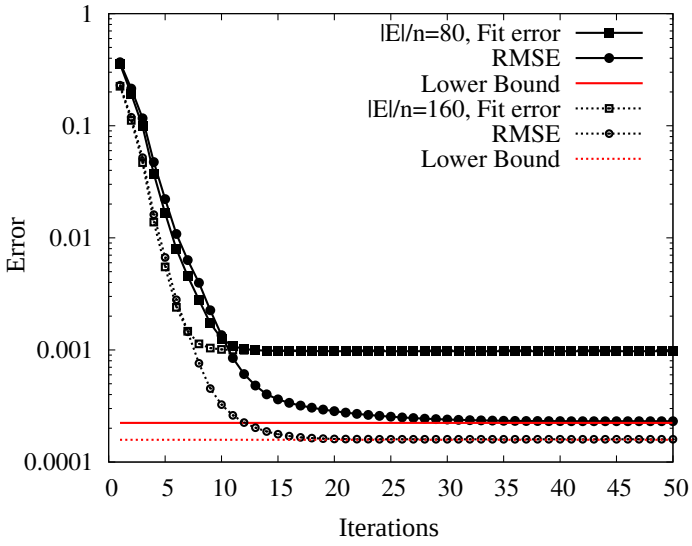

Figure 3: Numerical simulation with random rank-2 600×600 matrices and number of observed entries $|E|/n = 80$ and 160. The standard deviation of the i.i.d. Gaussian noise is 0.001. Fit error and root mean square error achieved by OPTSPACE are shown as functions of the number of line minimizations. Information theoretic lower bounds are also shown.

2), from which we took the data points for the convex relaxation approach, as well as the information theoretic lower bound described later in this section. After a few iterations, OPTSPACE has a smaller root mean square error than the one produced by convex relaxation. In about 10 iterations it becomes indistinguishable from the information theoretic lower bound for small ranks.

In Figure 3, we illustrate the rate of convergence of OPTSPACE. Two metrics, root mean squared error(RMSE) and fit error $\|\mathcal{P}_E(\widehat{M}-N)\|_F/\sqrt{|E|}$, are shown as functions of the number of iterations in the manifold optimization step. Note, that the fit error can be easily evaluated since $N^E = \mathcal{P}_E(N)$ is always available at the estimator. M is a random 600×600 rank-2 matrix generated as in the previous examples. The additive noise is distributed as $Z_{ij} \sim N(0, \sigma^2)$ with $\sigma = 0.001$ (A small noise level was used in order to trace the RMSE evolution over many iterations). Each point in the figure is the averaged over 20 random instances, and resulting errors for two different values of sample size $|E| = 80$ and $|E| = 160$ are shown. In both cases, we can see that the RMSE converges to the information theoretic lower bound described later in this section. The fit error decays exponentially with the number iterations and converges to the standard deviation of the noise which is 0.001. This is a lower bound on the fit error when $r \ll n$, since even if we have a perfect reconstruction of M, the average fit error is still 0.001.

For a more complete numerical comparison between various algorithms for matrix completion, including different noise models, real data sets and ill conditioned matrices, we refer to Keshavan and Oh (2009).

Next, let us compare our main result with the performance guarantee of Candès and Plan (2009, Theorem 7). Let us stress that we require the condition number κ to be bounded, while the analysis of Candès and Plan (2009) and Candès and Tao (2009) requires a stronger incoherence assumption

(compared to our **A1**). Therefore the assumptions are not directly comparable. As far as the error bound is concerned, Candès and Plan (2009) proved that the semidefinite programming approach returns an estimate $\widehat{M}$ which satisfies

$$\frac{1}{\sqrt{mn}}\|\widehat{M}_{\mathrm{SDP}}-M\|_F \le 7\sqrt{\frac{n}{|E|}}\,\|Z^E\|_F+\frac{2}{n\sqrt{\alpha}}\,\|Z^E\|_F\,. \tag{6}$$

(The constant in front of the first term is in fact slightly smaller than 7 in Candès and Plan (2009), but in any case larger than $4\sqrt{2}$. We choose to quote a result which is slightly less accurate but easier to parse.)

Theorem 1.2 improves over this result in several respects: (1) We do not have the second term on the right-hand side of (6), that actually increases with the number of observed entries; (2) Our error decreases as $n/|E|$ rather than $(n/|E|)^{1/2}$; (3) The noise enters Theorem 1.2 through the operator norm $\|Z^E\|_2$ instead of its Frobenius norm $\|Z^E\|_F \ge \|Z^E\|_2$. For E uniformly random, one expects $\|Z^E\|_F$ to be roughly of order $\|Z^E\|_2\sqrt{n}$. For instance, within the independent entries model with bounded variance σ, $\|Z^E\|_F=\Theta(\sqrt{|E|})$ while $\|Z^E\|_2$ is of order $\sqrt{|E|/n}$ (up to logarithmic terms).

Theorem 1.2 can also be compared to an information theoretic lower bound computed by Candès and Plan (2009). Suppose, for simplicity, $m=n$ and assume that an oracle provides us a linear subspace T where the correct rank r matrix $M=U\Sigma V^T$ lies. More precisely, we know that $M\in T$ where T is a linear space of dimension $2nr-r^2$ defined by

$$T=\{UY^T+XV^T \mid X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{n\times r}\}\,.$$

Notice that the rank constraint is therefore replaced by this simple linear constraint. The minimum mean square error estimator is computed by projecting the revealed entries onto the subspace T, which can be done by solving a least squares problem. Candès and Plan (2009) analyzed the root mean squared error of the resulting estimator $\widehat{M}$ and showed that

$$\frac{1}{\sqrt{mn}}\|\widehat{M}_{\mathrm{Oracle}}-M\|_F \approx \sqrt{\frac{1}{|E|}}\,\|Z^E\|_F\,.$$

Here $\approx$ indicates that the root mean squared error concentrates in probability around the right-hand side.

For the sake of comparison, suppose we have i.i.d. Gaussian noise with variance σ^2. In this case the oracle estimator yields (for $r=o(n)$)

$$\frac{1}{\sqrt{mn}}\|\widehat{M}_{\mathrm{Oracle}}-M\|_F \approx \sigma\sqrt{\frac{2nr}{|E|}}\,.$$

The bound (6) on the semidefinite programming approach yields

$$\frac{1}{\sqrt{mn}}\|\widehat{M}_{\mathrm{SDP}}-M\|_F \le \sigma\left(7\sqrt{n|E|}+\frac{2}{n}|E|\right).$$

Finally, using Theorems 1.2 and 1.3 we deduce that OPTSPACE achieves

$$\frac{1}{\sqrt{mn}}\|\widehat{M}_{\mathrm{OptSpace}}-M\|_F \le \sigma\sqrt{\frac{Cnr}{|E|}}\,.$$

Hence, when the noise is i.i.d. Gaussian with small enough σ, OPTSPACE is order-optimal.

1.7 Related Work on Gradient Descent

Local optimization techniques such as gradient descent of coordinate descent have been intensively studied in machine learning, with a number of applications. Here we will briefly review the recent literature on the use of such techniques within collaborative filtering applications.

Collaborative filtering was studied from a graphical models perspective in Salakhutdinov et al. (2007), which introduced an approach to prediction based on Restricted Boltzmann Machines (RBM). Exact learning of the model parameters is intractable for such models, but the authors studied the performances of a *contrastive divergence*, which computes an approximate gradient of the likelihood function, and uses it to optimize the likelihood locally. Based on empirical evidence, it was argued that RBM's have several advantages over spectral methods for collaborative filtering.

An objective function analogous to the one used in the present paper was considered early on in Srebro and Jaakkola (2003), which uses gradient descent in the factors to minimize a weighted sum of square residuals. Salakhutdinov and Mnih (2008) justified the use of such an objective function by deriving it as the (negative) log-posterior of an appropriate probabilistic model. This approach naturally lead to the use of quadratic regularization in the factors. Again, gradient descent in the factors was used to perform the optimization. Also, this paper introduced a logistic mapping between the low-rank matrix and the recorded ratings.

Recently, this line of work was pushed further in Salakhutdinov and Srebro (2010), which emphasize the advantage of using a non-uniform quadratic regularization in the factors. The basic objective function was again a sum of square residuals, and version of stochastic gradient descent was used to optimize it.

This rich and successful line of work emphasizes the importance of obtaining a rigorous understanding of methods based on local minimization of the sum of square residuals with respect to the factors. The present paper provides a first step in that direction. Hopefully the techniques developed here will be useful to analyze the many variants of this approach.

The relationship between the non-convex objective function and convex relaxation introduced by Fazel (2002) was further investigated by Srebro et al. (2005) and Recht et al. (2007). The basic relation is provided by the identity

$$\|M\|_* = \frac{1}{2} \min_{M=XY^T} \left\{ \|X\|_F^2 + \|Y\|_F^2 \right\}, \tag{7}$$

where $\|M\|_*$ denotes the nuclear norm of M (the sum of its singular values). In other words, adding a regularization term that is quadratic in the factors (as the one used in much of the literature reviewed above) is equivalent to weighting M by its nuclear norm, that can be regarded as a convex surrogate of its rank.

In view of the identity (7) it might be possible to use the results in this paper to prove stronger guarantees on the nuclear norm minimization approach. Unfortunately this implication is not immediate. Indeed in the present paper we assume the correct rank r is known, while on the other hand we do not use a quadratic regularization in the factors. (See Keshavan and Oh, 2009 for a procedure that estimates the rank from the data and is provably successful under the hypotheses of Theorem 1.2.) Trying to establish such an implication, and clarifying the relation between the two approaches is nevertheless a promising research direction.

1.8 On the Spectrum of Sparse Matrices and the Role of Trimming

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

Let us now turn to the example. Assume, for the sake of simplicity, that $m = n$, there is no noise in the revealed entries, and M is the rank one matrix with $M_{ij} = 1$ for all i and j. Within the independent sampling model, the matrix M^E has i.i.d. entries, with distribution Bernoulli(ε/n). The number of non-zero entries in a column is Binomial$(n, \varepsilon/n)$ and is independent for different columns. It is not hard to realize that the column with the largest number of entries has more than $C \log n/\log\log n$ entries, with positive probability (this probability can be made as large as we want by reducing C). Let i be the index of this column, and consider the test vector $\underline{e}^{(i)}$ that has the i-th entry equal to 1 and all the others equal to 0. By computing $\|M^E \underline{e}^{(i)}\|$, we conclude that the largest singular value of M^E is at least $\sqrt{C \log n/\log\log n}$. In particular, this is very different from the largest singular value of $\mathbb{E}\{M^E\} = (\varepsilon/n)M$ which is ε. This suggests that approximating M with the $\mathsf{P}_r(M^E)$ leads to a large error. Hence trimming is crucial in proving Theorem 1.1. Also, the phenomenon is more severe in real data sets than in the present model, where each entry is revealed independently.

Trimming is also crucial in proving Theorem 1.3. Using the above argument, it is possible to show that under the worst case model,

$$\|Z^E\|_2 \geq C'(\varepsilon)\, Z_{\max} \sqrt{\frac{\log n}{\log\log n}}\, .$$

This suggests that the largest singular value of the noise matrix Z^E is quite different from the largest singular value of $\mathbb{E}\{Z^E\}$ which is $\varepsilon Z_{\max}$.

To summarize, Theorems 1.1 and 1.3 (for the worst case model) simply do not hold without trimming or a similar procedure to normalize rows/columns of N^E. Trimming allows to overcome the above phenomenon by setting to 0 over-represented rows/columns.

2. Proof of Theorem 1.1

As explained in the introduction, the crucial idea is to consider the singular value decomposition of the trimmed matrix $\widetilde{N}^E$ instead of the original matrix N^E. Apart from a trivial rescaling, these singular values are close to the ones of the original matrix M.

Lemma 1 *There exists a numerical constant C such that, with probability greater than $1 - 1/n^3$,*

$$\left|\frac{\sigma_q}{\varepsilon} - \Sigma_q\right| \leq C M_{\max} \sqrt{\frac{\alpha}{\varepsilon}} + \frac{1}{\varepsilon}\|\widetilde{Z}^E\|_2\, ,$$

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

Proof For any matrix A, let $\sigma_q(A)$ denote the qth singular value of A. Then, $\sigma_q(A+B) \le \sigma_q(A) + \sigma_1(B)$, whence

$$\begin{aligned}\left|\frac{\sigma_q}{\varepsilon} - \Sigma_q\right| &\le \left|\frac{\sigma_q(\widetilde{M}^E)}{\varepsilon} - \Sigma_q\right| + \frac{\sigma_1(\widetilde{Z}^E)}{\varepsilon} \\ &\le CM_{\max}\sqrt{\frac{\alpha}{\varepsilon}} + \frac{1}{\varepsilon}\|\widetilde{Z}^E\|_2\,,\end{aligned}$$

where the second inequality follows from the next Lemma as shown by Keshavan et al. (2010).

Lemma 2 (Keshavan, Montanari, Oh, 2009) *There exists a numerical constant C such that, with probability larger than $1-1/n^3$,*

$$\frac{1}{\sqrt{mn}}\left|\left|M - \frac{\sqrt{mn}}{\varepsilon}\widetilde{M}^E\right|\right|_2 \le CM_{\max}\sqrt{\frac{\alpha}{\varepsilon}}\,.$$

■

We will now prove Theorem 1.1.

Proof (Theorem 1.1) For any matrix A of rank at most $2r$, $\|A\|_F \le \sqrt{2r}\|A\|_2$, whence

$$\begin{aligned}\frac{1}{\sqrt{mn}}\|M - \mathsf{P}_r(\widetilde{N}^E)\|_F &\le \frac{\sqrt{2r}}{\sqrt{mn}}\left|\left|M - \frac{\sqrt{mn}}{\varepsilon}\Big(\widetilde{N}^E - \sum_{i\ge r+1}\sigma_i x_i y_i^T\Big)\right|\right|_2 \\ &= \frac{\sqrt{2r}}{\sqrt{mn}}\left|\left|M - \frac{\sqrt{mn}}{\varepsilon}\Big(\widetilde{M}^E + \widetilde{Z}^E - \sum_{i\ge r+1}\sigma_i x_i y_i^T\Big)\right|\right|_2 \\ &= \frac{\sqrt{2r}}{\sqrt{mn}}\left|\left|\Big(M - \frac{\sqrt{mn}}{\varepsilon}\widetilde{M}^E\Big) + \frac{\sqrt{mn}}{\varepsilon}\left(\widetilde{Z}^E - \Big(\sum_{i\ge r+1}\sigma_i x_i y_i^T\Big)\right)\right|\right|_2 \\ &\le \frac{\sqrt{2r}}{\sqrt{mn}}\left(\left|\left|M - \frac{\sqrt{mn}}{\varepsilon}\widetilde{M}^E\right|\right|_2 + \frac{\sqrt{mn}}{\varepsilon}\|\widetilde{Z}^E\|_2 + \frac{\sqrt{mn}}{\varepsilon}\sigma_{r+1}\right) \\ &\le 2CM_{\max}\sqrt{\frac{2\alpha r}{\varepsilon}} + \frac{2\sqrt{2r}}{\varepsilon}\|\widetilde{Z}^E\|_2 \\ &\le C'M_{\max}\left(\frac{nr\alpha^{3/2}}{|E|}\right)^{1/2} + 2\sqrt{2}\left(\frac{n\sqrt{r\alpha}}{|E|}\right)\|\widetilde{Z}^E\|_2\,.\end{aligned}$$

where on the fourth line, we have used the fact that for any matrices A_i, $\|\sum_i A_i\|_2 \le \sum_i \|A_i\|_2$. This proves our claim. ■

3. Proof of Theorem 1.2

Recall that the cost function is defined over the Riemannian manifold $\mathsf{M}(m,n) \equiv \mathsf{G}(m,r)\times\mathsf{G}(n,r)$. The proof of Theorem 1.2 consists in controlling the behavior of F in a neighborhood of $\mathbf{u} = (U,V)$ (the point corresponding to the matrix M to be reconstructed). Throughout the proof we let $\mathcal{K}(\mu)$ be the set of matrix couples $(X,Y) \in \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r}$ such that $\|X^{(i)}\|^2 \le \mu r$, $\|Y^{(j)}\|^2 \le \mu r$ for all i,j.

3.1 Preliminary Remarks and Definitions

Given $\mathbf{x}_1 = (X_1, Y_1)$ and $\mathbf{x}_2 = (X_2, Y_2) \in \mathsf{M}(m,n)$, two points on this manifold, their distance is defined as $d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{d(X_1, X_2)^2 + d(Y_1, Y_2)^2}$, where, letting $(\cos\theta_1, \dots, \cos\theta_r)$ be the singular values of $X_1^T X_2/m$,

$$d(X_1, X_2) = \|\theta\|_2 .$$

The next remark bounds the distance between two points on the manifold. In particular, we will use this to bound the distance between the original matrix $M = U\Sigma V^T$ and the starting point of the manifold optimization $\widehat{M} = X_0 S_0 Y_0^T$.

Remark 3 (Keshavan, Montanari, Oh, 2009) *Let $U, X \in \mathbb{R}^{m\times r}$ with $U^T U = X^T X = m\mathbf{I}$, $V, Y \in \mathbb{R}^{n\times r}$ with $V^T V = Y^T Y = n\mathbf{I}$, and $M = U\Sigma V^T$, $\widehat{M} = XSY^T$ for $\Sigma = \mathrm{diag}(\Sigma_1, \dots, \Sigma_r)$ and $S \in \mathbb{R}^{r\times r}$. If $\Sigma_1, \dots, \Sigma_r \geq \Sigma_{\min}$, then*

$$d(U, X) \leq \frac{\pi}{\sqrt{2\alpha} n \Sigma_{\min}} \|M - \widehat{M}\|_F \quad , \quad d(V, Y) \leq \frac{\pi}{\sqrt{2\alpha} n \Sigma_{\min}} \|M - \widehat{M}\|_F$$

Given S achieving the minimum in Eq. (2), it is also convenient to introduce the notations

$$d_-(\mathbf{x}, \mathbf{u}) \equiv \sqrt{\Sigma_{\min}^2 d(\mathbf{x}, \mathbf{u})^2 + \|S - \Sigma\|_F^2} ,$$
$$d_+(\mathbf{x}, \mathbf{u}) \equiv \sqrt{\Sigma_{\max}^2 d(\mathbf{x}, \mathbf{u})^2 + \|S - \Sigma\|_F^2} .$$

3.2 Auxiliary Lemmas and Proof of Theorem 1.2

The proof is based on the following two lemmas that generalize and sharpen analogous bounds in Keshavan et al. (2010).

Lemma 4 *There exist numerical constants C_0, C_1, C_2 such that the following happens. Assume $\varepsilon \geq C_0 \mu_0 r \sqrt{\alpha} \max\{\log n; \mu_0 r \sqrt{\alpha} (\Sigma_{\min}/\Sigma_{\max})^4\}$ and $\delta \leq \Sigma_{\min}/(C_0 \Sigma_{\max})$. Then,*

$$F(\mathbf{x}) - F(\mathbf{u}) \geq C_1 n\varepsilon\sqrt{\alpha}\, d_-(\mathbf{x}, \mathbf{u})^2 - C_1 n\sqrt{r\alpha}\|Z^E\|_2 d_+(\mathbf{x}, \mathbf{u}) , \quad (8)$$
$$F(\mathbf{x}) - F(\mathbf{u}) \leq C_2 n\varepsilon\sqrt{\alpha}\,\Sigma_{\max}^2 d(\mathbf{x}, \mathbf{u})^2 + C_2 n\sqrt{r\alpha}\|Z^E\|_2 d_+(\mathbf{x}, \mathbf{u}) , \quad (9)$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ such that $d(\mathbf{x}, \mathbf{u}) \leq \delta$, with probability at least $1 - 1/n^4$. Here $S \in \mathbb{R}^{r\times r}$ is the matrix realizing the minimum in Eq. (2).

Corollary 3.1 *There exist a constant C such that, under the hypotheses of Lemma 4*

$$\|S - \Sigma\|_F \leq C\Sigma_{\max} d(\mathbf{x}, \mathbf{u}) + C\frac{\sqrt{r}}{\varepsilon}\|Z^E\|_2 .$$

Further, for an appropriate choice of the constants in Lemma 4, we have

$$\sigma_{\max}(S) \leq 2\Sigma_{\max} + C\frac{\sqrt{r}}{\varepsilon}\|Z^E\|_2 , \quad (10)$$
$$\sigma_{\min}(S) \geq \frac{1}{2}\Sigma_{\min} - C\frac{\sqrt{r}}{\varepsilon}\|Z^E\|_2 . \quad (11)$$

Lemma 5 *There exist numerical constants C_0, C_1, C_2 such that the following happens. Assume $\varepsilon \geq C_0 \mu_0 r \sqrt{\alpha} (\Sigma_{\max}/\Sigma_{\min})^2 \max\{\log n; \mu_0 r \sqrt{\alpha} (\Sigma_{\max}/\Sigma_{\min})^4\}$ and $\delta \leq \Sigma_{\min}/(C_0 \Sigma_{\max})$. Then,*

$$\|\mathrm{grad}\, \widetilde{F}(\mathbf{x})\|^2 \geq C_1 n \varepsilon^2 \Sigma_{\min}^4 \left[d(\mathbf{x},\mathbf{u}) - C_2 \frac{\sqrt{r}\Sigma_{\max}}{\varepsilon \Sigma_{\min}} \frac{\|Z^E\|_2}{\Sigma_{\min}} \right]_+^2 , \tag{12}$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \leq \delta$, with probability at least $1 - 1/n^4$. (Here $[a]_+ \equiv \max(a,0)$.)

We can now turn to the proof of our main theorem.

Proof (Theorem 1.2). Let $\delta = \Sigma_{\min}/C_0\Sigma_{\max}$ with C_0 large enough so that the hypotheses of Lemmas 4 and 5 are verified.

Call $\{\mathbf{x}_k\}_{k\geq 0}$ the sequence of pairs $(X_k, Y_k) \in \mathsf{M}(m,n)$ generated by gradient descent. By assumption the right-hand side of Eq. (3) is smaller than $\Sigma_{\min}$. The following is therefore true for some numerical constant C:

$$\|Z^E\|_2 \leq \frac{\varepsilon}{C\sqrt{r}} \left(\frac{\Sigma_{\min}}{\Sigma_{\max}} \right)^2 \Sigma_{\min} . \tag{13}$$

Notice that the constant appearing here can be made as large as we want by modifying the constant appearing in the statement of the theorem. Further, by using Corollary 3.1 in Eqs. (8) and (9) we get

$$F(\mathbf{x}) - F(\mathbf{u}) \geq C_1 n \varepsilon \sqrt{\alpha} \Sigma_{\min}^2 \left\{ d(\mathbf{x},\mathbf{u})^2 - \delta_{0,-}^2 \right\} , \tag{14}$$

$$F(\mathbf{x}) - F(\mathbf{u}) \leq C_2 n \varepsilon \sqrt{\alpha} \Sigma_{\max}^2 \left\{ d(\mathbf{x},\mathbf{u})^2 + \delta_{0,+}^2 \right\} , \tag{15}$$

with C_1 and C_2 different from those in Eqs. (8) and (9), where

$$\delta_{0,-} \equiv C \frac{\sqrt{r}\Sigma_{\max}}{\varepsilon \Sigma_{\min}} \frac{\|Z^E\|_2}{\Sigma_{\min}} , \qquad \delta_{0,+} \equiv C \frac{\sqrt{r}\Sigma_{\max}}{\varepsilon \Sigma_{\min}} \frac{\|Z^E\|_2}{\Sigma_{\max}} .$$

By Eq. (13), with large enough C, we can assume $\delta_{0,-} \leq \delta/20$ and $\delta_{0,+} \leq (\delta/20)(\Sigma_{\min}/\Sigma_{\max})$.

Next, we provide a bound on $d(\mathbf{u},\mathbf{x}_0)$. Using Remark 3, we have $d(\mathbf{u},\mathbf{x}_0) \leq (\pi/n\sqrt{\alpha}\Sigma_{\min}) \|M - X_0 S_0 Y_0^T\|_F$. Together with Theorem 1.1 this implies

$$d(\mathbf{u},\mathbf{x}_0) \leq \frac{C M_{\max}}{\Sigma_{\min}} \left(\frac{r\alpha}{\varepsilon} \right)^{1/2} + \frac{C'\sqrt{r}}{\varepsilon \Sigma_{\min}} \|\widetilde{Z}^E\|_2 .$$

Since $\varepsilon \geq C'' \alpha \mu_1^2 r^2 (\Sigma_{\max}/\Sigma_{\min})^4$ as per our assumptions and $M_{\max} \leq \mu_1 \sqrt{r}\Sigma_{\max}$ for incoherent M, the first term in the above bound is upper bounded by $\Sigma_{\min}/20C_0\Sigma_{\max}$, for large enough C''. Using Eq. (13), with large enough constant C, the second term in the above bound is upper bounded by $\Sigma_{\min}/20C_0\Sigma_{\max}$. Hence we get

$$d(\mathbf{u},\mathbf{x}_0) \leq \frac{\delta}{10} .$$

We make the following claims :

1. $\mathbf{x}_k \in \mathcal{K}(4\mu_0)$ for all k.

   First we notice that we can assume $\mathbf{x}_0 \in \mathcal{K}(3\mu_0)$. Indeed, if this does not hold, we can 'rescale' those rows of X_0, Y_0 that violate the constraint. A proof that this rescaling is possible was given in Keshavan et al. (2010) (cf. Remark 6.2 there). We restate the result here for the reader's convenience in the next Remark.

   Remark 6 *Let $U, X \in \mathbb{R}^{n\times r}$ with $U^TU = X^TX = n\mathbf{I}$ and $U \in \mathcal{K}(\mu_0)$ and $d(X,U) \le \delta \le \frac{1}{16}$. Then there exists $X' \in \mathbb{R}^{n\times r}$ such that $X'^TX' = n\mathbf{I}$, $X' \in \mathcal{K}(3\mu_0)$ and $d(X',U) \le 4\delta$. Further, such an X' can be computed from X in a time of $O(nr^2)$.*

   Since $\mathbf{x}_0 \in \mathcal{K}(3\mu_0)$, $\widetilde{F}(\mathbf{x}_0) = F(\mathbf{x}_0) \le 4C_2 n\varepsilon\sqrt{\alpha}\Sigma_{\max}^2\delta^2/100$. On the other hand $\widetilde{F}(\mathbf{x}) \ge \rho(e^{1/9}-1)$ for $\mathbf{x} \notin \mathcal{K}(4\mu_0)$. Since $\widetilde{F}(\mathbf{x}_k)$ is a non-increasing sequence, the thesis follows provided we take $\rho \ge C_2 n\varepsilon\sqrt{\alpha}\Sigma_{\min}^2$.

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

   Since $\varepsilon \ge C\alpha\mu_1^2 r^2(\Sigma_{\max}/\Sigma_{\min})^6$ as per our assumptions in Theorem 1.2, we have $d(\mathbf{x}_0,\mathbf{u})^2 \le (C_1\Sigma_{\min}^2/C_2\Sigma_{\max}^2)(\delta/20)^2$. Also assuming Eq. (13) with large enough C, we have $\delta_{0,-} \le \delta/20$ and $\delta_{0,+} \le (\delta/20)(\Sigma_{\min}/\Sigma_{\max})$. Then, by Eq. (15),

   $$F(\mathbf{x}_0) \le F(\mathbf{u}) + C_1 n\varepsilon\sqrt{\alpha}\Sigma_{\min}^2\frac{2\delta^2}{400}.$$

   Also, using Eq. (14), for all $\mathbf{x}_k$ such that $d(\mathbf{x}_k,\mathbf{u}) \in [\delta/10,\delta]$, we have

   $$F(\mathbf{x}) \ge F(\mathbf{u}) + C_1 n\varepsilon\sqrt{\alpha}\Sigma_{\min}^2\frac{3\delta^2}{400}.$$

   Hence, for all $\mathbf{x}_k$ such that $d(\mathbf{x}_k,\mathbf{u}) \in [\delta/10,\delta]$, we have $\widetilde{F}(\mathbf{x}) \ge F(\mathbf{x}) \ge F(\mathbf{x}_0)$. This contradicts the monotonicity of $\widetilde{F}(\mathbf{x})$, and thus proves the claim.

Since the cost function is twice differentiable, and because of the above two claims, the sequence $\{\mathbf{x}_k\}$ converges to

$$\Omega = \left\{\mathbf{x} \in \mathcal{K}(4\mu_0) \cap \mathsf{M}(m,n) : d(\mathbf{x},\mathbf{u}) \le \delta, \operatorname{grad}\widetilde{F}(\mathbf{x}) = 0\right\}.$$

By Lemma 5 for any $\mathbf{x} \in \Omega$,

$$d(\mathbf{x},\mathbf{u}) \le C\frac{\sqrt{r}\Sigma_{\max}}{\varepsilon\Sigma_{\min}}\frac{\|Z^E\|_2}{\Sigma_{\min}}. \tag{16}$$

Using Corollary 3.1, we have $d_+(\mathbf{x},\mathbf{u}) \le \Sigma_{\max}d(\mathbf{x},\mathbf{u}) + \|S-\Sigma\|_F \le C\Sigma_{\max}d(\mathbf{x},\mathbf{u}) + C(\sqrt{r}/\varepsilon)\|Z^E\|_2$. Together with Eqs. (18) and (16), this implies

$$\frac{1}{n\sqrt{\alpha}}\|M - XSY^T\|_F \le C\frac{\sqrt{r}\Sigma_{\max}^2\|Z^E\|_2}{\varepsilon\Sigma_{\min}^2},$$

which finishes the proof of Theorem 1.2. ■

3.3 Proof of Lemma 4 and Corollary 3.1

Proof (Lemma 4) The proof is based on the analogous bound in the noiseless case, that is, Lemma 5.3 in Keshavan et al. (2010). For readers' convenience, the result is reported in Appendix A, Lemma 7. For the proof of these lemmas, we refer to Keshavan et al. (2010).

In order to prove the lower bound, we start by noticing that

$$F(\mathbf{u}) \le \frac{1}{2}\|\mathcal{P}_E(Z)\|_F^2\,,$$

which is simply proved by using $S = \Sigma$ in Eq. (2). On the other hand, we have

$$\begin{aligned} F(\mathbf{x}) &= \frac{1}{2}\|\mathcal{P}_E(XSY^T - M - Z)\|_F^2 \\ &= \frac{1}{2}\|\mathcal{P}_E(Z)\|_F^2 + \frac{1}{2}\|\mathcal{P}_E(XSY^T - M)\|_F^2 - \langle \mathcal{P}_E(Z), (XSY^T - M)\rangle \qquad (17)\\ &\ge F(\mathbf{u}) + Cn\varepsilon\sqrt{\alpha}\, d_-(\mathbf{x},\mathbf{u})^2 - \sqrt{2r}\|Z^E\|_2\|XSY^T - M\|_F\,, \end{aligned}$$

where in the last step we used Lemma 7. Now by triangular inequality

$$\begin{aligned} \|XSY^T - M\|_F^2 &\le 3\|X(S-\Sigma)Y^T\|_F^2 + 3\|X\Sigma(Y-V)^T\|_F^2 + 3\|(X-U)\Sigma V^T\|_F^2 \\ &\le 3nm\|S-\Sigma\|_F^2 + 3n^2\alpha\Sigma_{\max}^2\Big(\frac{1}{m}\|X-U\|_F^2 + \frac{1}{n}\|Y-V\|_F^2\Big) \\ &\le Cn^2\alpha d_+(\mathbf{x},\mathbf{u})^2\,, \qquad (18) \end{aligned}$$

In order to prove the upper bound, we proceed as above to get

$$F(\mathbf{x}) \le \tfrac{1}{2}\|\mathcal{P}_E(Z)\|_F^2 + Cn\varepsilon\sqrt{\alpha}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2 + \sqrt{2r\alpha}\|Z^E\|_2 Cn d_+(\mathbf{x},\mathbf{u})\,.$$

Further, by replacing $\mathbf{x}$ with $\mathbf{u}$ in Eq. (17)

$$\begin{aligned} F(\mathbf{u}) &\ge \frac{1}{2}\|\mathcal{P}_E(Z)\|_F^2 - \langle \mathcal{P}_E(Z), (U(S-\Sigma)V^T)\rangle \\ &\ge \frac{1}{2}\|\mathcal{P}_E(Z)\|_F^2 - \sqrt{2r\alpha}\|Z^E\|_2 Cn d_+(\mathbf{x},\mathbf{u})\,. \end{aligned}$$

By taking the difference of these inequalities we get the desired upper bound. ■

Proof (Corollary 3.1) By putting together Eq. (8) and (9), and using the definitions of $d_+(\mathbf{x},\mathbf{u})$, $d_-(\mathbf{x},\mathbf{u})$, we get

$$\|S-\Sigma\|_F^2 \le \frac{C_1+C_2}{C_1}\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2 + \frac{(C_1+C_2)\sqrt{r}}{C_1\varepsilon}\|Z^E\|_2\sqrt{\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2 + \|S-\Sigma\|_F^2}\,.$$

Let $x \equiv \|S-\Sigma\|_F$, $a^2 \equiv \big((C_1+C_2)/C_1\big)\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2$, and $b \equiv \big((C_1+C_2)\sqrt{r}/C_1\varepsilon\big)\|Z^E\|_2$. The above inequality then takes the form

$$x^2 \le a^2 + b\sqrt{x^2+a^2} \le a^2 + ab + bx\,,$$

which implies our claim $x \le a+b$.

The singular value bounds (10) and (11) follow by triangular inequality. For instance

$$\sigma_{\min}(S) \geq \Sigma_{\min} - C\Sigma_{\max} d(\mathbf{x},\mathbf{u}) - C\frac{\sqrt{r}}{\varepsilon}\|Z^E\|_2 \,.$$

which implies the inequality (11) for $d(\mathbf{x},\mathbf{u}) \leq \delta = \Sigma_{\min}/C_0\Sigma_{\max}$ and C_0 large enough. An analogous argument proves Eq. (10). ■

3.4 Proof of Lemma 5

Without loss of generality we will assume $\delta \leq 1$, $C_2 \geq 1$ and

$$\frac{\sqrt{r}}{\varepsilon}\|Z^E\|_2 \leq \Sigma_{\min}\,, \tag{19}$$

because otherwise the lower bound (12) is trivial for all $d(\mathbf{x},\mathbf{u}) \leq \delta$.

Denote by $t \mapsto \mathbf{x}(t)$, $t \in [0,1]$, the geodesic on $\mathsf{M}(m,n)$ such that $\mathbf{x}(0) = \mathbf{u}$ and $\mathbf{x}(1) = \mathbf{x}$, parametrized proportionally to the arclength. Let $\widehat{\mathbf{w}} = \dot{\mathbf{x}}(1)$ be its final velocity, with $\widehat{\mathbf{w}} = (\widehat{W},\widehat{Q})$. Obviously $\widehat{\mathbf{w}} \in \mathsf{T}_{\mathbf{x}}$ (with $\mathsf{T}_{\mathbf{x}}$ the tangent space of $\mathsf{M}(m,n)$ at $\mathbf{x}$) and

$$\frac{1}{m}\|\widehat{W}\|^2 + \frac{1}{n}\|\widehat{Q}\|^2 = d(\mathbf{x},\mathbf{u})^2,$$

because $t \mapsto \mathbf{x}(t)$ is parametrized proportionally to the arclength.

Explicit expressions for $\widehat{\mathbf{w}}$ can be obtained in terms of $\mathbf{w} \equiv \dot{\mathbf{x}}(0) = (W,Q)$ (Keshavan et al., 2010). If we let $W = L\Theta R^T$ be the singular value decomposition of W, we obtain

$$\widehat{W} = -UR\Theta\sin\Theta R^T + L\Theta\cos\Theta R^T\,. \tag{20}$$

It was proved in Keshavan et al. (2010) that $\langle \operatorname{grad} G(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq 0$. It is therefore sufficient to lower bound the scalar product $\langle \operatorname{grad} F, \widehat{\mathbf{w}}\rangle$. By computing the gradient of F we get

$$\begin{aligned}
\langle \operatorname{grad} F(\mathbf{x}), \widehat{\mathbf{w}}\rangle &= \langle \mathcal{P}_E(XSY^T - N), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle \\
&= \langle \mathcal{P}_E(XSY^T - M), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle - \langle \mathcal{P}_E(Z), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle \\
&= \langle \operatorname{grad} F_0(\mathbf{x}), \widehat{\mathbf{w}}\rangle - \langle \mathcal{P}_E(Z), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle
\end{aligned} \tag{21}$$

where $F_0(\mathbf{x})$ is the cost function in absence of noise, namely

$$F_0(X,Y) = \min_{S\in\mathbb{R}^{r\times r}} \left\{ \frac{1}{2}\sum_{(i,j)\in E} \left((XSY^T)_{ij} - M_{ij}\right)^2 \right\}. \tag{22}$$

As proved in Keshavan et al. (2010),

$$\langle \operatorname{grad} F_0(\mathbf{x}), \widehat{\mathbf{w}}\rangle \geq Cn\varepsilon\sqrt{\alpha}\Sigma_{\min}^2 d(\mathbf{x},\mathbf{u})^2 \tag{23}$$

(see Lemma 9 in Appendix).

We are therefore left with the task of upper bounding $\langle \mathcal{P}_E(Z), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle$. Since $XS\widehat{Q}^T$ has rank at most r, we have

$$\langle \mathcal{P}_E(Z), XS\widehat{Q}^T\rangle \leq \sqrt{r}\,\|Z^E\|_2\,\|XS\widehat{Q}^T\|_F\,.$$

Since $X^T X = m\mathbf{I}$, we get

$$\begin{aligned} \|XS\widehat{Q}^T\|_F^2 &= m\mathrm{Tr}(S^T S\widehat{Q}^T\widehat{Q}) \le n\alpha\sigma_{\max}(S)^2\|\widehat{Q}\|_F^2 \\ &\le Cn^2\alpha\Big(\Sigma_{\max} + \frac{\sqrt{r}}{\varepsilon}\|Z^E\|_F\Big)^2 d(\mathbf{x},\mathbf{u})^2 \qquad (24) \\ &\le 4Cn^2\alpha\Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2\,, \end{aligned}$$

where, in inequality (24), we used Corollary 3.1 and in the last step, we used Eq. (19). Proceeding analogously for $\langle \mathcal{P}_E(Z), \widehat{W}SY^T\rangle$, we get

$$\langle \mathcal{P}_E(Z), (XS\widehat{Q}^T + \widehat{W}SY^T)\rangle \le C'n\Sigma_{\max}\sqrt{r\alpha}\,\|Z^E\|_2\, d(\mathbf{x},\mathbf{u})\,.$$

Together with Eq. (21) and (23) this implies

$$\langle \mathrm{grad}F(\mathbf{x}), \widehat{\mathbf{w}}\rangle \ge C_1 n\varepsilon\sqrt{\alpha}\Sigma_{\min}^2 d(\mathbf{x},\mathbf{u})\Big\{d(\mathbf{x},\mathbf{u}) - C_2\frac{\sqrt{r}\Sigma_{\max}}{\varepsilon\Sigma_{\min}}\frac{\|Z^E\|_2}{\Sigma_{\min}}\Big\}\,,$$

which implies Eq. (12) by Cauchy-Schwartz inequality.

4. Proof of Theorem 1.3

Proof (*Independent entries model*) We start with a claim that for any sampling set E, we have

$$\|\widetilde{Z}^E\|_2 \le \|Z^E\|_2\,.$$

To prove this claim, let x^* and y^* be m and n dimensional vectors, respectively, achieving the optimum in $\max_{\|x\|\le 1,\|y\|\le 1}\{x^T\widetilde{Z}^E y\}$, that is, such that $\|\widetilde{Z}^E\|_2 = x^{*T}\widetilde{Z}^E y^*$. Recall that, as a result of the trimming step, all the entries in trimmed rows and columns of $\widetilde{Z}^E$ are set to zero. Then, there is no gain in maximizing $x^T\widetilde{Z}^E y$ to have a non-zero entry x_i^* for i corresponding to the rows which are trimmed. Analogously, for j corresponding to the trimmed columns, we can assume without loss of generality that $y_j^* = 0$. From this observation, it follows that $x^{*T}\widetilde{Z}^E y^* = x^{*T}Z^E y^*$, since the trimmed matrix $\widetilde{Z}^E$ and the sample noise matrix Z^E only differ in the trimmed rows and columns. The claim follows from the fact that $x^{*T}Z^E y^* \le \|Z^E\|_2$, for any x^* and y^* with unit norm.

In what follows, we will first prove that $\|Z^E\|_2$ is bounded by the right-hand side of Eq. (4) for any range of $|E|$. Due to the above observation, this implies that $\|\widetilde{Z}^E\|_2$ is also bounded by $C\sigma\sqrt{\varepsilon\sqrt{\alpha}\log n}$, where $\varepsilon \equiv |E|/\sqrt{\alpha}n$. Further, we use the same analysis to prove a tighter bound in Eq. (5) when $|E| \ge n\log n$.

First, we want to show that $\|Z^E\|_2$ is bounded by $C\sigma\sqrt{\varepsilon\sqrt{\alpha}\log n}$, and Z_{ij}'s are i.i.d. random variables with zero mean and sub-Gaussian tail with parameter σ^2. The proof strategy is to show that $\mathbb{E}\big[\|Z^E\|_2\big]$ is bounded, using the result of Seginer (2000) on expected norm of random matrices, and use the fact that $\|\cdot\|_2$ is a Lipschitz continuous function of its arguments together with concentration inequality for Lipschitz functions on i.i.d. Gaussian random variables due to Talagrand (1996).

Note that $\|\cdot\|_2$ is a Lipschitz function with a Lipschitz constant 1. Indeed, for any M and M', $\big|\|M'\|_2 - \|M\|_2\big| \le \|M'-M\|_2 \le \|M'-M\|_F$, where the first inequality follows from triangular inequality and the second inequality follows from the fact that $\|\cdot\|_F^2$ is the sum of the squared singular values.

To bound the probability of large deviation, we use the result on concentration inequality for Lipschitz functions on i.i.d. sub-Gaussian random variables due to Talagrand (1996). For a 1-Lipschitz function $\|\cdot\|_2$ on $m \times n$ i.i.d. random variables Z_{ij}^E with zero mean, and sub-Gaussian tails with parameter σ^2,

$$\mathbb{P}\big(\|Z^E\|_2 - \mathbb{E}[\|Z^E\|_2] > t\big) \leq \exp\Big\{-\frac{t^2}{2\sigma^2}\Big\}. \tag{25}$$

Setting $t = \sqrt{8\sigma^2 \log n}$, this implies that $\|Z^E\|_2 \leq \mathbb{E}\big[\|Z\|_2\big] + \sqrt{8\sigma^2 \log n}$ with probability larger than $1 - 1/n^4$.

Now, we are left to bound the expectation $\mathbb{E}\big[\|Z^E\|_2\big]$. First, we symmetrize the possibly asymmetric random variables Z_{ij}^E to use the result of Seginer (2000) on expected norm of random matrices with symmetric random variables. Let Z'_{ij}'s be independent copies of Z_{ij}'s, and ξ_{ij}'s be independent Bernoulli random variables such that $\xi_{ij} = +1$ with probability $1/2$ and $\xi_{ij} = -1$ with probability $1/2$. Then, by convexity of $\mathbb{E}\big[\|Z^E - Z'^E\|_2 | Z'^E\big]$ and Jensen's inequality,

$$\mathbb{E}\big[\|Z^E\|_2\big] \leq \mathbb{E}\big[\|Z^E - Z'^E\|_2\big] = \mathbb{E}\big[\|(\xi_{ij}(Z_{ij}^E - Z_{ij}'^E))\|_2\big] \leq 2\mathbb{E}\big[\|(\xi_{ij} Z_{ij}^E)\|_2\big],$$

where $(\xi_{ij} Z_{ij}^E)$ denotes an $m \times n$ matrix with entry $\xi_{ij} Z_{ij}^E$ in position (i, j). Thus, it is enough to show that $\mathbb{E}\big[\|Z^E\|_2\big]$ is bounded by $C\sigma\sqrt{\varepsilon\sqrt{\alpha}\log n}$ in the case of symmetric random variables Z_{ij}'s.

To this end, we apply the following bound on expected norm of random matrices with i.i.d. symmetric random entries, proved by Seginer (2000, Theorem 1.1).

$$\mathbb{E}\big[\|Z^E\|_2\big] \leq C\Big(\mathbb{E}\big[\max_{i\in[m]} \|Z_{i\bullet}^E\|\big] + \mathbb{E}\big[\max_{j\in[n]} \|Z_{\bullet j}^E\|\big]\Big), \tag{26}$$

where $Z_{i\bullet}^E$ and $Z_{\bullet j}^E$ denote the ith row and jth column of A respectively. For any positive parameter β, which will be specified later, the following is true.

$$\mathbb{E}\big[\max_j \|Z_{\bullet j}^E\|^2\big] \leq \beta\sigma^2\varepsilon\sqrt{\alpha} + \int_0^\infty \mathbb{P}\big(\max_j \|Z_{\bullet j}^E\|^2 \geq \beta\sigma^2\varepsilon\sqrt{\alpha} + z\big)\,dz. \tag{27}$$

To bound the second term, we can apply union bound on each of the n columns, and use the following bound on each column $\|Z_{\bullet j}^E\|^2$ resulting from concentration of measure inequality for the i.i.d. sub-Gaussian random matrix Z.

$$\mathbb{P}\Big(\sum_{k=1}^m (Z_{kj}^E)^2 \geq \beta\sigma^2\varepsilon\sqrt{\alpha} + z\Big) \leq \exp\Big\{-\frac{3}{8}\Big((\beta-3)\varepsilon\sqrt{\alpha} + \frac{z}{\sigma^2}\Big)\Big\}. \tag{28}$$

To prove the above result, we apply Chernoff bound on the sum of independent random variables. Recall that $Z_{kj}^E = \tilde{\xi}_{kj} Z_{kj}$ where $\tilde{\xi}$'s are independent Bernoulli random variables such that $\tilde{\xi} = 1$ with probability $\varepsilon/\sqrt{mn}$ and zero with probability $1 - \varepsilon/\sqrt{mn}$. Then, for the choice of $\lambda = 3/8\sigma^2 < 1/2\sigma^2$,

$$\begin{aligned}
\mathbb{E}\Big[\exp\Big(\lambda \sum_{k=1}^m (\tilde{\xi}_{kj} Z_{kj})^2\Big)\Big] &= \Big(1 - \frac{\varepsilon}{\sqrt{mn}} + \frac{\varepsilon}{\sqrt{mn}}\mathbb{E}[e^{\lambda Z_{kj}^2}]\Big)^m \\
&\leq \Big(1 - \frac{\varepsilon}{\sqrt{mn}} + \frac{\varepsilon}{\sqrt{mn(1-2\sigma^2\lambda)}}\Big)^m \\
&= \exp\Big\{m\log\Big(1 + \frac{\varepsilon}{\sqrt{mn}}\Big)\Big\} \\
&\leq \exp\big\{\varepsilon\sqrt{\alpha}\big\},
\end{aligned}$$

where the first inequality follows from the definition of Z_{kj} as a zero mean random variable with sub-Gaussian tail, and the second inequality follows from $\log(1+x) \le x$. By applying Chernoff bound, Eq. (28) follows. Note that an analogous result holds for the Euclidean norm on the rows $\|Z^E_{i\bullet}\|^2$.

Substituting Eq. (28) and $\mathbb{P}\big(\max_j \|Z^E_{\bullet j}\|^2 \ge z\big) \le m\,\mathbb{P}\big(\|Z^E_{\bullet j}\|^2 \ge z\big)$ in Eq. (27), we get

$$\mathbb{E}\big[\max_j \|Z^E_{\bullet j}\|^2\big] \le \beta\sigma^2\varepsilon\sqrt{\alpha} + \frac{8\sigma^2 m}{3} e^{-\frac{3}{8}(\beta-3)\varepsilon\sqrt{\alpha}}\,. \tag{29}$$

The second term can be made arbitrarily small by taking $\beta = C\log n$ with large enough C. Since $\mathbb{E}\big[\max_j \|Z^E_{\bullet j}\|\big] \le \sqrt{\mathbb{E}\big[\max_j \|Z^E_{\bullet j}\|^2\big]}$, applying Eq. (29) with $\beta = C\log n$ in Eq. (26) gives

$$\mathbb{E}\big[\|Z^E\|_2\big] \le C\sigma\sqrt{\varepsilon\sqrt{\alpha}\log n}\,.$$

Together with Eq. (25), this proves the desired thesis for any sample size $|E|$.

In the case when $|E| \ge n\log n$, we can get a tighter bound by similar analysis. Since $\varepsilon \ge C'\log n$, for some constant C', the second term in Eq. (29) can be made arbitrarily small with a large constant β. Hence, applying Eq. (29) with $\beta = C$ in Eq. (26), we get

$$\mathbb{E}\big[\|Z^E\|_2\big] \le C\sigma\sqrt{\varepsilon\sqrt{\alpha}}\,.$$

Together with Eq. (25), this proves the desired thesis for $|E| \ge n\log n$.

■

Proof (*Worst Case Model*) Let D be the $m \times n$ all-ones matrix. Then for any matrix Z from the *worst case model*, we have $\|\widetilde{Z}^E\|_2 \le Z_{\max}\|\widetilde{D}^E\|_2$, since $x^T\widetilde{Z}^E y \le \sum_{i,j} Z_{\max}|x_i|\widetilde{D}^E_{ij}|y_j|$, which follows from the fact that Z_{ij}'s are uniformly bounded. Further, $\widetilde{D}^E$ is an adjacency matrix of a corresponding bipartite graph with bounded degrees. Then, for any choice of E the following is true for all positive integers k:

$$\|\widetilde{D}^E\|_2^{2k} \le \max_{x,\|x\|=1} \big|x^T((\widetilde{D}^E)^T\widetilde{D}^E)^k x\big| \le \mathrm{Tr}\big(((\widetilde{D}^E)^T\widetilde{D}^E)^k\big) \le n(2\varepsilon)^{2k}\,.$$

Now $\mathrm{Tr}\big(((\widetilde{D}^E)^T\widetilde{D}^E)^k\big)$ is the number of paths of length $2k$ on the bipartite graph with adjacency matrix $\widetilde{D}^E$, that begin and end at i for every $i \in [n]$. Since this graph has degree bounded by 2ε, we get

$$\|\widetilde{D}^E\|_2^{2k} \le n(2\varepsilon)^{2k}\,.$$

Taking k large, we get the desired thesis. ■

Acknowledgments

This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211. SO was supported by a fellowship from the Samsung Scholarship Foundation.

Appendix A. Three Lemmas on the Noiseless Problem

Lemma 7 *There exists numerical constants C_0, C_1, C_2 such that the following happens. Assume $\varepsilon \geq C_0 \mu_0 r \sqrt{\alpha} \max\{\log n; \mu_0 r \sqrt{\alpha} (\Sigma_{\min}/\Sigma_{\max})^4\}$ and $\delta \leq \Sigma_{\min}/(C_0 \Sigma_{\max})$. Then,*

$$C_1 \sqrt{\alpha} \Sigma_{\min}^2 d(\mathbf{x},\mathbf{u})^2 + C_1 \sqrt{\alpha} \|S_0 - \Sigma\|_F^2 \leq \frac{1}{n\varepsilon} F_0(\mathbf{x}) \leq C_2 \sqrt{\alpha} \Sigma_{\max}^2 d(\mathbf{x},\mathbf{u})^2 ,$$

for all $\mathbf{x} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ such that $d(\mathbf{x},\mathbf{u}) \leq \delta$, with probability at least $1 - 1/n^4$. Here $S_0 \in \mathbb{R}^{r \times r}$ is the matrix realizing the minimum in Eq. (22).

Lemma 8 *There exists numerical constants C_0 and C such that the following happens. Assume $\varepsilon \geq C_0 \mu_0 r \sqrt{\alpha} (\Sigma_{\max}/\Sigma_{\min})^2 \max\{\log n; \mu_0 r \sqrt{\alpha} (\Sigma_{\max}/\Sigma_{\min})^4\}$ and $\delta \leq \Sigma_{min}/(C_0 \Sigma_{\max})$. Then*

$$\|\operatorname{grad} \widetilde{F}_0(\mathbf{x})\|^2 \geq C n \varepsilon^2 \Sigma_{\min}^4 d(\mathbf{x},\mathbf{u})^2 ,$$

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

Lemma 9 *Define $\widehat{\mathbf{w}}$ as in Eq. (20). Then there exists numerical constants C_0 and C such that the following happens. Under the hypothesis of Lemma 8*

$$\langle \operatorname{grad} F_0(\mathbf{x}), \widehat{\mathbf{w}} \rangle \geq C n \varepsilon \sqrt{\alpha} \Sigma_{\min}^2 d(\mathbf{x},\mathbf{u})^2 ,$$

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

References

P.-A. Absil, R. Mahony, and R. Sepulchrer. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54 (2):9, 2007.

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

E. J. Candès and Y. Plan. Matrix completion with noise. `arXiv:0903.3131`, 2009.

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

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

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

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004. ISSN 0004-5411.

R. H. Keshavan and S. Oh. Optspace: A gradient descent algorithm on the grassman manifold for matrix completion. arXiv:0910.5260, 2009.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, June 2010.

K. Lee and Y. Bresler. Admira: Atomic decomposition for minimum rank approximation. arXiv:0905.0044, 2009.

S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. arXiv:0905.1643, 2009.

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

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. arXiv:1002.2780, 2010.

R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the International Conference on Machine Learning*, volume 24, pages 791–798, 2007.

Y. Seginer. The expected norm of random matrices. *Comb. Probab. Comput.*, 9:149–166, March 2000. ISSN 0963-5483. doi: 10.1017/S096354830000420X. URL http://portal.acm.org/citation.cfm?id=971471.971475.

N. Srebro and T. S. Jaakkola. Weighted low-rank approximations. In *In 20th International Conference on Machine Learning*, pages 720–727. AAAI Press, 2003.

N. Srebro, J. D. M. Rennie, and T. S. Jaakola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, 2005.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1–34, 1996. ISSN 00911798. URL http://www.jstor.org/stable/2244830.

K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. http://www.math.nus.edu.sg/∼matys, 2009.