Deterministic Algorithms for Matrix Completion

Eyal Heiman,[1] Gideon Schechtman,[2] Adi Shraibman[3]
[1] Department of Computer Science, Hebrew University, Jerusalam, Israel
[2] Department of Mathematics, Weizmann Institute of Science, Rehovot, Israel
[3] Department of Computer Science, Tel Aviv-Yaffo Academic College, Tel Aviv, Israel

Received 25 September 2011; accepted 12 March 2012; received in final form 9 July 2012
Published online in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/rsa.20483

ABSTRACT: The goal of the *matrix completion problem* is to retrieve an unknown real matrix from a small subset of its entries. This problem comes up in many application areas, and has received a great deal of attention in the context of the *Netflix challenge*. This setup usually represents our partial knowledge of some information domain. Unknown entries may be due to the unavailability of some relevant experimental data.

One approach to this problem starts by selecting a complexity measure of matrices, such as rank or trace norm. The corresponding algorithm outputs a matrix of lowest possible complexity that agrees with the partially specified matrix. The performance of the above algorithm under the assumption that the revealed entries are sampled randomly has received considerable attention (e.g., Refs. Srebro et al., 2005; COLT, 2005; Foygel and Srebro, 2011; Candes and Tao, 2010; Recht, 2009; Keshavan et al., 2010; Koltchinskii et al., 2010). Here we ask what can be said if the observed entries are chosen deterministically. We prove generalization error bounds for such deterministic algorithms, that resemble the results of Refs. Srebro et al. (2005); COLT (2005); Foygel and Srebro (2011) for the randomized algorithms.

We still do not understand which sets of entries in a given matrix can be used to properly reconstruct it. Our hope is that the present work sheds some light on this problem. © 2013 Wiley Periodicals, Inc. Random Struct. Alg., 00, 000–000, 2013

Keywords: matrix completion; expander graphs; graph sparsifiers; factorization norms; deterministic guarantees

Correspondence to: Adi Shraibman

1. INTRODUCTION

Consider the problem of approximating a partially observed target matrix Y with another matrix X. This problem, known as the *matrix completion problem*, arises often in practice. A well known instance of this problem is the famous *Netflix challenge* in which we seek to predict people's preferences in films based on their past choices in viewing films. Say y_{ij} is the score given by viewer i to film j. These numbers can be considered as a very sparse sample of the matrix Y which we seek to reconstruct.

More formally, we have *oracle access* to a real $n \times n$ matrix $Y = (y_{ij})$. Namely, given a pair of indices (i,j) the oracle returns y_{ij}. We consider an algorithm that is given such access to Y and an error parameter ϵ. The algorithm should use a small number of calls to the oracle and return a real matrix X such that $\sum_{i,j}(x_{ij} - y_{ij})^2 \leq \epsilon$. Clearly, the number of oracle calls that we require depends on the properties of Y. Without any assumption (or restriction) on Y, the above question is meaningless.

A common general scheme for solving such problems is to select a matrix X that minimizes some combination of the *complexity* of X and the *distance* between X and Y on the observed part. In particular, one can insist that X agrees with Y on the queried entries. This general scheme follows the principle of Occam's razor, namely that the "simplest" solution yields the best performance on new instances. The heart of the matter is therefore our interpretation of "simplicity," namely our choice of the complexity measure for X.

The most commonly used notion of complexity in such tasks is matrix rank. For example, it is not hard to see why small rank makes sense in the Netflix example. It stands to reason that users' preferences depend on a small set of parameters. More recently, the trace-norm and γ_2 were suggested as alternative measures of complexity [5, 16]. Whereas the search for minimal rank usually results in NP-hard problems, the problems of minimizing the trace-norm and γ_2 can be solved in polynomial time using *convex programming*. Figure 1 describes the outline of the algorithm with γ_2 as the complexity measure.

The γ_2 norm originated in the study of factorization norms in Banach space theory, and is defined for a real matrix X as:

$$\gamma_2(X) = \min_{UV=X} \|U\|_{\ell_2 \to \ell_\infty^m} \|V\|_{\ell_1^n \to \ell_2}.$$

For a more detailed expository of the γ_2 norm, see Section 2.2.

The γ_2 norm was first utilized in the context of matrix completion by Srebro et al. [16][1]. They analyzed the algorithm of Figure (1) when the initial set S is chosen at random, and proved the following bound on the generalization error [2]:

Theorem 1 ([16]). *Let Y be an $n \times n$ real matrix, $\delta > 0$, and P a probability distribution on pairs $(i,j) \in [n]^2$. Choose a sample S of $|S| > n \log n$ entries according to P. Then, with probability at least $1 - \delta$ over the sample selection, the following holds:*

$$\sum_{i,j} p_{ij}|x_{ij} - y_{ij}| \leq c\gamma_2(X)\sqrt{\frac{n - \log \delta}{|S|}}.$$

Where X is the output of the algorithm with sample S, and c is a universal constant.

[1] Note that in Refs. [6, 16, 17] the γ_2 norm is referred to as "the max norm."

[2] Their result is more general than stated here and applies to every Lipschitz loss function. We opted for this simplified statement for ease of comparison. For most purposes this simplified version is not less powerful.

```
1.  Choose a subset $S \subset [n]^2$ and query the oracle for the value of Y, on $S$.
2.  Return a matrix $X$ of smallest possible $\gamma_2(X)$ under the condition that
    $x_{ij} = y_{ij}$ for all $(i,j) \in S$.
```

Fig. 1. The basic scheme.

The statement and proof of Theorem 1 in Ref. [16] only deal with the case of sampling from the uniform distribution. The general statement is from [17].

Papers studying the performance of the algorithm of Fig. 1 can be divided in two categories. The first family of papers (e.g., [3,4,9,15]) study conditions under which this simple approach for matrix completion retrieves the underlying matrix, exactly. In this family of papers the trace norm is used as a complexity measure. It is an interesting open problem whether this kind of result holds also for γ_2.

In the second family of papers no conditions are posed on the matrix, but the output of the algorithm is an approximation of the underlying matrix. Upper bounds on the degree of approximation are proved, as in Theorem 1. Our work continue this line of results, in particular that of [16,17] and related papers. In these papers the sample is chosen at random, but in practice we are typically limited in our choice of sampled entries. As suggested above, it may require some experimental work to reveal an entry of Y, and some entries are harder to determine than others. It is therefore of practical interest to have a good estimate for the level of approximation that a given set of samples is guaranteed to yield. This issue is the motivating force of our study.

In addition, studying deterministic versions of randomized algorithms usually shed new light on the underlying structure, especially when explicit constructions are involved. We consider the following (deterministic) choice of the initial set S that is specified in terms of an expander graph G. We examine an entry (i,j) iff it is an edge in G. We prove the following bound on the generalization error of our basic algorithm in this case.

Theorem 2. *Let S be the set of edges of a d-regular graph with second eigenvalue [3] bound λ. For every $n \times n$ real matrix Y, if X is the output of our algorithm with initial subset S, then*

$$\frac{1}{n^2}\sum_{i,j}(x_{ij} - y_{ij})^2 \leq c\gamma_2(Y)^2\frac{\lambda}{d},$$

where c is a small universal constant.

It is known that λ can be made as small as $O(\sqrt{d})$ (e.g., a Ramanujan graph). In this case Theorem 2 yields an error bound of

$$\begin{aligned}\frac{1}{n^2}\sum_{i,j}(x_{ij} - y_{ij})^2 &\leq c'\gamma_2(Y)^2\frac{1}{\sqrt{d}}\\ &= c'\gamma_2(Y)^2\left(\frac{n}{|S|}\right)^{1/2}\end{aligned}$$

[3] The eigenvalues are eigenvalues of the adjacency matrix of the graph.

We recall that d-regular graphs with $\lambda = O(\sqrt{d})$ can be constructed in linear time using e.g. the well-known LPS Ramanujan graphs [13].

Our generalization error bounds are not as strong as the bounds proved for randomized sampling [16] and we believe that better bounds can be proved using only the properties of expander graphs. Namely we suggest the following conjecture:

Conjecture 3. *Let S be the set of edges of a d-regular graph with $\lambda = O(\sqrt{d})$. For every $n \times n$ real matrix Y, if X the output of our algorithm when S is picked in the first step, then*

$$\frac{1}{n^2}\sum_{i,j}|x_{ij} - y_{ij}| \le c\gamma_2(Y)\frac{\lambda}{d},$$

for some constant c.

1.1. Non-Uniform Weights

As mentioned, Theorem 1 holds when the initial sample is drawn from any probability distribution. Our construction, based on expander graphs, is good only when entries are chosen uniformly. Obviously, we cannot expect a deterministic construction to work well with an arbitrary distribution. And indeed, expander graphs need not yield good samples for matrix completion against non-uniform distributions. Nevertheless, for any probability distribution, we provide explicit constructions of an initial sample, that work well against that probability distribution. These explicit constructions are based on other graph sparsifiers, such as the ones given by Refs. [1,2]. Formally:

Theorem 4. *Let P be a probability distribution on pairs $(i,j) \in [n]^2$, and $d > 1$. There is an efficiently constructed set $S \subset [n]^2$ of size at most dn, such that for every $n \times n$ real target matrix Y, if X is the output of our algorithm with initial subset S, then*

$$\sum_{i,j} p_{ij}(x_{ij} - y_{ij})^2 \le c\gamma_2(Y)^2\frac{1}{\sqrt{d}}.$$

The efficiency of constructing the initial set of queries, and the generalization bounds in Theorem 4 depend on the notion of graph sparsifiers (Section 4). We require cut preserving or quadratic form preserving sparsifiers. There are several possible sparsifiers which may be used and different applications call for different choices. For example, the above statement assumes the sparsifiers of Ref. [1] which provide very good guarantees, but are not extremely efficient. For better efficiency but slightly worse guarantees the sparsifiers of Ref. [2] can be invoked.

2. BACKGROUND

2.1. Expander Graphs

Let $G = (V,E)$ be a d-regular graph on n vertices, and $d = \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n$ its eigenvalues (i.e., eigenvalues of the adjacency matrix of G). We denote the second eigenvalue bound $\lambda = \lambda(G) = \max_{2\le i\le n}|\lambda_i|$. We often use a shorthand and say that G is a (d,λ) *graph* in this case. As usual, a family of d-regular graphs $\{G_t\}$ on $n_t \to \infty$ vertices is called a family of *expander graphs* if their *spectral gap* $d - \lambda$ is bounded from below when $t \to \infty$. It is known that $\lambda \geq 2\sqrt{d-1} - o(1)$. When $\lambda \leq 2\sqrt{d-1}$ we say that G is a *Ramanujan* graph.

One important property of expander graphs is the *expander mixing lemma*, which states that the number of edges between every two sets of vertices in an expander graph is close to what we expect in a random graph. More formally, for every $A, B \subset V$

$$\left| \frac{|A||B|}{n^2} - \frac{E(A,B)}{|E|} \right| \leq \frac{\lambda}{d} \sqrt{\frac{|A||B|}{n^2}}$$

Here $E(A,B)$ is the number of edges between A and B. For more details on expander graphs see Ref. [7].

2.2. γ_2 and Grothendieck's Inequality

As mentioned in the introduction, the γ_2 norm originated in the study of factorization norms in Banach space theory. The γ_2 norm of a real matrix X is defined as follows:

$$\gamma_2(X) = \min_{UV=X} \|U\|_{\ell_2 \to \ell_\infty^m} \|V\|_{\ell_1^n \to \ell_2}. \tag{1}$$

Here are a few comments that may add some insight regarding the intuition underlying this definition. Recall the following definition for rank of a real matrix X

$$rank(X) = \min_{UV=X} \sqrt{dim_R(U)} \cdot \sqrt{dim_C(V)},$$

where $dim_R(U)$ is the number of columns in U (i.e., the dimension of the row space of U). Similarly $dim_C(V)$ is the dimension of V's column space.

We can informally describe the γ_2 norm as a semi-definite relaxation of matrix rank. To see this, note the following two simple facts: the operator norm $\|V\|_{\ell_1^n \to \ell_2}$ is the largest ℓ_2 norm of a column of V. Likewise, $\|U\|_{\ell_2 \to \ell_\infty^m}$ is the largest ℓ_2 norm of a row of U. Thus γ_2 is defined by modifying the definition of rank, where length substitutes dimension. One useful feature of this definition is that γ_2 can be viewed as the optimum of an optimization problem that is solvable by semi-definite programming. Specifically, variations on this definition that involve various linear restrictions can (unlike matrix rank) be still conveniently characterized and efficiently computed.

The relation between γ_2 and rank is also expressed in the following inequality: for every real matrix X it holds

$$\gamma_2(X) \leq \sqrt{rank(X)} \|X\|_\infty$$

This inequality is tight, e.g., for a Hadamard matrix. Also it is tight up to a constant for a random $n \times n$ sign matrix.

There is no matching lower bound though. Consider the $n \times n$ identity matrix I_n, then $\gamma_2(I_n) = 1$ while $rank(I_n) = n$. If we allow some slackness and ask for the minimal rank (res. γ_2) in an ℓ_∞ environment then the two notions do become strongly related [11].

The γ_2 norm has many other interesting properties of which we mention Grothendieck's inequality (e.g., Refs. [14, pg. 64] and [18]).

Theorem 5 (Grothendieck's inequality). *There is a universal constant* $1.5 \leq K_G \leq 1.8$ *such that for every real* $m \times n$ *matrix* X

$$\max \sum_{ij} x_{ij} \langle u_i, v_j \rangle \leq K_G \max \sum_{ij} x_{ij} \epsilon_i \delta_j. \tag{2}$$

Here $u_1, \ldots, u_m, v_1, \ldots, v_n$ *are arbitrary unit vectors in* ℓ_2 *and* $\epsilon_1, \ldots, \epsilon_m, \delta_1, \ldots, \delta_n$ *take values in* $\{\pm 1\}$.

Grothendieck's inequality can be stated in terms of γ_2 and the nuclear norm.

Definition 6 (Nuclear norm). *Let* X *be a real matrix. The* (ℓ_1 *to* ℓ_∞) *nuclear norm is defined as*

$$\nu(X) = \min_{\alpha_i \in \mathbb{R}} \left\{ \sum_i |\alpha_i| : X = \sum_i \alpha_i \epsilon_i \delta_i^t, \textit{ for sign vectors } \epsilon_i, \delta_i \right\}$$

Grothendieck's inequality states that the dual norms of γ_2 and ν are equivalent up to a small constant, K_G. Alternatively:

Theorem 7. *For every real matrix* X*:*

$$\gamma_2(X) \leq \nu(X) \leq K_G \gamma_2(X).$$

Nuclear norms are dual to operator norms. Specifically, ν is the *nuclear norm* from ℓ_1 to ℓ_∞ [8]. Observe that the unit ball of ν is the convex polytope whose vertices are rank one sign matrices. Thus, Grothendieck's inequality says that the unit ball of γ_2 coincides, up to a factor of K_G, with the convex hull of rank one sign matrices.

3. PROOF OF THEOREM 2

We start by proving the following theorem, which might be of independent interest. It says that the average of all entries of a matrix is approximated by the average over the edges of an expander graph. The degree of approximation depends on the nuclear (equivalently γ_2) norm of the matrix.

Theorem 8. *For every real* $n \times n$ *matrix* R*, and* (d, λ) *graph* $G = (V, E)$

$$\left| \frac{1}{n^2} \sum_{i,j} r_{ij} - \frac{1}{|E|} \sum_{(i,j) \in E} r_{ij} \right| \leq 2\nu(R) \frac{\lambda}{d}.$$

By Grothendieck's inequality (Theorem 7), this implies:

Corollary 9. *For every real* $n \times n$ *matrix* R*, and* (d, λ) *graph* $G = (V, E)$

$$\left| \frac{1}{n^2} \sum_{i,j} r_{ij} - \frac{1}{|E|} \sum_{(i,j) \in E} r_{ij} \right| \leq 2K_G \gamma_2(R) \frac{\lambda}{d}$$

where K_G *is the Grothendieck's constant.*

Proof. First we prove the theorem for a rank-1 sign-matrix S. For every sign-matrix S we define the corresponding $(0,1)$-matrix $S' = \frac{1}{2}(S+J)$, where J is the all-ones matrix. Clearly $S' = 1_A \times 1_B + 1_{A^c} \times 1_{B^c}$ for some subsets $A, B \subset V = [n]$, where 1_Z is the characteristic vector of the set Z. We rewrite the error expression in these terms

$$\begin{aligned}\left|\frac{1}{n^2}\sum_{i,j} s_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} s_{ij}\right| &= \left|\frac{1}{n^2}\sum_{i,j}(2s'_{ij}-1) - \frac{1}{|E|}\sum_{(i,j)\in E}(2s'_{ij}-1)\right| \\ &= 2\left|\frac{1}{n^2}\sum_{i,j} s'_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} s'_{ij}\right| \\ &= 2\left|\frac{|A||B| + |A^c||B^c|}{n^2} - \frac{E(A,B)+E(A^c,B^c)}{|E|}\right| \\ &\leq 2\left|\frac{|A||B|}{n^2} - \frac{E(A,B)}{|E|}\right| + 2\left|\frac{|A^c||B^c|}{n^2} - \frac{E(A^c,B^c)}{|E|}\right|\end{aligned}$$

By applying the expander mixing lemma we get

$$\begin{aligned}\left|\frac{1}{n^2}\sum_{i,j} s_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} s_{ij}\right| &\leq 2\left|\frac{|A||B|}{n^2} - \frac{E(A,B)}{|E|}\right| + 2\left|\frac{|A^c||B^c|}{n^2} - \frac{E(A^c,B^c)}{|E|}\right| \\ &\leq \frac{2\lambda}{d}\left(\sqrt{\frac{|A||B|}{n^2}} + \sqrt{\frac{|A^c||B^c|}{n^2}}\right) \\ &\leq \frac{2\lambda}{d}.\end{aligned}$$

In the last inequality we use the fact that $f(x,y) = \sqrt{xy} + \sqrt{(1-x)(1-y)} \leq 1$ for $0 \leq x, y \leq 1$ with equality when $x = y$.

In the general case we represent a real matrix R as a linear combination of rank-1 sign matrices $R = \sum_k \alpha_k S^k$, with $\nu(R) = \sum_k |\alpha_k|$. This yields

$$\begin{aligned}\left|\frac{1}{n^2}\sum_{i,j} r_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} r_{ij}\right| &= \left|\sum_k \alpha_k\left(\frac{1}{n^2}\sum_{i,j} s^k_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} s^k_{ij}\right)\right| \\ &\leq \sum_k |\alpha_k|\left|\frac{1}{n^2}\sum_{i,j} s^k_{ij} - \frac{1}{|E|}\sum_{(i,j)\in E} s^k_{ij}\right| \\ &\leq 2\sum_k |\alpha_k|\frac{\lambda}{d} \\ &= 2\nu(R)\frac{\lambda}{d}\end{aligned}$$

■

Consider the matrix $R = (X - Y) \circ (X - Y)$, where $\circ$ is the Hadamard (or entry-wise) product, namely $r_{ij} = (x_{ij} - y_{ij})^2$. Theorem 2 follows by applying Corollary 9 to this matrix:

$$\left| \frac{1}{n^2} \sum_{i,j} (x_{ij} - y_{ij})^2 - \frac{1}{|E|} \sum_{(i,j)\in E} (x_{ij} - y_{ij})^2 \right| \leq 2K_g \gamma_2(R) \frac{\lambda}{d}.$$

But γ_2 is multiplicative under Hadamard product [12], so that

$$\gamma_2(R) \leq \gamma_2(X - Y)^2 \leq (\gamma_2(X) + \gamma_2(Y))^2.$$

For the last inequality recall that γ_2 is a norm. Since the matrix X is the output of our algorithm we have that $\gamma_2(X) \leq \gamma_2(Y)$. Therefore

$$\gamma_2(R) \leq 4\gamma_2(Y)^2.$$

We conclude that

$$\left| \frac{1}{n^2} \sum_{i,j} (x_{ij} - y_{ij})^2 - \frac{1}{|E|} \sum_{(i,j)\in E} (x_{ij} - y_{ij})^2 \right| \leq 8K_g \gamma_2(Y)^2 \frac{\lambda}{d}.$$

The theorem now follows, since our algorithm satisfies (see Fig. 1)

$$\frac{1}{|E|} \sum_{(i,j)\in E} (x_{ij} - y_{ij})^2 = 0.$$

■

4. PROOF OF THEOREM 4

So far we considered a sample S which is the edge set of a d-regular expander G. We derived for this case an upper bound on the generalization error with respect to the uniform distribution in terms of $\gamma_2(Y)$, d and $\lambda(G)$. The proof is based on Theorem 8, which uses properties of expander graphs. A good sample w.r.t. non-uniform distributions requires slightly different graphs, called sparsifiers.

A *sparsifier* of a graph $G = (V, E, w)$ is a sparse graph H that is similar to G in some useful manner. For example, expander graphs are sparsifiers of the complete graph. They are similar to the complete graph in the fraction of edges they contain in every cut. Or, as we saw, they are also similar in estimating the average over the entries of a matrix with low γ_2 norm.

Batson et al. [1] consider a spectral notion of similarity. They prove

Theorem 10 ([1]). *For every $d > 1$, every undirected weighted graph $G = (V, E, w)$ on n vertices contains a weighted subgraph $H = (V, F, \tilde{w})$ with $d(n-1)$ edges that satisfies:*

$$\sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2 \leq \sum_{(i,j)\in F} \tilde{w}_{ij}(x_i - x_j)^2 \leq \frac{d+1+2\sqrt{d}}{d+1-2\sqrt{d}} \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2, \qquad (3)$$

for every vector of real numbers $(x_1, x_2, \ldots, x_n)$.

Notice that Eq. (3) implies

$$\left|\sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2 - \sum_{(i,j)\in F} \tilde{w}_{ij}(x_i - x_j)^2\right| \le \frac{4\sqrt{d}}{d+1-2\sqrt{d}} \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2$$
$$= \Theta(\frac{1}{\sqrt{d}}) \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2.$$

We use the sparsifiers of Batson et al. to query a matrix that we wish to complete. Instead of Theorem 8 we use:

Theorem 11. *Let P be a probability distribution on pairs $(i,j) \in [n]^2$, and $d > 1$. There is an efficiently constructed set $S \subset [n]^2$ of cardinality at most dn, and a weight function $w : S \to \mathbb{R}^+$, such that for every $n \times n$ real matrix R:*

$$\left|\sum_{i,j} p_{ij} r_{ij} - \sum_{(i,j)\in S} w_{ij} r_{ij}\right| \le O\left(\frac{\nu(R)}{\sqrt{d}}\right).$$

Like before, Grothendieck's inequality implies the corollary:

Corollary 12. *Let P be a probability distribution on pairs $(i,j) \in [n]^2$, and $d > 1$. There is an efficiently constructed set $S \subset [n]^2$ of size at most dn, and a weight function $w : S \to \mathbb{R}^+$, such that for every $n \times n$ real matrix R:*

$$\left|\sum_{i,j} p_{ij} r_{ij} - \sum_{(i,j)\in S} w_{ij} r_{ij}\right| \le O\left(\frac{\gamma_2(R)}{\sqrt{d}}\right).$$

4.1. Constructing the Sample

Before we prove Theorem 11 let us describe the construction of the initial sample: let P be a probability distribution on pairs $(i,j) \in [n]^2$, and $d > 1$. Let $V = [2n]$, and $G = (V, E, P)$ be the complete bipartite graph having n vertices in each side, with weights given by P. That is, p_{ij} is the weight assigned to the edge (i,j). The left[right] set of vertices of G correspond to the rows[columns] of the matrix that we want to recover, respectively. Let $H = (V, F, w)$ be the subgraph of G guaranteed by Theorem 10. Then, the sample is taken as $S = F$, the set of edges of H. By Theorem 10

$$\left|\sum_{i,j} p_{ij}(x_i - y_j)^2 - \sum_{(i,j)\in S} w_{ij}(x_i - y_j)^2\right| = \Theta(\frac{1}{\sqrt{d}}) \sum_{i,j} p_{ij}(x_i - y_j)^2, \tag{4}$$

for every vector of real numbers $(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$.

Remark 13. *In the construction of S, we have considered G as the complete bipartite graph. We can instead take the bipartite graph with n vertices in each side, and edge set that corresponds to the support of P. This way we avoid any dependence on entries (i,j) for which $p_{ij} = 0$.*

Proof of Theorem 11. As in the proof of Theorem 8, it is enough to prove the theorem for a rank-1 sign matrix xy^t. The general theorem then follows from basic properties of the nuclear norm.

Thus, let xy^t be a rank-1 sign matrix. By plugging the vector $(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$ in Eq. (4), we get

$$\left|\sum_{i,j} p_{ij}(x_i - y_j)^2 - \sum_{(i,j)\in S} w_{ij}(x_i - y_j)^2\right| = \Theta(\frac{1}{\sqrt{d}})\sum_{i,j} p_{ij}(x_i - y_j)^2, \tag{5}$$

By plugging the vector $(x_1, x_2, \ldots, x_n, -y_1, -y_2, \ldots, -y_n)$ again in Eq. (4), we get

$$\left|\sum_{i,j} p_{ij}(x_i + y_j)^2 - \sum_{(i,j)\in S} w_{ij}(x_i + y_j)^2\right| = \Theta(\frac{1}{\sqrt{d}})\sum_{i,j} p_{ij}(x_i + y_j)^2, \tag{6}$$

Notice that

$$|x_i - y_j| = \begin{cases} 0 & \textit{if} \;\; x_i = y_j \\ -2x_iy_j & \textit{if} \;\; x_i \neq y_j \end{cases}$$

and

$$|x_i + y_j| = \begin{cases} 2x_iy_j & \textit{if} \;\; x_i = y_j \\ 0 & \textit{if} \;\; x_i \neq y_j \end{cases}$$

Therefore

$$\begin{aligned}
\left|\sum_{i,j} p_{ij}x_iy_j - \sum_{(i,j)\in S} w_{ij}x_iy_j\right| &\leq \left|\sum_{i,j:x_i=y_j} p_{ij}x_iy_j - \sum_{(i,j)\in S:x_i=y_j} w_{ij}x_iy_j\right| \\
&\quad + \left|\sum_{i,j:x_i\neq y_j} p_{ij}x_iy_j - \sum_{(i,j)\in S:x_i\neq y_j} w_{ij}x_iy_j\right| \\
&= \frac{1}{4}\left|\sum_{i,j} p_{ij}(x_i - y_j)^2 - \sum_{(i,j)\in S} w_{ij}(x_i - y_j)^2\right| \\
&\quad + \frac{1}{4}\left|\sum_{i,j} p_{ij}(x_i + y_j)^2 - \sum_{(i,j)\in S} w_{ij}(x_i + y_j)^2\right| \\
&\leq \Theta(\frac{1}{\sqrt{d}})\sum_{i,j} p_{ij}(x_i - y_j)^2 \\
&\quad + \Theta(\frac{1}{\sqrt{d}})\sum_{i,j} p_{ij}(x_i + y_j)^2 \\
&\leq \Theta(\frac{1}{\sqrt{d}}).
\end{aligned}$$

For the last inequality recall that P is a probability distribution. Also, $x_i - y_j$ and $x_i + y_j$ have disjoint support and are at most two in absolute value.

This completes the proof for Rank 1 sign matrices. The case of a general real matrix is now proved as in Theorem 8. ▪

The last step of the proof is similar to the proof of Theorem 2, with Theorem 11 replacing Theorem 8. We choose our sample as explained in the proof of Theorem 11 and apply the statement of the theorem with the matrix $r_{ij} = (x_{ij} - y_{ij})^2$. We get

$$\left| \sum_{i,j} p_{ij} r_{ij} - \sum_{(i,j)\in S} w_{ij} r_{ij} \right| \leq O\left(\frac{\gamma_2(R)}{\sqrt{d}} \right).$$

Since by definition of our algorithm $r_{ij} = 0$ for $(i,j) \in S$. And since, as explained before $\gamma_2(R) \leq 4\gamma_2(Y)^2$. We conclude that

$$\sum_{i,j} p_{ij}(x_{ij} - y_{ij})^2 \leq O\left(\frac{\gamma_2(Y)^2}{\sqrt{d}} \right).$$

■

ACKNOWLEDGMENTS

We thank Nati Srebro for insightful comments. We also thank Nati Linial for his careful reading of the manuscript, and his numerous valuable comments.

REFERENCES

[1] J. D. Batson, D. A. Spielman, and N. Srivastava, Twice-Ramanujan sparsifiers, STOC ’09: Proceedings of the 41st Annual ACM Symposium on Theory of Computing (May 2009), 2009, pp. 255–262.

[2] A. A. Benczúr and D. R. Karger, Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time, STOC ’1996: Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (May 1996), 1996, pp. 47–55.

[3] E. J. Candes and B. Recht, Exact matrix completion via convex optimization, Found Comput Math 9 (2009), 717–772.

[4] E. J. Candes and T. Tao, The power of convex relaxation: near-optimal matrix completion, IEEE Trans Infor Theo 56 (2010), 2053–2080.

[5] M. Fazel, H. Hindi, and S. P. Boyd, A rank minimization heuristic with application to minimum order system approximation, In Proceedings American Control Conference, Vol. 6, 2001, pp. 4734–4739.

[6] R. Foygel and N. Srebro, Concentration-based guarantees for low-rank matrix reconstruction, 24th Annual Conference on Learning Theory (COLT), 2011.

[7] S. Hoory, N. Linial, and A. Wigderson, Expander graphs and their applications, Bull Am Math Soc 43 (2006), 439–562.

[8] G. J. O. Jameson, Summing and nuclear norms in banach space theory, Cambridge University Press, 1987.

[9] R. H. Keshavan, A. Montanari, and S. Oh, Matrix completion from noisy entries, J Mach Learn Res 11 (2010), 2057–2078.

[10] V. Koltchinskii, A. B. Tsybakov, and K. Lounici, Nuclear norm penalization and optimal rates for noisy low rank matrix completion, Ann Statist 39 (2011), 2302–2329.

[11] T. Lee and A. Shraibman, An approximation algorithm for approximation rank, In Proceedings of the 24th IEEE Conference on Computational Complexity, IEEE, 2008, pp. 351–357.

[12] T. Lee, A. Shraibman, and R. Špalek, A direct product theorem for discrepancy, In Proceedings of the 23rd IEEE Conference on Computational Complexity, IEEE, 2008, pp. 71–80.

[13] A. Lubotzky, R. Phillips, and P. Sarnak, Ramanujan graphs, Combinatorica 8 (1988), 261–277.

[14] G. Pisier, Factorization of linear operators and geometry of Banach spaces, Vol. 60 of CBMS Regional Conference Series in Mathematics, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1986.

[15] B. Recht, A simpler approach to matrix completion, J Machine Learn Res 12 (2011), 3413–3430.

[16] N. Srebro, J. D. M. Rennie, and T. S. Jaakola, Maximum-margin matrix factorization, In Advances in Neural Information Processing Systems 17, Vol. 17, 2005, pp. 1329–1336.

[17] N. Srebro and A. Shraibman, Rank, trace-norm and max-norm, In 18th Annual Conference on Computational Learning Theory (COLT), 2005, pp. 545–560.

[18] N. Tomczak-Jaegermann, Banach-Mazur distances and finite-dimensional operator ideals, Vol. 38 of Pitman Monographs and Surveys in Pure and Applied Mathematics, Longman Scientific & Technical, Harlow, 1989.