Journal of Machine Learning Research 16 (2015) 2999-3034 Submitted 6/14; Revised 4/15; Published 7/15

Completing Any Low-rank Matrix, Provably*

Yudong Chen YUDONG.CHEN@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94704, USA

Srinadh Bhojanapalli BSRINADH@UTEXAS.EDU
Sujay Sanghavi SANGHAVI@MAIL.UTEXAS.EDU
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712, USA

Rachel Ward RWARD@MATH.UTEXAS.EDU
Department of Mathematics and ICES
The University of Texas at Austin
Austin, TX 78712, USA

Editor: Tong Zhang

Abstract

Matrix completion, i.e., the exact and provable recovery of a low-rank matrix from a small subset of its elements, is currently only known to be possible if the matrix satisfies a restrictive structural constraint—known as *incoherence*—on its row and column spaces. In these cases, the subset of elements is assumed to be sampled uniformly at random.

In this paper, we show that *any* rank-r n-by-n matrix can be exactly recovered from as few as $O(nr\log^2 n)$ randomly chosen elements, provided this random choice is made according to a *specific biased distribution* suitably dependent on the coherence structure of the matrix: the probability of any element being sampled should be at least a constant times the sum of the leverage scores of the corresponding row and column. Moreover, we prove that this specific form of sampling is nearly necessary, in a natural precise sense; this implies that many other perhaps more intuitive sampling schemes fail.

We further establish three ways to use the above result for the setting when leverage scores are not known *a priori*. (a) We describe a provably-correct sampling strategy for the case when only the column space is incoherent and no assumption or knowledge of the row space is required. (b) We propose a two-phase sampling procedure for general matrices that first samples to estimate leverage scores followed by sampling for exact recovery. These two approaches assume control over the sampling procedure. (c) By using our main theorem in a reverse direction, we provide an analysis showing the advantages of the (empirically successful) weighted nuclear/trace-norm minimization approach over the vanilla un-weighted formulation given non-uniformly distributed observed elements. This approach does not require controlled sampling or knowledge of the leverage scores.

Keywords: matrix completion, coherence, leverage score, nuclear norm, weighted nuclear norm

∗. Partial preliminary results appeared at the International Conference on Machine Learning (ICML) 2014 under the title "Coherent Matrix Completion".

1. Introduction

Low-rank matrix completion has been the subject of much recent study due to its application in myriad tasks: collaborative filtering, dimensionality reduction, clustering, non negative matrix factorization and localization in sensor networks. Clearly, the problem is ill-posed in general; correspondingly, analytical work on the subject has focused on the joint development of algorithms, and sufficient conditions under which such algorithms are able to recover the matrix.

While they differ in scaling/constant factors, all existing sufficient conditions (Candès and Recht, 2009; Candès and Tao, 2010; Recht, 2011; Keshavan et al., 2010; Gross, 2011; Jain et al., 2013; Negahban and Wainwright, 2012)—with a couple of exceptions we describe in Section 2—require that *(a)* the subset of observed elements should be uniformly randomly chosen, independent of the values of the matrix elements, and *(b)* the low-rank matrix be "incoherent" or "not spiky"—i.e., its row and column spaces should be diffuse, having low inner products with the standard basis vectors. Under these conditions, the matrix has been shown to be provably recoverable—via methods based on convex optimization (Candès and Recht, 2009), alternating minimization (Jain et al., 2013), iterative thresholding (Cai et al., 2010), etc.—using as few as $\Theta(nr \log n)$ observed elements for an $n \times n$ matrix of rank r.

Actually, the incoherence assumption *is required because of the uniform sampling:* coherent matrices are those which have most of their mass in a relatively small number of elements. By sampling entries uniformly and independently at random, most of the mass of a coherent low-rank matrix will be missed; this could (and *does*) throw off most existing methods for exact matrix completion. One could imagine that if the sampling is adapted to the matrix, roughly in a way that ensures that elements with more mass are more likely to be observed, then it may be possible for *existing* methods to recover the full matrix.

In this paper, we show that the incoherence requirement can be eliminated completely, provided the sampling distribution is dependent on the matrix to be recovered in the right way. Specifically, we have the following results.

1. If the probability of an element being observed is proportional to the sum of the corresponding row and column leverage scores (which are local versions of the standard incoherence parameter) of the underlying matrix, then an *arbitrary* rank-r matrix can be exactly recovered from $\Theta(nr \log^2 n)$ observed elements with high probability, using nuclear norm minimization (Theorem 2 and Corollary 3). In the case when all leverage scores are uniformly bounded from above, our results reduce to existing guarantees for incoherent matrices using uniform sampling. Our sample complexity bound $\Theta(nr \log^2 n)$ is optimal up to a single factor of $\log^2 n$, since the degrees of freedom in an $n \times n$ matrix of rank r is in general in the order of nr. Moreover, we show that to complete a coherent matrix, it is *necessary* (in certain precise sense) to sample according to the leverage scores as above (Theorem 6).

2. For a matrix whose column space is incoherent and row space is arbitrarily coherent, our results immediately lead to a provably correct sampling scheme which *requires no prior knowledge of the leverage scores of the underlying matrix* and has near optimal sample complexity (Corollary 4).

3. We provide numerical evidence that a two-phase adaptive sampling strategy, which assumes no prior knowledge about the leverage scores of the underlying matrix, can perform on par with the optimal sampling strategy in completing coherent matrices, and significantly outperforms uniform sampling (Section 4). Specifically, we consider a two-phase sampling strategy whereby given a fixed budget of m samples, we first draw a fixed proportion of samples uniformly at random, and then draw the remaining samples according to the leverage scores of the resulting sampled matrix.

4. As a corollary of our main theorem, we are able to obtain the first exact recovery guarantee for the *weighted* nuclear norm minimization approach, which can be viewed as adjusting the leverage scores to align with the given sampling distribution. Our results provide a strategy for choosing the weights when non-uniformly distributed samples are *given* so as to order-wise reduce the sample complexity of the weighted approach to that of the standard *unweighted* formulation (Theorem 7). Our theorem quantifies the benefit of the weighted approach, thus providing theoretical justification for its good empirical performance observed in Srebro and Salakhutdinov (2010); Foygel et al. (2011); Negahban and Wainwright (2012).

These results provide a deeper and more general theoretical understanding of the relation between the sampling procedure and the matrix coherence/leverage-score structure, and how they affect the recovery performance. While in practice one may not have complete control over the sampling procedure, or exact knowledge of the matrix leverage scores, partial control and knowledge are often possible, and we believe our theory provides useful approximations and insights. We expect that the ideas and results in this paper will serve as the foundation for developing algorithms for more general settings and applications.

Our theoretical results are achieved by a new analysis based on concentration bounds involving the *weighted* $\ell_{\infty,2}$ matrix norm, defined as the maximum of the appropriately weighted row and column norms of the matrix. This differs from previous approaches that use ℓ_∞ or unweighted $\ell_{\infty,2}$ norm bounds (Gross, 2011; Recht, 2011; Chen, 2015). In some sense, using the weighted $\ell_{\infty,2}$-type bounds is natural for the analysis of low-rank matrix recover/approximation when the observations are in the form of entries of rows/columns of the matrix, because the rank is a property of the rows and columns of the matrix rather than its individual elements, and the weighted norm captures the relative importance of the rows/columns. Therefore, our techniques based on the $\ell_{\infty,2}$ norm might be of independent interest beyond the specific settings and algorithms considered here.

1.1 Organization

In Section 2 we briefly survey the relevant literature. We present our main results for coherent matrix completion in Section 3. In Section 4 we propose a two-phase algorithm that requires no prior knowledge about the underlying matrix's leverage scores. In Section 5 we provide guarantees for weighted nuclear norm minimization. The paper concludes with a discussion of future work in Section 6. We provide the proofs of the main theorems in the appendix.

2. Related Work

There is now a vast body of literature on matrix completion, and an even bigger body of literature on matrix approximations; we restrict our literature review here to papers that are most directly related.

Completion of incoherent and row-coherent matrices: The first algorithm and theoretical guarantees for exact low-rank matrix completion appeared in Candès and Recht (2009); there it was shown that nuclear norm minimization works when the low-rank matrix is incoherent, and the sampling is uniform random and independent of the matrix. Subsequent works have refined provable completion results for incoherent matrices under the uniform random sampling model, both via nuclear norm minimization (Candès and Tao, 2010; Recht, 2011; Gross, 2011; Chen, 2015), and other methods like SVD followed by local descent (Keshavan et al., 2010) and alternating minimization (Jain et al., 2013), etc. The setting with sparse errors and additive noise is also considered (Candès and Plan, 2010; Chandrasekaran et al., 2011; Chen et al., 2013; Candès et al., 2011; Negahban and Wainwright, 2012).

The recent work in Krishnamurthy and Singh (2013) considers matrix completion when the row space is allowed to be coherent but the column space is still required to be incoherent with parameter μ_0. Their proposed adaptive sampling algorithm selects columns to observe in their entirety and requires a total of $O(\mu_0 r^{3/2} n \log(2r/\delta))$ observed elements with a success probability $1-\delta$, which is superlinear in r. A corollary of our results guarantees a sample complexity that is linear in r in this row-coherent setting. The sample complexity was recently improved to $O(\mu_0 rn \log^2(r^2/\delta))$ in Krishnamurthy and Singh (2014).

Matrix approximations via sub-sampling: Weighted sampling methods have been widely considered in the related context of matrix *sparsification*, where one aims to approximate a given large dense matrix with a sparse matrix. The strategy of element-wise matrix sparsification was introduced in Achlioptas and McSherry (2007). They propose and provide bounds for the ℓ_2 *element-wise sampling* model, where elements of the matrix are sampled with probability proportional to their squared magnitude. These bounds were later refined in Drineas and Zouzias (2011). Alternatively, Arora et al. (2006) propose the ℓ_1 *element-wise sampling* model, where elements are sampled with probabilities proportional to their magnitude. This model was further investigated in Achlioptas et al. (2013) and argued to be almost always preferable to ℓ_2 sampling.

Closely related to the matrix sparsification problem is the matrix *column selection* problem, where one aims to find the "best" k column subset of a matrix to use as an approximation. State-of-the-art algorithms for column subset selection (Boutsidis et al., 2009; Mahoney, 2011) involve randomized sampling strategies whereby columns are selected proportionally to their *statistical leverage scores*—the squared Euclidean norms of projections of the canonical unit vectors on the column subspaces. The statistical leverage scores of a matrix can be approximated efficiently, faster than the time needed to compute an SVD (Drineas et al., 2012). Statistical leverage scores are also used extensively in statistical regression analysis for outlier detection (Chatterjee and Hadi, 1986). More recently, statistical leverage scores were used in the context of graph sparsification under the name of graph resistance (Spielman and Srivastava, 2011). The sampling distribution we use for the

matrix completion guarantees of this paper is *elemen-wise* and based on statistical leverage scores. As shown both theoretically (Theorem 6) and empirically (Section 4.1), sampling as such outperforms both ℓ_1 and ℓ_2 element-wise sampling, at least in the context of matrix completion.

Weighted sampling in compressed sensing: This paper is similar in spirit to recent work in compressed sensing which shows that sparse recovery guarantees traditionally requiring mutual incoherence can be extended to systems which are only *weakly* incoherent, without any loss of approximation power, provided measurements from the sensing basis are subsampled according to their coherence with the sparsity basis. This notion of *local coherence sampling* seems to have originated in Rauhut and Ward (2012) in the context of sparse orthogonal polynomial expansions, and has found applications in uncertainty quantification (Yang and Karniadakis, 2013), interpolation with spherical harmonics (Burq et al., 2012), and MRI compressive imaging (Krahmer and Ward, 2014).

3. Main Results

The results in this paper hold for what is arguably the most popular approach to matrix completion: nuclear norm minimization. If the true matrix is M with its (i,j)-th element denoted by M_{ij}, and the set of observed elements is Ω, this method estimates M via the optimum of the convex program:

$$\begin{aligned} \min_X \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega. \end{aligned} \tag{1}$$

where the nuclear norm $\|\cdot\|_*$ of a matrix is the sum of its singular values.[1]

We focus on the setting where matrix elements are revealed according an underlying probability distribution. To introduce the distribution of interest, we first need a definition.

Definition 1 (Leverage Scores) *For an $n_1 \times n_2$ real-valued matrix M of rank r whose rank-r SVD is given by $U\Sigma V^\top$, its (normalized) leverage scores—$\mu_i(M)$ for any row i, and $\nu_j(M)$ for any column j—are defined as*

$$\begin{aligned} \mu_i(M) &:= \frac{n_1}{r}\left\|U^\top e_i\right\|_2^2, \quad i = 1, 2, \ldots, n_1, \\ \nu_j(M) &:= \frac{n_2}{r}\left\|V^\top e_j\right\|_2^2, \quad j = 1, 2, \ldots, n_2, \end{aligned} \tag{2}$$

where e_i denotes the i-th standard basis element with appropriate dimension.[2]

Note that the leverage scores are non-negative, and are functions of the column and row spaces of the matrix M. Since U and V have orthonormal columns, we always have relationship $\sum_i \mu_i(M) r/n_1 = \sum_j \nu_j(M) r/n_2 = r$. The standard *incoherence parameter* μ_0

1. This becomes the trace norm for positive-definite matrices. It is now well-recognized to be a convex surrogate for the rank function (Fazel, 2002).
2. In the matrix sparsification literature (Drineas et al., 2012; Boutsidis et al., 2009) and beyond, the leverage scores of M often refer to the *un-normalized* quantities $\left\|U^\top e_i\right\|^2$ and $\left\|V^\top e_j\right\|^2$.

of M used in the previous literature corresponds to a global upper bound on the leverage scores:

$$\mu_0 \geq \max_{i,j}\{\mu_i(M), \nu_j(M)\}.$$

Therefore, the leverage scores can be considered as the localized versions of the standard incoherence parameter.

We are ready to state our main result, the theorem below.

Theorem 2 *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix of rank r, and suppose that its elements M_{ij} are observed only over a subset of elements $\Omega \subset [n_1]\times[n_2]$. There is a universal constant $c_0 > 0$ such that, if each element (i, j) is independently observed with probability p_{ij}, and p_{ij} satisfies*

$$\begin{aligned} p_{ij} &\geq \min\left\{ c_0 \frac{(\mu_i(M) + \nu_j(M))\, r \log^2(n_1+n_2)}{\min\{n_1, n_2\}},\quad 1 \right\}, \qquad (3)\\ p_{ij} &\geq \frac{1}{\min\{n_1, n_2\}^{10}}, \end{aligned}$$

then M is the unique optimal solution to the nuclear norm minimization problem (1) with probability at least $1 - 5(n_1 + n_2)^{-10}$.

We will refer to the sampling strategy (3) as *leveraged sampling*. Note that the expected number of observed elements is $\sum_{i,j} p_{ij}$, and this satisfies

$$\begin{aligned} \sum_{i,j} p_{ij} &\geq \max\left\{ c_0 \frac{r\log^2(n_1+n_2)}{\min\{n_1,n_2\}} \sum_{i,j} \left(\mu_i(M) + \nu_j(M)\right), \sum_{i,j} \frac{1}{\min\{n_1,n_2\}^{10}} \right\} \\ &= 2c_0 \max\{n_1, n_2\}\, r \log^2(n_1+n_2), \end{aligned}$$

which is independent of the leverage scores, or indeed any other property of the matrix. Hoeffding's inequality implies that the actual number of observed elements sharply concentrates around its expectation, leading to the following corollary:

Corollary 3 *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix of rank r. Draw a subset Ω of its elements by leveraged sampling according to the procedure described in Theorem 2. There is a universal constant $c_0 > 0$ such that the following holds with probability at least $1 - 10(n_1 + n_2)^{-10}$: the number m of revealed elements is bounded by*

$$|\Omega| \leq 3c_0 \max\{n_1, n_2\}\, r \log^2(n_1+n_2)$$

and M is the unique optimal solution to the nuclear norm minimization program (1).

We now provide comments and discussion.

(A) Roughly speaking, the condition given in (3) ensures that elements in important rows/columns (indicated by large leverage scores μ_i and ν_j) of the matrix should be observed more often. Note that Theorem 2 only stipulates that an *inequality* relation hold between p_{ij} and $\{\mu_i(M), \nu_j(M)\}$. This allows for there to be some discrepancy between the sampling

distribution and the leverage scores. It also has the natural interpretation that the more the sampling distribution $\{p_{ij}\}$ is "aligned" to the leverage score pattern of the matrix, the fewer observations are needed.

(B) Sampling based on leverage scores provides close to the optimal number of sampled elements required for exact recovery (when sampled with any distribution). In particular, recall that the number of degrees of freedom of an $n \times n$ matrix of rank r is $2nr(1 - r/2n)$, and knowing the leverage scores of the matrix reduces the degrees of freedom by $2n$ in the worst case. Hence, regardless of how the elements are sampled, a minimum of $\Theta(nr)$ elements is required to recover the matrix. Theorem 2 matches this lower bound, with an additional $O(\log^2(n))$ factor.

(C) Our work improves on existing results *even* in the case of uniform sampling and uniform incoherence. Recall that the original work of Candès and Recht (2009), and subsequent works (Candès and Tao, 2010; Recht, 2011; Gross, 2011) give recovery guarantees based on two parameters of the matrix $M \in \mathbb{R}^{n\times n}$ (assuming its SVD is $U\Sigma V^\top$): (a) the (above-defined) *incoherence parameter* μ_0, which is a uniform bound on the leverage scores, and (b) a *joint incoherence parameter* μ_{str} defined by $\|UV^\top\|_\infty = \sqrt{\frac{r\mu_{\text{str}}}{n^2}}$. With these definitions, the current state of the art states that if the sampling probability is uniform and satisfies

$$p_{ij} \equiv p \geq c\frac{\max\{\mu_0, \mu_{\text{str}}\} r \log^2 n}{n}, \quad \forall i, j,$$

where c is a constant, then M will be the unique optimum of (1) with high probability. A direct corollary of our work improves on this result, by removing the need for extra constraints on the joint incoherence; in particular, it is easy to see that our theorem implies that a uniform sampling probability of $p \geq c\frac{\mu_0 r \log^2 n}{n}$—that is, with no μ_{str}—guarantees recovery of M with high probability. Note that μ_{str} can be as high as $\mu_0 r$, for example, in the case when M is positive semi-definite; our corollary thus removes this sub-optimal dependence on the rank and on the incoherence parameter. This improvement was recently observed in Chen (2015).

3.1 Knowledge-Free Completion for Row Coherent Matrices

Theorem 2 immediately yields a useful result in scenarios where only the row space of a matrix is coherent and one has control over the sampling of the matrix. This setting is considered by Krishnamurthy and Singh (2013).

Suppose the column space of $M \in \mathbb{R}^{n\times n}$ is incoherent with $\max_i \mu_i(M) \leq \mu_0$ and the row space is arbitrary (we consider square matrix for simplicity). For a number $0 < \delta < 1$ to be prescribed by the user, We choose each row of M with probability $\frac{10\mu_0 r}{n} \log \frac{2r}{\delta}$, and observe all the elements of the chosen rows. We then compute the leverage scores $\{\tilde{\nu}_j\}$ of the space spanned by these rows, and use them as estimates for $\nu_j(M)$, the leverage scores of M. Based on these estimates, we can perform leveraged sampling according to (3) and then use nuclear norm minimization to recover M. Note that this procedure does not require any prior knowledge about the leverage scores of M. The following corollary shows that the procedure is *provably correct* and exactly recovers M with high probability, using a near-optimal number of samples.

Corollary 4 *For any number $0 < \delta < 1$ and some universal constants $c_0, c_1 > 0$, the following holds. With probability at least $1 - \delta$, the above procedure computes the column leverage scores of M exactly, i.e., $\tilde{\nu}_j = \nu_j(M), \forall j \in [n]$. If we set $\delta = 4n^{-10}$, and further sample a set Ω of elements of M with probabilities*

$$p_{ij} = \min\left\{c_0 \frac{(\mu_0 + \tilde{\nu}_j) r \log^2 n}{n}, 1\right\}, \quad \forall i, j,$$

then with probability at least $1 - 10n^{-10}$, M is the unique optimal solution to the nuclear norm minimization program (1), and we use a total of at most $c_1 \mu_0 rn \log^2 n$ samples.

The algorithm proposed in Krishnamurthy and Singh (2013) requires a sample complexity of $O(\mu_0 r^{3/2} n \log(2r/\delta))$ (and guarantees a success probability of $1 - \delta$). Our result in the corollary above removes the sub-optimal $r^{3/2}$ factor in the sample complexity. Very recently Krishnamurthy and Singh (2014) provide a new sample complexity bound $O(\mu_0 rn \log^2(r^2/\delta))$ using the same algorithm from their previous paper. We note that our sampling strategy is different from theirs: we sample entire rows of M, whereas they sample entire columns.

3.2 Necessity of Leveraged Sampling

In this subsection, we show that the leveraged sampling in (3) is necessary for completing a coherent matrix in a certain precise sense. For simplicity, we restrict ourselves to square matrices in $\mathbb{R}^{n \times n}$. Suppose each element (i, j) is observed independently with probability p_{ij}. We consider a family of sampling probabilities $\{p_{ij}\}$ with the following property.

Definition 5 (Location Invariance) *$\{p_{ij}\}$ is said to be location-invariant with respect to the matrix M if the following are satisfied: (1) For any two rows $i \neq i'$ that are identical, i.e., $M_{ij} = M_{i'j}$ for all j, we have $p_{ij} = p_{i'j}$ for all j; (2) For any two columns $j \neq j'$ that are identical, i.e., $M_{ij} = M_{ij'}$ for all i, we have $p_{ij} = p_{ij'}$ for all i.*

In other words, $\{p_{ij}\}$ is location-invariant with respect to M if identical rows (or columns) of M have identical sampling probabilities. We consider this assumption very mild, and it covers the leveraged sampling as well as many other typical sampling schemes, including:

- uniform sampling, where $p_{ij} \equiv p$,
- element-wise magnitude sampling, where $p_{ij} \propto |M_{ij}|$ (ℓ_1 sampling) or $p_{ij} \propto M_{ij}^2$ (ℓ_2 sampling), and
- row/column-wise magnitude sampling, where $p_{ij} \propto f\left(\|M_{i\cdot}\|_2, \|M_{\cdot j}\|_2\right)$ for some (usually coordinate-wise non-decreasing) function $f : \mathbb{R}_+^2 \mapsto [0, 1]$.

Given two n-dimensional vectors $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ and $\vec{\nu} = (\nu_1, \ldots, \nu_n)$, we use $\mathcal{M}_r(\vec{\mu}, \vec{\nu})$ to denote the set of rank-r matrices whose leverage scores are bounded by $\vec{\mu}$ and $\vec{\nu}$; that is,

$$\mathcal{M}_r(\vec{\mu}, \vec{\nu}) := \left\{M \in \mathbb{R}^{n \times n} : \mathrm{rank}(M) = r; \mu_i(M) \leq \mu_i, \nu_j(M) \leq \nu_j, \forall i, j\right\}.$$

We have the following results.

Theorem 6 *Suppose $n \geq r \geq 2$. Given any $2r$ numbers $a_1, \ldots, a_r$ and $b_1, \ldots, b_r$ with $\frac{r}{4} \leq \sum_{k=1}^{r} \frac{1}{a_k}, \sum_{k=1}^{r} \frac{1}{b_k} \leq r$ and $\frac{2}{r} \leq a_k, b_k \leq \frac{2n}{r}, \forall k \in [r]$, there exist two n-dimensional vectors $\vec{\mu}$ and $\vec{\nu}$ and the corresponding set $\mathcal{M}_r(\vec{\mu}, \vec{\nu})$ with the following properties:*

1. *For each $i, j \in [n]$, $\mu_i = a_k$ and $\nu_j = b_{k'}$ for some $k, k' \in [r]$. That is, the values of the leverage scores are given by $\{a_k\}$ and $\{b_{k'}\}$.*

2. *There exists a matrix $M^{(0)} \in \mathcal{M}_r(\vec{\mu}, \vec{\nu})$ for which the following holds. If $\{p_{ij}\}$ is location-invariant w.r.t. $M^{(0)}$, and for some (i_0, j_0),*

$$p_{i_0 j_0} \leq \frac{\mu_{i_0} + \nu_{j_0}}{4n} \cdot r \log\left(\frac{2n}{(\mu_{i_0} \vee \nu_{j_0}) r}\right), \text{[3]} \tag{4}$$

   then with probability at least $\frac{1}{4}$, the following conclusion holds: There are infinitely many matrices $M^{(1)} \neq M^{(0)}$ in $\mathcal{M}_r(\vec{\mu}, \vec{\nu})$ such that $\{p_{ij}\}$ is location-invariant w.r.t. $M^{(1)}$, and

$$M^{(0)}_{ij} = M^{(1)}_{ij}, \quad \forall (i, j) \in \Omega.$$

3. *If we replace the condition (4) with*

$$p_{i_0 j_0} \leq \frac{\mu_{i_0} + \nu_{j_0}}{4n} \cdot r \log\left(\frac{n}{2}\right), \tag{5}$$

   then the conclusion above holds with probability at least $\frac{1}{n}$.

In other words, if (4) holds, then with probability at least $1/4$, no method can distinguish between $M^{(0)}$ and $M^{(1)}$; similarly, if (5) holds, then with probability at least $1/n$ no method succeeds. We shall compare these results with Theorem 2, which guarantees that if we use leveraged sampling,

$$p_{ij} \geq c_0 \frac{\mu_i + \nu_j}{n} \cdot r \log n, \quad \forall i, j$$

for some universal constant c_0, then for any matrix $M^{(0)}$ in $\mathcal{M}_r(\vec{\mu}, \vec{\nu})$, the nuclear norm minimization approach (1) recovers $M^{(0)}$ from its observed elements with failure probability no more than $\frac{1}{n}$. Therefore, under the setting of Theorem 6, leveraged sampling is *sufficient and necessary* for matrix completion up to one logarithmic factor for a target failure probability $\frac{1}{n}$ (or up to two logarithmic factors for a target failure probability $\frac{1}{4}$).

Admittedly, the setting covered by Theorem 6 has several restrictions on the sampling distributions and the values of the leverage scores. Nevertheless, we believe this result captures some essential difficulties in recovering general coherent matrices, and highlights how the sampling probabilities should relate in a specific way to the leverage score structure of the underlying object.

4. A Two-Phase Sampling Procedure

We have seen that one can exactly recover an arbitrary $n \times n$ rank-r matrix using $\Theta(nr \log^2 n)$ elements if sampled in accordance with the leverage scores. In practical applications of

3. We use the notation $a \vee b = \max\{a, b\}$.

matrix completion, even when the user is free to choose how to sample the matrix elements, she may not be privy to the leverage scores $\{\mu_i(M), \nu_j(M)\}$. In this section we propose a two-phase sampling procedure, described below and in Algorithm 1, which assumes no a priori knowledge about the matrix leverage scores, yet is observed to be competitive with the "oracle" leveraged sampling distribution (3).

Suppose we are given a total budget of m samples. The first step of the algorithm is to use the first β fraction of the budget to estimate the leverage scores of the underlying matrix, where $\beta \in [0, 1]$. Specifically, take a set of indices Ω sampled uniformly without replacement[4] such that $|\Omega| = \beta m$, and let $\mathcal{P}_\Omega(\cdot)$ be the sampling operator which maps the matrix elements not in Ω to 0. Take the rank-r SVD of $\mathcal{P}_\Omega(M)$, $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$, where $\tilde{U}, \tilde{V} \in \mathbb{R}^{n\times r}$ and $\tilde{\Sigma} \in \mathbb{R}^{r\times r}$, and then use the leverage scores $\tilde{\mu}_i := \mu_i(\tilde{U}\tilde{\Sigma}\tilde{V}^\top)$ and $\tilde{\nu}_j := \nu_j(\tilde{U}\tilde{\Sigma}\tilde{V}^\top)$ as estimates for the column and row leverage scores of M. Now as the second step, generate the remaining $(1-\beta)m$ samples of the matrix M by sampling without replacement with distribution

$$\tilde{p}_{ij} \propto \frac{(\tilde{\mu}_i + \tilde{\nu}_j) r \log^2(2n)}{n}. \tag{6}$$

Let $\tilde{\Omega}$ denote the new set of samples. Using the combined set of samples $\mathcal{P}_{\Omega\cup\tilde{\Omega}}(M)$ as constraints, run the nuclear norm minimization program (1). Let $\hat{M}$ be the optimum of this program.

This approach of adjusting the sampling distribution based on leverage scores is relevant whenever we have some freedom in choosing the observed entries. For example, many recommendation systems do actively solicit users' opinions on some items chosen by the system, e.g., by asking them to fill out a survey or to choose from a list of items. While our assumptions are not strictly satisfied in practice, they are useful approximations and provide guidance for designing/analyzing practical systems. For example, in many systems there exist popular items that are viewed/rated by a large number of users, and "heavy" users that view/rate a large number of items. Our row-wise sampling procedure discussed in Section 3.1 can be viewed as an approximation of such settings.

To understand the performance of the two-phase algorithm, assume that the initial set of $m_1 = \beta m$ samples $\mathcal{P}_\Omega(M)$ are generated uniformly at random. If the underlying matrix

4. Note that sampling without replacement has lesser failure probability than the equivalent binomial sampling with replacement (Recht, 2011).

Algorithm 1 Two-phase sampling for coherent matrix completion

input Rank parameter r, sample budget m, and parameter $\beta \in [0, 1]$

Step 1: Obtain the initial set Ω by sampling uniformly without replacement such that $|\Omega| = \beta m$. Compute best rank-r approximation to $\mathcal{P}_\Omega(M)$, $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$, and its leverage scores $\{\tilde{\mu}_i\}$ and $\{\tilde{\nu}_j\}$.

Step 2: Generate set of $(1-\beta)m$ new samples $\tilde{\Omega}$ by sampling without replacement with distribution (6). Set

$$\hat{M} = \arg\min_X \|X\|_* \text{ s.t } \mathcal{P}_{\Omega\cup\tilde{\Omega}}(X) = \mathcal{P}_{\Omega\cup\tilde{\Omega}}(M).$$

output Completed matrix $\hat{M}$.

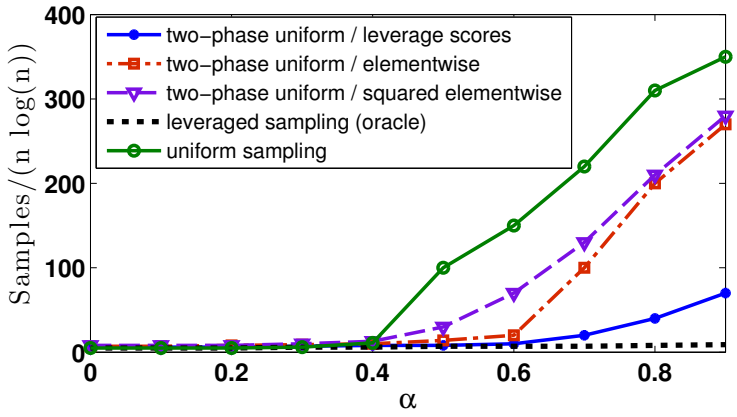


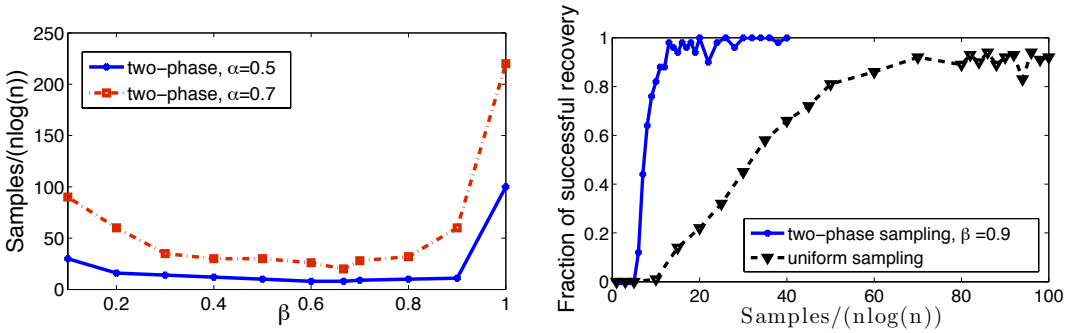
Figure 1: Performance of Algorithm 1 for power-law matrices: We consider rank-5 matrices of the form $M = DUV^\top D$, where elements of the matrices U and V are generated independently from a Gaussian distribution $\mathcal{N}(0,1)$ and D is a diagonal matrix with $D_{ii} = \frac{1}{i^\alpha}$. Higher values of α correspond to more non-uniform leverage scores and less incoherent matrices. The above simulations are run with two-phase parameter $\beta = 2/3$. Leveraged sampling (3) gives the best results of successful recovery using roughly $10n\log(n)$ samples for all values of α in accordance with Theorem 2. Surprisingly, sampling according to (6) with estimated leverage scores has almost the same sample complexity for $\alpha \leq 0.7$. Uniform sampling and sampling proportional to element and element squared perform well for low values of α, but their performance degrades quickly for $\alpha > 0.6$.

M is incoherent, then already the algorithm will recover M if $m_1 = \Theta(nr\log^2(2n))$. On the other hand, if M is *highly* coherent, having almost all energy concentrated on just a few elements, then the estimated leverage scores (6) from uniform sampling in the first step will be poor and hence the recovery algorithm suffers. Between these two extremes, there is reason to believe that the two-phase sampling procedure will provide a better estimate to the underlying matrix than if all m elements were sampled uniformly. Indeed, numerical experiments suggest that the two-phase procedure can indeed significantly outperform uniform sampling for completing coherent matrices.

4.1 Numerical Experiments

We now study the performance of the two-phase sampling procedure outlined in Algorithm 1 through numerical experiments. For this, we consider rank-5 matrices of size 500×500 of the form $M = DUV^\top D$, where the elements of the matrices U and V are i.i.d. Gaussian $\mathcal{N}(0,1)$ and D is a diagonal matrix with power-law decay, $D_{ii} = i^{-\alpha}, 1 \leq i \leq 500$. We refer to such constructions as *power-law* matrices. The parameter α adjusts the leverage scores (and hence the coherence level) of M with $\alpha = 0$ being maximal incoherence $\mu_0 = \Theta(1)$ and $\alpha = 1$ corresponding to maximal coherence $\mu_0 = \Theta(n)$.

Figure 1 plots the number of samples required for successful recovery (y-axis) for different values of α (x-axis) and $\beta = 2/3$ using Algorithm 1 with the initial samples Ω



(a) (b)

Figure 2: We consider power-law matrices with parameter $\alpha = 0.5$ and $\alpha = 0.7$. (a): This plot shows that Algorithm 1 successfully recovers coherent low-rank matrices with fewest samples ($\approx 10n\log(n)$) when the proportion of initial samples drawn from the uniform distribution is in the range $\beta \in [0.5, 0.8]$. In particular, the sample complexity is significantly lower than that for uniform sampling ($\beta = 1$). Note the x-axis starts at 0.1. (b): Even by drawing 90% of the samples uniformly and using the estimated leverage scores to sample the remaining 10% samples, one observes a marked improvement in the rate of recovery.

taken uniformly at random. Successful recovery is defined as when at least 95% of trials have relative errors in the Frobenius norm $\|M - \hat{M}\|_F / \|M\|_F$ not exceeding 0.01. To put the results in perspective, we plot it in Figure 1 against the performance of pure uniform sampling, as well as other popular sampling distributions from the matrix sparsification literature (Achlioptas and McSherry, 2007; Achlioptas et al., 2013; Arora et al., 2006; Drineas and Zouzias, 2011), namely, in step 2 of the algorithm, sampling proportional to element ($\tilde{p}_{ij} \propto |\tilde{M}_{ij}|$) and sampling proportional to element squared ($\tilde{p}_{ij} \propto \tilde{M}_{ij}^2$), as opposed to sampling from the distribution (6). In all cases, the estimated matrix $\tilde{M}$ is constructed from the rank-r SVD of $\mathcal{P}_\Omega(M)$, $\tilde{M} = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$. Performance of nuclear norm minimization using samples generated according to the "oracle" distribution (3) serves as baseline for the best possible recovery, as theoretically justified by Theorem 2. We use the Augmented Lagrangian Method (ALM) based solver in Lin et al. (2009) to solve the convex optimization program (1).

Figure 1 suggests that the two-phase algorithm performs comparably to the theoretically optimal leverage scores-based distribution (3), despite not having access to the underlying leverage scores, in the regime of mild to moderate coherence. While the element-wise sampling strategies perform comparably for low values of α, the number of samples for successful recovery increases quickly for $\alpha > 0.6$. Completion from purely uniformly sampled elements requires significantly more samples at higher values of α.

Choosing β: Recall that the parameter β in Algorithm 1 is the fraction of uniform samples used to estimate the leverage scores. Figure 2(a) plots the number of samples required for successful recovery (y-axis) as β (x-axis) varies from 0 to 1 for different values of α. Setting $\beta = 1$ reduces to purely uniform sampling, and for small values of β, the leverage scores estimated in (6) will be far from the actual leverage scores. Then, as expected, the

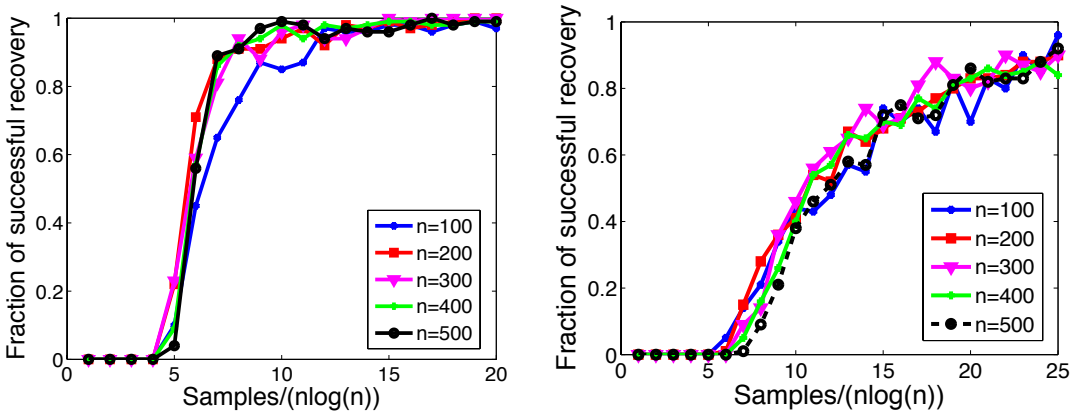


Figure 3: Scaling of sample complexity of Algorithm 1 with n. We consider power-law matrices with $\alpha = 0.5$ in plot (a) and 0.7 in plot (b). We set $\beta = 2/3$ for this set of simulations. The plots suggest that the sample complexity of Algorithm 1 scales roughly as $\Theta(n \log(n))$.

sample complexity goes up for β near 0 and $\beta = 1$. We find the algorithm performs well for a wide range of β, and setting $\beta \approx 2/3$ results in the lowest sample complexity. Surprisingly, even taking $\beta = 0.9$ as opposed to pure uniform sampling $\beta = 1$ results in a significant decrease in the sample complexity; see Figure 2(b) for more details. That is, even budgeting just a small fraction of samples to be drawn from the estimated leverage scores can significantly improve the success rate in low-rank matrix recovery as long as the underlying matrix is not completely coherent. In applications like collaborative filtering, this would imply that incentivizing just a small fraction of users to rate a few selected movies according to the estimated leverage score distribution obtained by previous samples has the potential to greatly improve the quality of the recovered matrix of preferences.

In Figure 3 we compare the performance of the two-phase algorithm for different values of the matrix dimension n, and notice for each n a phase transition occurring at $\Theta(n \log(n))$ samples. In Figure 4 we consider the scenario where the samples are noisy and compare the performance of Algorithm 1 to uniform sampling and the theoretically-optimal leveraged sampling from Theorem 2. Specifically we assume that the samples are generated from $M + Z$ where Z is a Gaussian noise matrix. We consider two values for the noise $\sigma \stackrel{\text{def}}{=} \|Z\|_F / \|M\|_F$: $\sigma = 0.1$ and $\sigma = 0.2$. The figures plot relative error in Frobenius norm (y-axis), vs total number of samples m (x-axis). These plots demonstrate the robustness of the algorithm to noise and once again show that sampling with estimated leverage scores can be as good as sampling with exact leverage scores for matrix recovery using nuclear norm minimization for $\alpha \leq 0.7$.

The empirical results in this section demonstrate the advantage of the two-phase algorithm over uniform sampling. It is an interesting future problem to provide rigorous analysis on the sample complexity of the algorithm. We note that there is an $\Omega(n^2)$ lower bound on

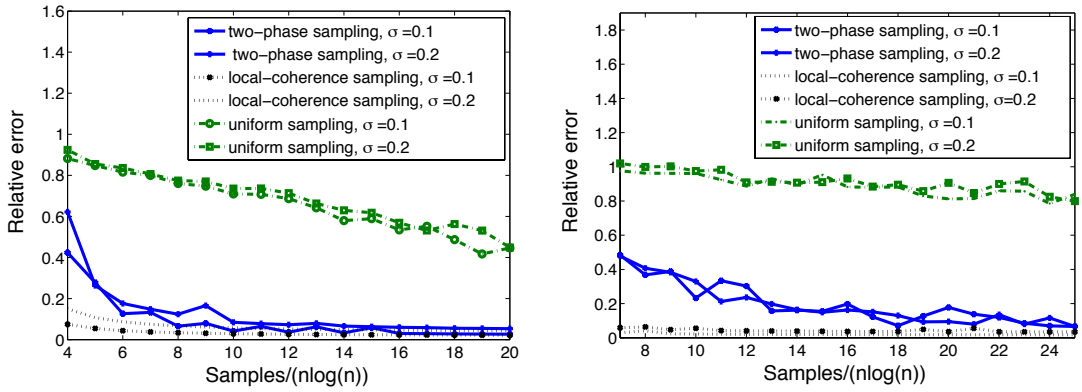


Figure 4: Performance of Algorithm 1 with noisy samples: We consider power-law matrices (with $\alpha = 0.5$ in plot (a) and $\alpha = 0.7$ in plot (b)), perturbed by a Gaussian noise matrix Z with $\|Z\|_F/\|M\|_F = \sigma$. The plots consider two different noise levels, $\sigma = 0.1$ and $\sigma = 0.2$. We compare two-phase sampling (Algorithm 1) with $\beta = 2/3$, sampling from the exact leverage scores, and uniform sampling. Algorithm 1 has relative error almost as low as the leveraged sampling without requiring any a priori knowledge of the low-rank matrix, while uniform sampling suffers dramatically.

the sample complexity for algorithms using passive sampling when the underlying matrix is maximally coherent (Krishnamurthy and Singh, 2014).

5. Weighted Nuclear Norm Minimization

Theorem 2 suggests that the performance of nuclear norm minimization will be better if the set of observed elements is aligned with the leverage scores of the matrix. Interestingly, Theorem 2 can also be used in a reverse way: *one may adjust the leverage scores to align with a given set of observed elements*. Here we demonstrate an application of this idea in quantifying the benefit of *weighted* nuclear norm minimization when the revealed elements are distributed non-uniformly.

Suppose the underlying matrix of interest is incoherent. In many applications, we do not have the freedom to choose which elements to observe. Instead, the revealed elements are *given* to us, and distributed non-uniformly among the rows and columns. As observed in Srebro and Salakhutdinov (2010), standard unweighted nuclear norm minimization (1) is inefficient in this setting. They propose to instead use weighted nuclear norm minimization for low-rank matrix completion:

$$\begin{aligned} \hat{X} = \arg \min_{X \in \mathbb{R}^{n_1 \times n_2}} \quad & \|RXC\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij}, \text{ for } (i,j) \in \Omega, \end{aligned} \tag{7}$$

where $R = \text{diag}(R_1, R_2, \ldots, R_{n_1}) \in \mathbb{R}^{n_1 \times n_1}$ and $C = \text{diag}(C_1, C_2, \ldots, C_{n_2}) \in \mathbb{R}^{n_2 \times n_2}$ are user-specified diagonal weight matrices with positive diagonal elements.

We now provide a theoretical guarantee for this method, and quantify its advantage over unweighted nuclear norm minimization. Our analysis is based on the observation that weighted nuclear norm minimization can be viewed as a way of scaling the rows and columns of the underlying matrix so that its leverage scores are adjusted to reflect the given non-uniform sampling distributions. Suppose $M \in \mathbb{R}^{n_1 \times n_2}$ has rank r and satisfies the standard incoherence condition $\max_{i,j} \{\mu_i(M), \nu_j(M)\} \le \mu_0$. Let $\lfloor x \rfloor$ denote the largest integer not exceeding x. Under this setting, we can apply Theorem 2 to establish the following:

Theorem 7 *Without loss of generality, assume $R_1 \le R_2 \le \cdots \le R_{n_1}$ and $C_1 \le C_2 \le \cdots \le C_{n_2}$. There exists a universal constant c_0 such that M is the unique optimum to (7) with probability at least $1 - 5(n_1+n_2)^{-10}$ provided that for all i, j, $p_{ij} \ge \frac{1}{\min\{n_1,n_2\}^{10}}$ and*

$$p_{ij} \ge c_0 \left(\frac{R_i^2}{\sum_{i'=1}^{\lfloor n_1/(\mu_0 r) \rfloor} R_{i'}^2} + \frac{C_j^2}{\sum_{j'=1}^{\lfloor n_2/(\mu_0 r) \rfloor} C_{j'}^2} \right) \log^2 n. \tag{8}$$

This theorem is proved by drawing a connection between the weighted nuclear norm formulation (7) and the leverage scores (2) of the target matrix. Define the scaled matrix $\bar{M} := RMC$. Observe that the weighted program (7) is equivalent to first solving the following *unweighted* problem with scaled observations

$$\begin{aligned} \bar{X} = \arg\min_X \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = \bar{M}_{ij}, \text{ for } (i,j) \in \Omega, \end{aligned} \tag{9}$$

and then rescaling the solution $\bar{X}$ to return $\hat{X} = R^{-1}\bar{X}C^{-1}$. In other words, through the use of the weighted nuclear norm, we convert the problem of completing M to that of completing the scaled matrix $\bar{M}$. This leads to the following observation, which underlines the proof of Theorem 7:

> *If we can choose the weights R and C in such a way that the leverage scores of scaled matrix $\bar{M}$, denoted as $\bar{\mu}_i := \mu_i(\bar{M}), \bar{\nu}_j := \nu_i(\bar{M}), i, j \in [n]$, are aligned with the given non-uniform observations in a way that roughly satisfies the relation* (3)*, then we gain in sample complexity compared to the unweighted approach.*

We now quantify this observation more precisely for a particular class of matrix completion problems.

5.1 Comparison to Unweighted Nuclear Norm.

Assume for simplicity $n_1 = n_2 = n$ and $n/(\mu_0 r)$ is an integer. Suppose the sampling probabilities have a product form: $p_{ij} = p_i^{\mathrm{r}} p_j^{\mathrm{c}}$, with $p_1^{\mathrm{r}} \le p_2^{\mathrm{r}} \le \cdots \le p_n^{\mathrm{r}}$ and $p_1^{\mathrm{c}} \le p_2^{\mathrm{c}} \le \cdots \le p_n^{\mathrm{c}}$. If we choose $R_i = \sqrt{\frac{1}{n} p_i^{\mathrm{r}} \sum_{j'} p_{j'}^{\mathrm{c}}}$ and $C_j = \sqrt{\frac{1}{n} p_j^{\mathrm{c}} \sum_{i'} p_{i'}^{\mathrm{r}}}$—which is suggested by the condition (8)—Theorem 7 asserts that the following set of conditions are sufficient for recovery of M with high probability:

$$p_j^{\mathrm{c}} \cdot \left(\frac{\mu_0 r}{n} \sum_{i=1}^{n/(\mu_0 r)} p_i^{\mathrm{r}} \right) \gtrsim \frac{\mu_0 r}{n} \log^2 n, \ \forall j; \qquad p_i^{\mathrm{r}} \cdot \left(\frac{\mu_0 r}{n} \sum_{j=1}^{n/(\mu_0 r)} p_j^{\mathrm{c}} \right) \gtrsim \frac{\mu_0 r}{n} \log^2 n, \ \forall i. \tag{10}$$

We can compare the above condition to the condition for the unweighted approach: by Theorem 2, the unweighted nuclear norm minimization formulation (1) recovers M if

$$p_i^{\mathrm{r}} \cdot p_j^{\mathrm{c}} \gtrsim \frac{\mu_0 r}{n} \log^2 n, \quad \forall i, j. \tag{11}$$

Therefore, the weighted nuclear norm approach succeeds under less restrictive conditions: the condition (11) for the unweighted approach requires a lower bound on *minimum* sampling probability over the rows and columns, whereas the condition (10) for the weighted approach involves the *average* sampling probability of the $n/(\mu_0 r)$ least sampled rows/columns. This benefit is most significant precisely when the observed samples are very non-uniformly distributed.

We provide a concrete example of the gain of weighting in Section E.

Our theoretical results are consistent with the empirical study in Srebro and Salakhutdinov (2010); Foygel et al. (2011), which demonstrate the advantage of the weighted approach with the weights R and C chosen as above (using the empirical sampling distribution). We remark that Theorem 7 is the first exact recovery guarantee for weighted nuclear norm minimization. It provides a theoretical explanation, complementary to those in Srebro and Salakhutdinov (2010); Foygel et al. (2011); Negahban and Wainwright (2012), for why the weighted approach is advantageous over the unweighted approach for non-uniform observations. It also serves as a testament to the power of Theorem 2 as a general result on the relationship between sampling and the coherence/leverage score structure of a matrix.

In Theorem 7 and the discussion above we assume the underlying matrix M is incoherent. Clearly, one can still use the weighted nuclear norm approach when M is coherent: as long as the weights are chosen such that the leverage scores of the scaled matrix $\bar{M}$ are aligned with the distributions of the revealed entries, Theorem 2 can be applied and we expect improvements of the recovery performance using the weighted approach. How to choose the weights in this setting, and how it affects the performance, are left to future work.

6. Conclusion

In this paper we study the problem of matrix completion with no assumptions on the incoherence of the underlying matrix. We show that if the sampling of entries suitably depends on leverage scores of the matrix, then it can be recovered from $O(nr \log^2(n))$ entries using nuclear norm minimization. We further establish the necessity of leverage score sampling within the class of location invariant sampling distributions. Based on these results, we present a new two-phase sampling algorithm which does not require knowledge of underlying structure of the matrix and provide simulation results to verify its performance. As a corollary of our main theorem, we provide exact recovery guarantees for the weighted nuclear norm minimization approach when the observed entries are given and distributed non-uniformly.

It is an interesting open problem to provide rigorous theoretical analysis of the number of samples needed by the two-phase sampling algorithm. It is also of interest to develop and analyze algorithms that sample with more stages and iteratively improve the leverage score estimates. More generally, it is useful to develop and study other methods for estimating/adjusting the leverage scores and tuning the sampling procedure. Extending the

results in this paper to other low-rank recovery settings and applications will be of great value.

Acknowledgments

We would like to thank Petros Drineas, Michael Mahoney and Aarti Singh for helpful discussions, and the anonymous reviewers for their insightful comments and suggestions. Y. Chen was supported by NSF grant CIF-31712-23800 and ONR MURI grant N00014-11-1-0688. R. Ward was supported in part by an NSF CAREER award, AFOSR Young Investigator Program award, and ONR Grant N00014-12-1-0743. S. Sanghavi would like to acknowledge NSF grants 1302435, 1320175 and 0954059 for supporting this work.

Appendix A. Proof of Theorem 2

We prove our main result Theorem 2 in this section. The overall outline of the proof is a standard convex duality argument. The main difference in establishing our results is that, while other proofs relied on bounding the ℓ_∞ norm of certain random matrices, we instead bound the weighted $\ell_{\infty,2}$, norm (to be defined).

The high level road map of the proof is a standard one: by convex analysis, to show that M is the unique optimal solution to (1), it suffices to construct a *dual certificate* Y obeying certain sub-gradient optimality conditions. One of the conditions requires the spectral norm $\|Y\|$ to be small. Previous work bounds $\|Y\|$ by the the ℓ_∞ norm $\|Y'\|_\infty := \sum_{i,j} \left|Y'_{ij}\right|$ of a certain matrix Y', which gives rise to the standard and joint incoherence conditions involving uniform bounds by μ_0 and μ_{str}. Here, we derive a new bound using the weighted $\ell_{\infty,2}$ norm of Y', which is the maximum of the weighted row and column norms of Y'. These bounds lead to a tighter bound of $\|Y\|$ and hence less restrictive conditions for matrix completion.

We now turn to the details. To simplify the notion, we prove the results for square matrices ($n_1 = n_2 = n$). The results for non-square matrices are proved in exactly the same fashion. In the sequel by *with high probability* (*w.h.p.*) we mean with probability at least $1 - n^{-20}$. The proof below involves no more than $5n^{10}$ random events, each of which will be shown to hold with high probability. It follows from the union bound that all the events simultaneously hold with probability at least $1 - 5n^{-10}$, which is the success probability in the statement of Theorem 2.

A few additional notations are needed. We drop the dependence of $\mu_i(M)$ and $\nu_j(M)$ on M and simply use μ_i and ν_j. We use c and its derivatives (c', c_0, etc.) for universal positive constants, which may differ from place to place. The inner product between two matrices is given by $\langle Y, Z\rangle = \mathrm{trace}(Y^\top Z)$. Recall that U and V are the left and right singular vectors of the underlying matrix M. We need several standard projection operators for matrices. The projections P_T and $P_{T^\perp}$ are given by

$$P_T(Z) := UU^\top Z + ZVV^\top - UU^\top VZZ^\top$$

and $P_{T^\perp}(Z) := Z - P_T(Z)$. $P_\Omega(Z)$ is the matrix with $(P_\Omega(Z))_{ij} = Z_{ij}$ if $(i,j) \in \Omega$ and zero otherwise, and $P_{\Omega^c}(Z) := Z - P_\Omega(Z)$. As usual, $\|z\|_2$ is the ℓ_2 norm of the vector z, and $\|Z\|_F$ and $\|Z\|$ are the Frobenius norm and spectral norm of the matrix

Z, respectively. For a linear operator R on matrices, its operator norm is defined as $\|R\|_{op} = \sup_{X \in \mathbb{R}^{n\times n}} \|R(X)\|_F / \|X\|_F$. For each $1 \le i, j \le n$, we define the random variable $\delta_{ij} := \mathbb{I}((i,j) \in \Omega)$, where $\mathbb{I}(\cdot)$ is the indicator function. The matrix operator $R_\Omega : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ is defined as

$$R_\Omega(Z) = \sum_{i,j} \frac{1}{p_{ij}} \delta_{ij} \left\langle e_i e_j^\top, Z \right\rangle e_i e_j^\top. \tag{12}$$

A.1 Optimality Condition

Following our proof road map, we now state a sufficient condition for M to be the unique optimal solution to the optimization problem (1). This is the content of Proposition 8 below (proved in Section A.7 to follow).

Proposition 8 *Suppose $p_{ij} \ge \frac{1}{n^{10}}$. The matrix M is the unique optimal solution to* (1) *if the following conditions hold.*

1. $\|P_T R_\Omega P_T - P_T\|_{op} \le \frac{1}{2}$.
2. *There exists a dual certificate $Y \in \mathbb{R}^{n\times n}$ which satisfies $P_\Omega(Y) = Y$ and*
   (a) $\left\|P_T(Y) - UV^\top\right\|_F \le \frac{1}{4n^5}$,
   (b) $\|P_{T^\perp}(Y)\| \le \frac{1}{2}$.

A.2 Validating the Optimality Condition

We begin by proving that Condition 1 in Proposition 8 is satisfied under the conditions of Theorem 2. This is done in the following lemma, which is proved in Section A.8 to follow. The lemma shows that R_Ω is close to the identity operator on T.

Lemma 9 *If $p_{ij} \ge \min\{c_0 \frac{(\mu_i+\nu_j)r}{n} \log n, 1\}$ for all (i,j) and a sufficiently large c_0, then w.h.p.*

$$\|P_T R_\Omega P_T - P_T\|_{op} \le \frac{1}{2}. \tag{13}$$

A.3 Constructing the Dual Certificate

It remains to construct a matrix Y (the dual certificate) that satisfies Condition 2 in Proposition 8. We do this using the golfing scheme (Gross, 2011; Candès et al., 2011). Set $k_0 := 20 \log n$. For each $k = 1, \ldots, k_0$, let $\Omega_k \subseteq \mathbb{R}^{n\times n}$ be a random set of matrix elements such that for each (i,j), $\mathbb{P}[(i,j) \in \Omega_k] = q_{ij} := 1 - (1-p_{ij})^{1/k_0}$, independently of all others. We may assume that the set Ω of observed elements is generated as $\Omega = \bigcup_{k=1}^{k_0} \Omega_k$, which is equivalent to the original Bernoulli sampling model. Let $W_0 := 0$ and for $k = 1, \ldots, k_0$,

$$W_k := W_{k-1} + R_{\Omega_k} P_T (UV^\top - P_T W_{k-1}), \tag{14}$$

where the operator R_{Ω_k} is given by

$$R_{\Omega_k}(Z) = \sum_{i,j} \frac{1}{q_{ij}} \mathbb{I}((i,j) \in \Omega_k) \left\langle e_i e_j^\top, Z \right\rangle e_i e_j^\top.$$

The dual certificate is given $Y := W_{k_0}$. Clearly $P_\Omega(Y) = Y$ by construction. The proof of Theorem 2 is completed if we show that under the condition in the theorem, Y satisfies Conditions 2(a) and 2(b) in Proposition 8 w.h.p.

A.4 Concentration Properties

The key step in our proof is to show that Y satisfies Condition 2(b) in Proposition 8, i.e., we need to bound $\|P_{T^\perp}(Y)\|$. Here our proof departs from existing ones, as we establish concentration bounds on this quantity in terms of (an appropriately weighted version of) the $\ell_{\infty,2}$ norm, which we now define. The $\mu(\infty,2)$-norm of a matrix $Z \in \mathbb{R}^{n\times n}$ is defined as

$$\|Z\|_{\mu(\infty,2)} := \max\left\{ \max_i \sqrt{\frac{n}{\mu_i r}\sum_b Z_{ib}^2}, \max_j \sqrt{\frac{n}{\nu_j r}\sum_a Z_{aj}^2} \right\},$$

which is the maximum of the weighted column and row norms of Z. We also need the $\mu(\infty)$-norm of Z, which is a weighted version of the matrix ℓ_∞ norm. This is given as

$$\|Z\|_{\mu(\infty)} := \max_{i,j} |Z_{ij}| \sqrt{\frac{n}{\mu_i r}}\sqrt{\frac{n}{\nu_j r}}.$$

which is the weighted element-wise magnitude of Z. We now state three new lemmas concerning the concentration properties of these norms. The first lemma is crucial to our proof; it bounds the spectral norm of $(R_\Omega - I)\, Z$ in terms of the $\mu(\infty,2)$ and $\mu(\infty)$ norms of Z. This obviates the intermediate lemmas required by previous approaches (Candès and Tao, 2010; Gross, 2011; Recht, 2011; Keshavan et al., 2010) which use the ℓ_∞ norm of Z.

Lemma 10 *Suppose Z is a fixed $n \times n$ matrix. For some universal constant $c > 1$, we have w.h.p.*

$$\|(R_\Omega - I)\, Z\| \le c\left(\max_{i,j}\left|\frac{Z_{ij}}{p_{ij}}\right| \log n + \sqrt{\max\left\{ \max_i \sum_{j=1}^n \frac{Z_{ij}^2}{p_{ij}}, \max_j \sum_{i=1}^n \frac{Z_{ij}^2}{p_{ij}} \right\} \log n} \right).$$

If $p_{ij} \ge \min\{c_0 \frac{(\mu_i+\nu_j)r}{n} \log n, 1\}$ for all (i,j) and a sufficiently large constant c_0, then we further have w.h.p.

$$\|(R_\Omega - I)\, Z\| \le \frac{c}{\sqrt{c_0}} \left(\|Z\|_{\mu(\infty)} + \|Z\|_{\mu(\infty,2)} \right).$$

The next two lemmas further control the $\mu(\infty,2)$ and $\mu(\infty)$ norms of a matrix after certain random transformation.

Lemma 11 *Suppose Z is a fixed $n \times n$ matrix. If $p_{ij} \ge \min\{c_0 \frac{(\mu_i+\nu_j)r}{n} \log n, 1\}$ for all i, j and a sufficiently large constant c_0, then w.h.p.*

$$\|(P_T R_\Omega - P_T)Z\|_{\mu(\infty,2)} \le \frac{1}{2}\left(\|Z\|_{\mu(\infty)} + \|Z\|_{\mu(\infty,2)} \right)$$

Lemma 12 *Suppose Z is a fixed $n \times n$ matrix. If $p_{ij} \geq \min\{c_0 \frac{(\mu_i+\nu_j)r}{n} \log n, 1\}$ for all i, j and a sufficiently large constant c_0, then w.h.p.*

$$\|(P_T R_\Omega - P_T)\, Z\|_{\mu(\infty)} \leq \frac{1}{2} \|Z\|_{\mu(\infty)}\,.$$

We prove Lemmas 10–12 in Section A.8. Equipped with the three lemmas above, we are now ready to validate that Y satisfies Condition 2 in Proposition 8.

A.5 Validating Condition 2(a)

Set $\Delta_k = UV^\top - P_T(W_k)$ for $k = 1, \ldots, k_0$; note that $\Delta_{k_0} = UV^\top - P_T(Y)$. By definition of W_k, we have

$$\Delta_k = (P_T - P_T R_{\Omega_k} P_T)\, \Delta_{k-1}. \tag{15}$$

Note that Ω_k is independent of Δ_{k-1} and $q_{ij} \geq p_{ij}/k_0 \geq c_0'(\mu_i + \nu_j) r \log(n)/n$ under the condition in Theorem 2. Applying Lemma 9 with Ω replaced by Ω_k , we obtain that w.h.p.

$$\|\Delta_k\|_F \leq \|P_T - P_T R_{\Omega_k} P_T\|\, \|\Delta_{k-1}\|_F \leq \frac{1}{2} \|\Delta_{k-1}\|_F\,.$$

Applying the above inequality recursively with $k = k_0, k_0 - 1, \ldots, 1$ gives

$$\left\|P_T(Y) - UV^\top\right\|_F = \|\Delta_{k_0}\|_F \leq \left(\frac{1}{2}\right)^{k_0} \left\|UV^\top\right\|_F \leq \frac{1}{4n^6} \cdot \sqrt{r} \leq \frac{1}{4n^5},$$

where we use our definition of k_0 and $\left\|UV^\top\right\|_F = \sqrt{r}$ in the second inequality.

A.6 Validating Condition 2(b)

By definition, Y can be rewritten as $Y = \sum_{k=1}^{k_0} R_{\Omega_k} P_T \Delta_{k-1}$. It follows that

$$\|P_{T^\perp}(Y)\| = \left\|P_{T^\perp} \sum_{k=1}^{k_0} (R_{\Omega_k} P_T - P_T)\, \Delta_{k-1}\right\| \leq \sum_{k=1}^{k_0} \|(R_{\Omega_k} - I)\, \Delta_{k-1}\|\,.$$

We apply Lemma 10 with Ω replaced by Ω_k to each summand in the last RHS to obtain w.h.p.

$$\|P_{T^\perp}(Y)\| \leq \frac{c}{\sqrt{c_0}} \sum_{k=1}^{k_0} \|\Delta_{k-1}\|_{\mu(\infty)} + \frac{c}{\sqrt{c_0}} \sum_{k=1}^{k_0} \|\Delta_{k-1}\|_{\mu(\infty,2)}\,. \tag{16}$$

We bound each summand in the last RHS. Applying $(k-1)$ times (15) and Lemma 12 (with Ω replaced by Ω_k), we have w.h.p.

$$\|\Delta_{k-1}\|_{\mu(\infty)} = \left\|\left(P_T - P_T R_{\Omega_{k-1}} P_T\right) \Delta_{k-2}\right\|_{\mu(\infty)} \leq \left(\frac{1}{2}\right)^{k-1} \left\|UV^\top\right\|_{\mu(\infty)}\,.$$

for each k. Similarly, repeatedly applying (15), Lemma 11 and the inequality we just proved above, we obtain w.h.p.

$$\begin{aligned}
&\|\Delta_{k-1}\|_{\mu(\infty,2)} && (17)\\
=&\left\|\left(P_T - P_T R_{\Omega_{k-1}} P_T\right)\Delta_{k-2}\right\|_{\mu(\infty,2)} && (18)\\
\leq&\frac{1}{2}\|\Delta_{k-2}\|_{\mu(\infty)} + \frac{1}{2}\|\Delta_{k-2}\|_{\mu(\infty,2)} && (19)\\
\leq&\left(\frac{1}{2}\right)^{k-1}\left\|UV^\top\right\|_{\mu(\infty)} + \frac{1}{2}\|\Delta_{k-2}\|_{\mu(\infty,2)} && (20)\\
\leq&k\left(\frac{1}{2}\right)^{k-1}\left\|UV^\top\right\|_{\mu(\infty)} + \left(\frac{1}{2}\right)^{k-1}\|UV\|_{\mu(\infty,2)}\,. && (21)
\end{aligned}$$

It follows that w.h.p.

$$\begin{aligned}
\|P_{T^\perp}(Y)\| &\leq \frac{c}{\sqrt{c_0}}\sum_{k=1}^{k_0}(k+1)\left(\frac{1}{2}\right)^{k-1}\left\|UV^\top\right\|_{\mu(\infty)} + \frac{c}{\sqrt{c_0}}\sum_{k=1}^{k_0}\left(\frac{1}{2}\right)^{k-1}\left\|UV^\top\right\|_{\mu(\infty,2)} && (22)\\
&\leq \frac{6c}{\sqrt{c_0}}\left\|UV^\top\right\|_{\mu(\infty)} + \frac{2c}{\sqrt{c_0}}\left\|UV^\top\right\|_{\mu(\infty,2)}\,. && (23)
\end{aligned}$$

Note that for all (i,j), we have $\left|\left(UV^\top\right)_{ij}\right| = \left|e_i^\top UV^\top e_j\right| \leq \sqrt{\frac{\mu_i r}{n}}\sqrt{\frac{\nu_j r}{n}}$, $\left\|e_i^\top UV^\top\right\|_2 = \sqrt{\frac{\mu_i r}{n}}$ and $\left\|UV^\top e_j\right\|_2 = \sqrt{\frac{\nu_j r}{n}}$. Hence $\left\|UV^\top\right\|_{\mu(\infty)} \leq 1$ and $\left\|UV^\top\right\|_{\mu(\infty,2)} = 1$. We conclude that

$$\|P_{T^\perp}(Y)\| \leq \frac{6c}{\sqrt{c_0}} + \frac{2c}{\sqrt{c_0}} \leq \frac{1}{2}$$

provided that the constant c_0 in Theorem 2 is sufficiently large. This completes the proof of Theorem 2.

A.7 Proof of Proposition 8 (Optimality Condition)

Proof Consider any feasible solution X to (1) with $P_\Omega(X) = P_\Omega(M)$. Let G be an $n \times n$ matrix which satisfies $\|P_{T^\perp}G\| = 1$, and $\langle P_{T^\perp}G, P_{T^\perp}(X-M)\rangle = \|P_{T^\perp}(X-M)\|_*$. Such G always exists by duality between the nuclear norm and spectral norm. Because $UV^\top + P_{T^\perp}G$ is a sub-gradient of the function $f(Z) = \|Z\|_*$ at $Z = M$, we have

$$\|X\|_* - \|M\|_* \geq \left\langle UV^\top + P_{T^\perp}G, X - M\right\rangle. \tag{24}$$

But $\langle Y, X-M\rangle = \langle P_\Omega(Y), P_\Omega(X-M)\rangle = 0$ since $P_\Omega(Y) = Y$. It follows that

$$\begin{aligned}
&\|X\|_* - \|M\|_* \geq \left\langle UV^\top + P_{T^\perp}G - Y, X - M\right\rangle\\
&= \|P_{T^\perp}(X-M)\|_* + \left\langle UV^\top - P_T Y, X - M\right\rangle - \langle P_{T^\perp}Y, X - M\rangle\\
&\geq \|P_{T^\perp}(X-M)\|_* - \left\|UV^\top - P_T Y\right\|_F \|P_T(X-M)\|_F - \|P_{T^\perp}Y\|\,\|P_{T^\perp}(X-M)\|_*\\
&\geq \frac{1}{2}\|P_{T^\perp}(X-M)\|_* - \frac{1}{4n^5}\|P_T(X-M)\|_F\,,
\end{aligned}$$

where in the last inequality we use conditions $2a$ and $2b$ in the proposition. Using Lemma 13 below, we obtain

$$\|X\|_* - \|M\|_* \geq \frac{1}{2}\|P_{T^\perp}(X-M)\|_* - \frac{1}{4n^5}\cdot\sqrt{2}n^5\|P_{T^\perp}(X-M)\|_* > \frac{1}{8}\|P_{T^\perp}(X-M)\|_*.$$

The RHS is strictly positive for all X with $P_\Omega(X-M)=0$ and $X \neq M$. Otherwise we must have $P_T(X-M) = X-M$ and $P_T P_\Omega P_T(X-M) = 0$, contradicting the assumption $\|P_T R_\Omega P_T - P_T\|_{op} \leq \frac{1}{2}$. This proves that M is the unique optimum. ■

Lemma 13 *If $p_{ij} \geq \frac{1}{n^{10}}$ for all (i,j) and $\|P_T R_\Omega P_T - P_T\|_{op} \leq \frac{1}{2}$, then we have*

$$\|P_T Z\|_F \leq \sqrt{2}n^5\|P_{T^\perp}(Z)\|_*, \forall Z \in \{Z' : P_\Omega(Z') = 0\}. \tag{25}$$

Proof Define the operator $R_\Omega^{1/2} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ by

$$R_\Omega^{1/2}(Z) := \sum_{i,j}\frac{1}{\sqrt{p_{ij}}}\delta_{ij}\left\langle e_i e_j^\top, Z\right\rangle e_i e_j^\top.$$

Note that $R_\Omega^{1/2}$ is self-adjoint and satisfies $R_\Omega^{1/2}R_\Omega^{1/2} = R_\Omega$. Hence we have

$$\begin{aligned}
\left\|R_\Omega^{1/2}P_T(Z)\right\|_F &= \sqrt{\langle P_T R_\Omega P_T Z, P_T Z\rangle} \\
&= \sqrt{\langle (P_T R_\Omega P_T - P_T)Z, P_T(Z)\rangle + \langle P_T(Z), P_T(Z)\rangle} \\
&\geq \sqrt{\|P_T(Z)\|_F^2 - \|P_T R_\Omega P_T - P_T\|\,\|P_T(Z)\|_F^2} \\
&\geq \frac{1}{\sqrt{2}}\|P_T(Z)\|_F,
\end{aligned}$$

where the last inequality follows from the assumption $\|P_T R_\Omega P_T - P_T\|_{op} \leq \frac{1}{2}$. On the other hand, $P_\Omega(Z) = 0$ implies $0 = R_\Omega^{1/2}(Z) = R_\Omega^{1/2}P_T(Z) + R_\Omega^{1/2}P_{T^\perp}(Z)$ and thus

$$\left\|R_\Omega^{1/2}P_T(Z)\right\|_F = \left\|-R_\Omega^{1/2}P_{T^\perp}(Z)\right\|_F \leq \left(\max_{i,j}\frac{1}{\sqrt{p_{ij}}}\right)\|P_{T^\perp}(Z)\|_F \leq n^5\|P_{T^\perp}(Z)\|_F.$$

Combining the last two display equations gives

$$\|P_T(Z)\|_F \leq \sqrt{2}n^5\|P_{T^\perp}(Z)\|_F \leq \sqrt{2}n^5\|P_{T^\perp}(Z)\|_*.$$

■

A.8 Proof of Technical Lemmas

We prove the four technical lemmas that are used in the proof of our main theorem. The proofs use the matrix Bernstein inequality given as Theorem 16 in Section F. We also make frequent use of the following facts: for all i and j, we have $\max\left\{\frac{\mu_i r}{n}, \frac{\nu_j r}{n}\right\} \leq 1$ and

$$\frac{(\mu_i+\nu_j)r}{n} \geq \left\|P_T(e_i e_j^\top)\right\|_F^2. \tag{26}$$

We also use the shorthand $a \wedge b := \min\{a,b\}$.

A.8.1 Proof of Lemma 9

For any matrix Z, we can write

$$(P_T R_\Omega P_T - P_T)(Z) = \sum_{i,j} \left(\frac{1}{p_{ij}} \delta_{ij} - 1 \right) \left\langle e_i e_j^\top, P_T(Z) \right\rangle P_T(e_i e_j^\top) =: \sum_{i,j} \mathcal{S}_{ij}(Z).$$

Note that $\mathbb{E}\left[\mathcal{S}_{ij}\right] = 0$ and $\mathcal{S}_{ij}$'s are independent of each other. For all Z and (i,j), we have $\mathcal{S}_{ij} = 0$ if $p_{ij} = 1$. On the other hand, when $p_{ij} \geq c_0 \frac{(\mu_i + \nu_j) r \log n}{n}$, then it follows from (26) that

$$\left\| \mathcal{S}_{ij}(Z) \right\|_F \leq \frac{1}{p_{ij}} \left\| P_T(e_i e_j^\top) \right\|_F^2 \|Z\|_F \leq \max_{i,j} \left\{ \frac{1}{p_{ij}} \frac{(\mu_i + \nu_j) r}{n} \right\} \|Z\|_F \leq \frac{1}{c_0 \log n} \|Z\|_F .$$

Putting together, we have that $\|\mathcal{S}_{ij}\| \leq \frac{1}{c_0 \log n}$ under the condition of the lemma. On the other hand, we have

$$\begin{aligned}
\left\| \sum_{i,j} \mathbb{E}\left[\mathcal{S}_{ij}^2(Z) \right] \right\|_F &= \left\| \sum_{i,j} \mathbb{E}\left[\left(\frac{1}{p_{ij}} \delta_{ij} - 1 \right)^2 \left\langle e_i e_j^\top, P_T(Z) \right\rangle \left\langle e_i e_j^\top, P_T(e_i e_j^\top) \right\rangle P_T(e_i e_j^\top) \right] \right\|_F \\
&\leq \left(\max_{i,j} \frac{1 - p_{ij}}{p_{ij}} \left\| P_T(e_i e_j^\top) \right\|_F^2 \right) \left\| \sum_{i,j} \left\langle e_i e_j^\top, P_T(Z) \right\rangle P_T(e_i e_j^\top) \right\|_F \\
&\leq \max_{i,j} \left\{ \frac{1 - p_{ij}}{p_{ij}} \frac{(\mu_i + \nu_j) r}{n} \right\} \|P_T(Z)\|_F ,
\end{aligned}$$

This implies $\left\| \sum_{i,j} \mathbb{E}\left[\mathcal{S}_{ij}^2 \right] \right\| \leq \frac{1}{c_0 \log n}$ under the condition of the lemma. Applying the Matrix Bernstein inequality (Theorem 16), we obtain $\|P_T R_\Omega P_T - P_T\| = \left\| \sum_{i,j} \mathcal{S}_{ij} \right\| \leq \frac{1}{2}$ w.h.p. for sufficiently large c_0.

A.8.2 Proof of Lemma 10

We can write $(R_\Omega - I)\, Z$ as the sum of independent matrices:

$$(R_\Omega - I)\, Z = \sum_{i,j} \left(\frac{1}{p_{ij}} \delta_{ij} - 1 \right) Z_{ij} e_i e_j^\top =: \sum_{i,j} S_{ij}.$$

Note that $\mathbb{E}[S_{ij}] = 0$. For all (i,j), we have $S_{ij} = 0$ if $p_{ij} = 1$, and

$$\|S_{ij}\| \leq \frac{1}{p_{ij}} \left| Z_{ij} \right| .$$

Moreover, we have

$$\left\| \mathbb{E}\left[\sum_{i,j} S_{ij}^\top S_{ij} \right] \right\| = \left\| \sum_{i,j} Z_{ij}^2 e_i e_j^\top e_j e_i^\top \mathbb{E}\left(\frac{1}{p_{ij}} \delta_{ij} - 1 \right)^2 \right\| = \max_i \sum_{j=1}^n \frac{1 - p_{ij}}{p_{ij}} Z_{ij}^2.$$

The quantity $\left\|\mathbb{E}\left[\sum_{i,j} S_{ij}S_{ij}^\top\right]\right\|$ is bounded by $\max_j \sum_{i=1}^n (1-p_{ij})Z_{ij}^2/p_{ij}$ in a similar way. The first part of the lemma then follows from the matrix Bernstein inequality (Theorem 16). If $p_{ij} \geq 1 \wedge \frac{c_0(\mu_i+\nu_j)r\log n}{n} \geq 1 \wedge 2c_0\sqrt{\frac{\mu_i r}{n}\cdot\frac{\nu_j r}{n}}\log n$, we have for all i and j,

$$\|S_{ij}\|\log n \leq (1-\mathbb{I}(p_{ij}=1))\,\frac{1}{p_{ij}}\,|Z_{ij}|\log n \leq \frac{1}{c_0}\,\|Z\|_{\mu(\infty)}\,,$$

$$\sum_{i=1}^n \frac{1-p_{ij}}{p_{ij}} Z_{ij}^2 \log n \leq \frac{1}{c_0}\,\|Z\|_{\mu(\infty,2)}^2\,,$$

$$\sum_{j=1}^n \frac{1-p_{ij}}{p_{ij}} Z_{ij}^2 \log n \leq \frac{1}{c_0}\,\|Z\|_{\mu(\infty,2)}^2\,.$$

The second part of the lemma follows again from applying the matrix Bernstein inequality.

A.8.3 Proof of Lemma 11

Let $X = (P_T R_\Omega - P_T)\,Z$. By definition we have

$$\|X\|_{\mu(\infty,2)} = \max_{a,b}\left\{\sqrt{\frac{n}{\mu_a r}}\,\|X_{a\cdot}\|_2\,, \sqrt{\frac{n}{\nu_b r}}\,\|X_{\cdot b}\|_2\right\},$$

where $X_{a\cdot}$ and $X_{\cdot b}$ are the a-th row and b-th column of of X, respectively. We bound each term in the maximum. Observe that $\sqrt{\frac{n}{\nu_b r}}X_{\cdot b}$ can be written as the sum of independent column vectors:

$$\sqrt{\frac{n}{\nu_b r}}X_{\cdot b} = \sum_{i,j}\left(\frac{1}{p_{ij}}\delta_{ij}-1\right)Z_{ij}\left(P_T(e_i e_j^\top)e_b\right)\sqrt{\frac{n}{\nu_b r}} =: \sum_{i,j} S_{ij},$$

where $\mathbb{E}\,[S_{ij}] = 0$. To control $\|S_{ij}\|_2$ and $\left\|\mathbb{E}\left[\sum_{i,j} S_{ij}^\top S_{ij}\right]\right\|$, we first need a bound for $\left\|P_T(e_i e_j^\top)e_b\right\|_2$. If $j = b$, we have

$$\left\|P_T(e_i e_j^\top)e_b\right\|_2 = \left\|UU^\top e_i + (I-UU^\top)e_i\left\|V^\top e_b\right\|_2^2\right\|_2 \leq \sqrt{\frac{\mu_i r}{n}} + \sqrt{\frac{\nu_b r}{n}}, \tag{27}$$

where we use the triangle inequality and the definition of μ_i and ν_b . Similarly, if $j \neq b$, we have

$$\left\|P_T(e_i e_j^\top)e_b\right\|_2 = \left\|(I-UU^\top)e_i e_j^\top VV^\top e_b\right\|_2 \leq \left|e_j^\top VV^\top e_b\right|. \tag{28}$$

Now note that $\|S_{ij}\|_2 \leq (1-\mathbb{I}(p_{ij}=1))\,\frac{1}{p_{ij}}\,|Z_{ij}|\,\sqrt{\frac{n}{\nu_b r}}\left\|P_T(e_i e_j^\top)e_b\right\|_2$. Using the bounds (27) and (28), we obtain that for $j = b$,

$$\begin{aligned}\|S_{ij}\|_2 &\leq (1-\mathbb{I}(p_{ij}=1))\,\frac{1}{p_{ib}}\,|Z_{ib}|\,\sqrt{\frac{n}{\nu_b r}}\cdot\left(\sqrt{\frac{\mu_i r}{n}}+\frac{\nu_b r}{n}\right)\\ &\leq \frac{2}{c_0\sqrt{\frac{\mu_i r\nu_b r}{n^2}}\log n}\,|Z_{ib}| \leq \frac{2}{c_0\log n}\,\|Z\|_{\mu(\infty)}\,,\end{aligned}$$

where we use $p_{ib} \geq 1 \wedge \frac{c_0 \mu_i r \log n}{n}$ and $p_{ib} \geq 1 \wedge c_0 \sqrt{\frac{\mu_i r}{n} \frac{\nu_b r}{n}} \log n$ in the second inequality. For $j \neq b$, we have

$$\|S_{ij}\|_2 \leq (1 - \mathbb{I}(p_{ij} = 1)) \frac{1}{p_{ij}} |Z_{ij}| \sqrt{\frac{n}{\nu_b r}} \cdot \sqrt{\frac{\nu_j r}{n}} \sqrt{\frac{\nu_b r}{n}} \leq \frac{2}{c_0 \log n} \|Z\|_{\mu(\infty)},$$

where we use $p_{ij} \geq 1 \wedge c_0 \sqrt{\frac{\mu_i r}{n} \frac{\nu_j r}{n}} \log n$. We thus obtain $\|S_{ij}\|_2 \leq \frac{2}{c_0 \log n} \|Z\|_{\mu(\infty)}$ for all (i, j).

On the other hand, note that

$$\begin{aligned} \left| \mathbb{E} \left[\sum_{i,j} S_{ij}^\top S_{ij} \right] \right| &= \left| \sum_{i,j} \mathbb{E} \left[\left(\frac{1}{p_{ij}} \delta_{ij} - 1 \right)^2 \right] Z_{ij}^2 \left\| P_T(e_i e_j^\top) e_b \right\|_2^2 \cdot \frac{n}{\nu_b r} \right| \\ &= \left(\sum_{j=b,i} + \sum_{j \neq b,i} \right) \frac{1 - p_{ij}}{p_{ij}} Z_{ij}^2 \left\| P_T(e_i e_j^\top) e_b \right\|_2^2 \cdot \frac{n}{\nu_b r}. \end{aligned}$$

Applying (27), we can bound the first sum by

$$\sum_{j=b,i} \leq \sum_i \frac{1 - p_{ib}}{p_{ib}} Z_{ib}^2 \cdot 2 \left(\frac{\mu_i r}{n} + \frac{\nu_b r}{n} \right) \cdot \frac{n}{\nu_b r} \leq \frac{2}{c_0 \log n} \frac{n}{\nu_b r} \|Z_{\cdot b}\|_2^2 \leq \frac{2}{c_0 \log n} \|Z\|_{\mu(\infty,2)}^2,$$

where we use $p_{ib} \geq 1 \wedge \frac{c_0(\mu_i + \nu_b) r}{n} \log n$ in the second inequality. The second sum can be bounded using (28):

$$\begin{aligned} \sum_{j \neq b,i} &\leq \sum_{j \neq b,i} \frac{1 - p_{ij}}{p_{ij}} Z_{ij}^2 \left| e_j^\top V V^\top e_b \right|^2 \frac{n}{\nu_b r} \\ &= \frac{n}{\nu_b r} \sum_{j \neq b} \left| e_j^\top V V^\top e_b \right|^2 \sum_i \frac{1 - p_{ij}}{p_{ij}} Z_{ij}^2 \\ &\overset{(a)}{\leq} \frac{n}{\nu_b r} \sum_{j \neq b} \left| e_j^\top V V^\top e_b \right|^2 \left(\frac{1}{c_0 \log n} \sum_i Z_{ij}^2 \frac{n}{\nu_j r} \right) \\ &\leq \left(\frac{1}{c_0 \log n} \|Z\|_{\mu(\infty,2)}^2 \right) \frac{n}{\nu_b r} \sum_{j \neq b} \left| e_j^\top V V^\top e_b \right|^2 \\ &\overset{(b)}{\leq} \frac{1}{c_0 \log n} \|Z\|_{\mu(\infty,2)}^2, \end{aligned}$$

where we use $p_{ij} \geq 1 \wedge \frac{c_0 \nu_j r \log n}{n}$ in (a) and $\sum_{j \neq b} \left| e_j^\top V V^\top e_b \right|^2 \leq \left\| V V^\top e_b \right\|_2^2 \leq \frac{\nu_b r}{n}$ in (b). Combining the bounds for the two sums, we obtain $\left\| \mathbb{E} \left[\sum_{i,j} S_{ij}^\top S_{ij} \right] \right\| \leq \frac{3}{c_0 \log n} \|Z\|_{\mu(\infty,2)}^2$. We can bound $\left\| \mathbb{E} \left[\sum_{i,j} S_{ij} S_{ij}^\top \right] \right\|$ in a similar way. Applying the Matrix Bernstein inequality in Theorem 16, we have w.h.p.

$$\left\| \sqrt{\frac{n}{\nu_b r}} X_{\cdot b} \right\|_2 = \left\| \sum_{i,j} S_{ij} \right\|_2 \leq \frac{1}{2} \left(\|Z\|_{\mu(\infty)} + \|Z\|_{\mu(\infty,2)} \right)$$

for c_0 sufficiently large. Similarly we can bound $\left\| \sqrt{\frac{n}{\mu_a r}} X_{a \cdot} \right\|_2$ by the same quantity. We take a union bound over all a and b to obtain the desired results.

A.8.4 Proof of Lemma 12

Fix a matrix index (a,b) and let $w_{ab} = \sqrt{\frac{\mu_a r}{n}\frac{\nu_b r}{n}}$. We can write

$$[(P_T R_\Omega - P_T)\, Z]_{ab} \sqrt{\frac{n}{\mu_a r}}\sqrt{\frac{n}{\nu_b r}} = \sum_{i,j}\left(\frac{1}{p_{ij}}\delta_{ij} - 1\right) Z_{ij} \left\langle e_i e_j^\top, P_T(e_a e_b^\top)\right\rangle \frac{1}{w_{ab}} =: \sum_{i,j} s_{ij},$$

which is the sum of independent zero-mean variables. We first compute the following bound:

$$\begin{aligned}
&\left|\left\langle e_i e_j^\top, P_T(e_a e_b^\top)\right\rangle\right| \\
&= \left|e_i^\top U U^\top e_a e_b^\top e_j + e_i^\top (I - UU^\top) e_a e_b^\top V V^\top e_j\right| \\
&= \begin{cases}
\left|e_a^\top U U^\top e_a + e_a^\top (I-UU^\top) e_a e_b^\top V V^\top e_b\right| \le \frac{\mu_a r}{n} + \frac{\nu_b r}{n}, & i = a, j = b, \\
\left|e_a^\top (I-UU^\top) e_a e_b^\top V V^\top e_j\right| \le \left|e_b^\top V V^\top e_j\right|, & i = a, j \neq b, \\
\left|e_i^\top U U^\top e_a e_b^\top (I-VV^\top) e_b\right| \le \left|e_i^\top U U^\top e_a\right|, & i \neq a, j = b, \\
\left|e_i^\top U U^\top e_a e_b^\top V V^\top e_j\right| \le \left|e_i^\top U U^\top e_a\right| \left|e_b^\top V V^\top e_j\right|, & i \neq a, j \neq b,
\end{cases}
\end{aligned} \tag{29}$$

where we use the fact that the matrices $I - UU^\top$ and $I - VV^\top$ have spectral norm at most 1. We proceed to bound $|s_{ij}|$. Note that

$$|s_{ij}| \le (1 - \mathbb{I}(p_{ij} = 1))\, \frac{1}{p_{ij}} \cdot |Z_{ij}| \cdot \left|\left\langle e_i e_j^\top, P_T(e_a e_b^\top)\right\rangle\right| \cdot \frac{1}{w_{ab}}.$$

We distinguish four cases. When $i = a$ and $j = b$, we use (29) and $p_{ab} \ge 1 \wedge \frac{c_0(\mu_a+\nu_b) r \log^2(n)}{n}$ to obtain $|s_{ij}| \le |Z_{ij}| / (w_{ij} c_0 \log n) \le \|Z\|_{\mu(\infty)} / (c_0 \log n)$. When $i = a$ and $j \neq b$, we apply (29) to get

$$|s_{ij}| \le (1 - \mathbb{I}(p_{ij} = 1))\, \frac{|Z_{aj}|}{p_{aj}} \cdot \sqrt{\frac{\nu_b r}{n}\frac{\nu_j r}{n}} \cdot \sqrt{\frac{n}{\mu_a r}\frac{n}{\nu_b r}} \overset{(a)}{\le} |Z_{aj}| \cdot \sqrt{\frac{n}{\mu_a r}\frac{n}{\nu_j r}}\frac{1}{c_0 \log n} \le \frac{\|Z\|_{\mu(\infty)}}{c_0 \log n},$$

where (a) follows from $p_{aj} \ge \min\left\{c_0 \frac{\nu_j r \log n}{n}, 1\right\}$. In a similar fashion, we can show that the same bound holds when $i \neq a$ and $j = b$. When $i \neq a$ and $j \neq b$, we use (29) to get

$$\begin{aligned}
|s_{ij}| &\le (1 - \mathbb{I}(p_{ij} = 1))\, \frac{|Z_{ij}|}{p_{ij}} \cdot \sqrt{\frac{\mu_i r}{n}\frac{\mu_a r}{n}}\sqrt{\frac{\nu_b r}{n}\frac{\nu_j r}{n}} \cdot \sqrt{\frac{n}{\mu_a r}\frac{n}{\nu_b r}} \\
&\overset{(b)}{\le} |Z_{ij}| \cdot \sqrt{\frac{n}{\mu_i r}\frac{n}{\nu_j r}}\frac{1}{c_0 \log n} \le \frac{\|Z\|_{\mu(\infty)}}{c_0 \log n},
\end{aligned}$$

where (b) follows from $p_{ij} \ge 1 \wedge c_0 \sqrt{\frac{\mu_i r}{n}\frac{\nu_j r}{n}} \log n$ and $\max\left\{\sqrt{\frac{\mu_i r}{n}}, \sqrt{\frac{\nu_j r}{n}}\right\} \le 1$. We conclude that $|s_{ij}| \le \|Z\|_{\mu(\infty)} / (c_0 \log n)$ for all (i,j).

On the other hand, note that

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{i,j} s_{ij}^2\right]\right| &= \sum_{i,j} \mathbb{E}\left[\left(\frac{1}{p_{ij}}\delta_{ij} - 1\right)^2\right] \frac{Z_{ij}^2}{w_{ab}^2} \left\langle e_i e_j^\top, P_T(e_a e_b^\top)\right\rangle^2 \\
&= \sum_{i=a, j=b} + \sum_{i=a, j\neq b} + \sum_{i\neq a, j=b} + \sum_{i\neq a, j\neq b}.
\end{aligned}$$

We bound each of the four sums. By (29) and $p_{ab} \geq 1 \wedge \frac{c_0(\mu_a+\nu_b)r\log n}{n} \geq 1 \wedge \frac{c_0(\mu_a+\nu_b)^2 r^2 \log n}{2n^2}$, we have

$$\sum_{i=a,j=b} \leq \frac{1-p_{ab}}{p_{ab}w_{ab}^2} Z_{ab}^2 \left(\frac{\mu_a r}{n} + \frac{\nu_b r}{n}\right)^2 \leq \frac{2\left\|Z\right\|_{\mu(\infty)}^2}{c_0 \log n}.$$

By (29) and $p_{aj}w_{ab}^2 \geq w_{ab}^2 \wedge \left(c_0 w_{aj}^2 \frac{\nu_b r}{n} \log n\right)$, we have

$$\sum_{i=a,j\neq b} \leq \sum_{,j\neq b} \frac{1-p_{aj}}{p_{aj}w_{ab}^2} Z_{aj}^2 \left|e_b^\top V V^\top e_j\right| \leq \frac{\left\|Z\right\|_{\mu(\infty)}^2}{c_0 \log n} \cdot \frac{n}{\nu_b r} \sum_{j\neq b} \left|e_b^\top V V^\top e_j\right|,$$

which implies $\sum_{i=a,j\neq b} \leq \left\|Z\right\|_{\mu(\infty)}^2 / (c_0 \log n)$. Similarly we can bound $\sum_{i\neq a,j=b}$ by the same quantity. Finally, by (29) and $p_{ij} \geq 1 \wedge \left(c_0 \frac{\mu_i r}{n} \frac{\nu_j r}{n} \log n\right)$, we have

$$\begin{aligned}\sum_{i\neq a,j\neq b} &\leq \frac{1}{w_{ab}^2} \sum_{i\neq a,j\neq b} \frac{(1-p_{ij})Z_{ij}^2}{p_{ij}} \cdot \left|e_i^\top U U^\top e_a\right| \left|e_b^\top V V^\top e_j\right| \\ &\leq \frac{\left\|Z\right\|_{\mu(\infty)}^2}{c_0 \log n} \cdot \frac{1}{w_{ab}^2} \sum_{i\neq a} \left|e_i^\top U U^\top e_a\right| \sum_{j\neq b} \left|e_b^\top V V^\top e_j\right|,\end{aligned}$$

which implies $\sum_{i\neq a,j\neq b} \leq \left\|Z\right\|_{\mu(\infty)}^2 / (c_0 \log n)$. Combining pieces, we obtain

$$\left|\mathbb{E}\left[\sum_{ij} s_{ij}^2\right]\right| \leq 5\left\|Z\right\|_{\mu(\infty)}^2 / (c_0 \log n).$$

Applying the Bernstein inequality (Theorem 16), we conclude that

$$\left|\left[\left(P_T R_\Omega P_T - P_T\right) Z\right]_{ab} \sqrt{\frac{n}{\mu_a r}} \sqrt{\frac{n}{\nu_b r}}\right| = \left|\sum_{i,j} s_{ij}\right| \leq \frac{1}{2}\left\|Z\right\|_{\mu(\infty)}$$

w.h.p. for c_0 sufficiently large. The desired result follows from a union bound over all (a, b).

Appendix B. Proof of Corollary 4

Recall the setting: for each row of M, we pick it with some probability p and observe all its elements. We need a simple lemma. Let $J \subseteq [n]$ be the (random) set of the indices of the row picked, and $P_J(Z)$ be the matrix that is obtained from Z by zeroing out the rows outside J. Recall that $U\Sigma V^\top$ is the SVD of M.

Lemma 14 *If $\mu_i(M) := \frac{n}{r}\left\|U^\top e_i\right\|^2 \leq \mu_0$ for all $i \in [n]$, and $p \geq 10\frac{\mu_0 r}{n}\log\frac{2r}{\delta}$, then with probability at least $1-\delta$,*

$$\left\|U^\top P_J(U) - I_{r\times r}\right\| \leq \frac{1}{2},$$

where $I_{r\times r}$ is the identity matrix in $\mathbb{R}^{r\times r}$.

Proof Define the random variable $\eta_j := \mathbb{I}(i \in J)$ for $i = 1, 2, \ldots, n$, where $\mathbb{I}(\cdot)$ is the indicator function. Note that

$$U^\top P_J(U) - I_{r\times r} = U^\top P_J(U) - U^\top U = \sum_{i=1}^n S_{(i)} := \sum_{i=1}^n \left(\frac{1}{p}\eta_i - 1\right) U^\top e_i e_i^\top U.$$

The matrices $\{S_{(i)}\}$ are mutually independent and satisfy $\mathbb{E}\left[S_{(i)}\right] = 0$, $\left\|S_{(i)}\right\| \le \frac{1}{p}\left\|U^\top e_i\right\|_2^2 \le \frac{\mu_0 r}{pn}$, and

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{i=1}^n S_{(i)}S_{(i)}^\top\right]\right\| = \left\|\mathbb{E}\left[\sum_{i=1}^n S_{(i)}^\top S_{(i)}\right]\right\| &= \frac{1-p}{p}\left\|\sum_{i=1}^n U^\top e_i e_i^\top U U^\top e_i e_i^\top U\right\| \\
&= \frac{1-p}{p}\left\|U^\top\left(\sum_{i=1}^n e_i e_i^\top \left\|U^\top e_i\right\|_2^2\right) U\right\| \\
&\le \frac{1}{p}\left\|\sum_{i=1}^n e_i e_i^\top \left\|U^\top e_i\right\|_2^2\right\| \\
&= \frac{1}{p}\max_i \left\|U^\top e_i\right\|_2^2 \le \frac{\mu_0 r}{pn}.
\end{aligned}$$

Note that $S_{(i)}$ are $r \times r$ matrices. It follows from the matrix Bernstein (Theorem 16) that when $p \ge \frac{10\mu_0 r}{n}\log\frac{2r}{\delta}$, we have

$$\mathbb{P}\left\{\left\|U^\top P_J(U) - I_{r\times r}\right\| \ge \frac{1}{2}\right\} \le 2r\exp\left(\frac{-(1/2)^2/2}{\frac{\mu_0 r}{6pn} + \frac{\mu_0 r}{pn}}\right) \le \delta.$$

■

Note that $\left\|U^\top P_J(U) - I_{r\times r}\right\| \le \frac{1}{2}$ implies that $U^\top P_J(U)$ is invertible, which further implies $P_J(U) \in \mathbb{R}^{n\times r}$ has rank-r. The rows picked are $P_J(M) = P_J(U)\Sigma V^\top$, which thus have full rank-r and their row space must be the same as the row space of M. Therefore, the leverage scores $\{\tilde{\nu}_j\}$ of these rows are the same as the row leverage scores $\{\nu_j(M)\}$ of M. Also note that we must have $\mu_0 \ge 1$. Setting δ and sampling Ω as described in the corollary and applying Theorem 2, we are guaranteed to recover M exactly with probability at least $1 - 9n^{-10}$. The total number of elements we have observed is

$$pn + \sum_{i,j} p_{ij} = 10\mu_0 r\log\left(\frac{2r}{4n^{-10}}\right) + c_0(\mu_0 rn + rn)\log^2 n \le c_1\mu_0 rn\log^2 n$$

for some sufficiently large universal constant c_1, and by Hoeffding's inequality, the actual number of observations is at most two times the expectation with probability at least $1 - n^{-10}$ provided c_0 is sufficiently large. The corollary follows from the union bound.

Appendix C. Proof of Theorem 6

We prove the theorem assuming $\sum_{k=1}^r \frac{1}{a_k} = \sum_{k=1}^r \frac{1}{b_k} = r$; extension to the general setting in the theorem statement will only affect the pre-constant in (4) by a factor of at most 2.

For each $k \in [r]$, let $s_k := \frac{2n}{a_k r}$, $t_k := \frac{2n}{b_k r}$. We assume the s_k's and t_k's are all integers. Under the assumption on a_k and b_k, we have $1 \le s_k, t_k \le n$ and $\sum_{k=1}^r s_k = \sum_{k=1}^r t_k = n$. Define the sets $I_k := \left\{\sum_{l=1}^{r-1} s_l + i : i \in [s_k]\right\}$ and $J_k := \left\{\sum_{l=1}^{r-1} t_l + j : j \in [t_k]\right\}$; note that $\bigcup_{k=1}^r I_k = \bigcup_{k=1}^r J_k = [n]$. The vectors $\vec{\mu}$ and $\vec{\nu}$ are given by

$$\mu_i = a_k, \quad \forall k \in [r], i \in I_k,$$
$$\nu_j = b_k, \quad \forall k \in [r], j \in J_k.$$

It is clear that $\vec{\mu}$ and $\vec{\nu}$ satisfy the property 1 in the statement of the theorem.

Let the matrix $M^{(0)}$ be given by $M^{(0)} = AB^\top$, where $A, B \in \mathbb{R}^{n \times r}$ are specified below.

- For each $k \in [r]$, we set
$$A_{ik} = \sqrt{\frac{1}{s_k}}$$
for all $i \in I_k$. All other elements of A are set to zero. Therefore, the k-th column of A has s_k non-zero elements equal to $\sqrt{\frac{1}{s_k}}$, and the columns of A have disjoint supports.
- Similarly, for each $k \in [r]$, we set
$$B_{jk} = \sqrt{\frac{1}{t_k}}$$
for all $j \in J_k$. All other elements of B are set to zero.

Observe that A is an orthonormal matrix, so

$$\mu_i\left(M^{(0)}\right) = \frac{n}{r}\left\|A_{i\cdot}\right\|_2^2 = \frac{n}{r} \cdot \frac{1}{s_k} = \frac{a_k}{2} = \frac{\mu_i}{2} \le \mu_i, \forall k \in [r], i \in I_k, .$$

A similar argument shows that $\nu_j\left(M^{(0)}\right) \le \nu_j, \forall j \in [n]$. Hence $M^{(0)} \in \mathcal{M}_r\left(\vec{\mu}, \vec{\nu}\right)$. We note that $M^{(0)}$ is a block diagonal matrix with r blocks where the k-th block has size $s_k \times t_k$, and $\left\|M^{(0)}\right\|_F = \sqrt{r}$.

Consider the i_0 and j_0 in the statement of the theorem. There must exit some $k_1, k_2 \in [r]$ such that $i_0 \in I_{k_1}$ and $j_0 \in J_{k_2}$. Assume w.l.o.g. that $s_{k_1} \ge t_{k_2}$. then

$$p_{i_0 j_0} \le \frac{\mu_{i_0} + \nu_{j_0}}{4n} \cdot r \log\left(\frac{1}{\eta}\right) = \frac{a_{k_1} + b_{k_2}}{4n} \cdot r \log\left(\frac{1}{\eta}\right) = \frac{\log\left(1/\eta\right)}{4s_{k_1}} + \frac{\log\left(1/\eta\right)}{4t_{k_2}} \le \frac{\log\left(1/\eta\right)}{2t_{k_2}},$$

where $\eta = \frac{\mu_{i_0} r}{2n} = \frac{1}{s_{k_1}}$ in part 2 of the theorem and $\eta = \frac{2}{n}$ in part 3. Because $\{p_{ij}\}$ is location-invariant w.r.t. $M^{(0)}$, we have

$$p_{ij} = p_{i_0 j_0} \le \frac{\log\left(1/\eta\right)}{2t_{k_2}}, \quad \forall i \in I_{k_1}, j \in J_{k_2}.$$

Let $W_i := |(\{i\} \times J_{k_2}) \cap \Omega|$ be the number of observed elements on $\{i\} \times J_{k_2}$. Note that for each $i \in I_{k_1}$, we have

$$\mathbb{P}\left[W_i = 0\right] = \prod_{j \in J_{k_2}} (1 - p_{ij}) \ge \left(1 - \frac{\log(1/\eta)}{2t_{k_2}}\right)^{t_{k_2}} \ge \exp\left(\log \eta\right) = \eta,$$

where we use $1-x \geq e^{-2x}, \forall 0 \leq x \leq \frac{1}{2}$ in the second inequality. Therefore, there must exist $i^* \in I_{k_1}$ for which there is no observed element in $\{i^*\} \times J_{k_2}$ with probability

$$\begin{aligned}\mathbb{P}\left[W_{i^*} = 0, \exists i^* \in I_{k_1}\right] &= 1 - \mathbb{P}\left[W_i \geq 1, \forall i \in I_{k_1}\right] \\ &\geq 1 - (1-\eta)^{s_{k_1}} \geq 1 - e^{-\eta s_{k_1}} \geq \frac{1}{2}\eta s_{k_1} \geq \begin{cases} \frac{1}{2}, & \eta = \frac{\mu_{i_0} r}{4n} \\ \frac{1}{n}, & \eta = \frac{n}{2}. \end{cases}\end{aligned}$$

These are the probabilities that appear in part 2 and part 3 of the theorem statement, respectively.

Now choose a number $\bar{s} \geq s_{k_1}$. Let $M^{(1)} = \bar{A}B^\top$, where B is the same as before and $\bar{A}$ is given by

$$\bar{A}_{ik} = \begin{cases} \sqrt{\frac{1}{\bar{s}}}, & i = i^*, k = k_2 \\ A_{ik}, & \text{otherwise.} \end{cases}$$

By varying $\bar{s}$ we can construct infinitely many such $M^{(1)}$. Clearly $M^{(1)}$ is rank-r. Observe that $M^{(1)}$ differs from $M^{(0)}$ only in the elements with indices in $\{i^*\} \times J_{k_2}$, which are not observed, so

$$M^{(0)}_{ij} = M^{(1)}_{ij}, \quad \forall (i,j) \in \Omega.$$

Also observe that any $\{p_{ij}\}$ that is location-invariant w.r.t. $M^{(0)}$ is also location-invariant w.r.t. $M^{(1)}$. The following lemma guarantees that $M^{(1)} \in \mathcal{M}_r\left(\vec{\mu}, \vec{\nu}\right)$, which completes the proof of the theorem.

Lemma 15 *The matrix $M^{(1)}$ constructed above satisfies*

$$\begin{aligned}\mu_i\left(M^{(1)}\right) &\leq 2\mu_i\left(M^{(0)}\right), \quad \forall i \in [n], \\ \nu_j\left(M^{(1)}\right) &= \nu_j\left(M^{(0)}\right), \quad \forall j \in [n].\end{aligned}$$

Proof Note that by the definition, the leverage scores of a rank-r matrix M with SVD $M = U\Sigma V^\top$ can be expressed as

$$\mu_i\left(M\right) = \frac{n}{r}\left\|U^\top e_i\right\|_2^2 = \frac{n}{r}\left\|UU^\top e_i\right\|_2^2 = \frac{n}{r}\left\|\mathcal{P}_{\mathrm{col}(M)}(e_i)\right\|_2^2,$$

where $\mathrm{col}(M)$ denotes the column space of M and $\mathcal{P}_{\mathrm{col}(M)}(\cdot)$ is the Euclidean projection onto the column space of M. A similar relation holds for the row leverage scores and the row space of M. In other words, the column/row leverage scores of a matrix are determined by its column/row space. Because $M^{(0)}$ and $M^{(1)}$ have the same row space (which is the span of the columns of B), the second set of equalities in the lemma hold.

It remains to prove the first set of inequalities for the column leverage scores. If $k_1 = k_2$, then the columns of $\bar{A}$ have unit norms and are orthogonal to each other. Using the above expression for the leverage scores, we have

$$\mu_i\left(M^{(1)}\right) = \frac{n}{r}\left\|\bar{A}\bar{A}^\top e_i\right\|_2^2 = \frac{n}{r}\left\|\bar{A}^\top e_i\right\|_2^2 = \frac{n}{r}\left\|A^\top e_i\right\|_2^2 = \mu_i\left(M^{(0)}\right).$$

If $k_1 \neq k_2$, we may assume without loss of generality that $k_1 = 1$, $k_2 = 2$ and $i^* = 1$. In the sequel we use $\bar{A}_i$ to denote the i-th columns of $\bar{A}$. We now construct two vectors $\tilde{\alpha}$ and $\tilde{\beta}$

which have the same span with $\bar{A}_1$ and $\bar{A}_2$. Define two vectors $\alpha, \beta \in \mathbb{R}^n$, such that the first s_1 elements of α and the $\{s_1+1, \ldots, s_1+s_2\}$-th elements of β are one, the first element of β is $\sqrt{\frac{s_2}{\bar{s}}}$, and all other elements of α and β are zero. Clearly $\alpha = \sqrt{s_1}\bar{A}_1$ and $\beta = \sqrt{s_2}\bar{A}_2$, so $\text{span}(\alpha, \beta) = \text{span}(\bar{A}_1, \bar{A}_2)$. We next orthogonalize α and β by letting $\bar{\alpha} = \alpha$ and

$$\bar{\beta} = \beta - \frac{\langle \alpha, \beta \rangle}{\|\alpha\|^2}\alpha = \beta - \frac{\sqrt{s_2}}{s_1\sqrt{\bar{s}}}\alpha = \begin{cases} \frac{(s_1-1)\sqrt{s_2}}{s_1\sqrt{\bar{s}}}, & i = 1 \\ -\frac{\sqrt{s_2}}{s_1\sqrt{\bar{s}}}, & i = 2, \ldots, s_1 \\ 1, & i = s_1+1, \ldots, s_1+s_2 \\ 0, & i = s_1+s_2+1, \ldots, n. \end{cases}$$

Note that $\text{span}(\bar{\alpha}, \bar{\beta}) = \text{span}(\alpha, \beta)$ and $\langle \bar{\alpha}, \bar{\beta} \rangle = 0$. Simple calculation shows that $\|\bar{\alpha}\|_2^2 = \|\alpha\|_2^2 = s_1$ and $\|\bar{\beta}\|_2^2 = \left(\frac{s_1-1}{s_1\bar{s}} + 1\right) s_2$. Finally, we normalize $\bar{\alpha}$ and $\bar{\beta}$ by letting $\tilde{\alpha} = \bar{\alpha}/\|\bar{\alpha}\|$ and $\tilde{\beta} = \bar{\beta}/\|\bar{\beta}\|$. It is clear that $\text{span}(\tilde{\alpha}, \tilde{\beta}) = \text{span}(\bar{A}_1, \bar{A}_2)$, and $\langle \tilde{\alpha}, \bar{A}_k \rangle = \left\langle \tilde{\beta}, \bar{A}_k \right\rangle = 0, \forall k = 3, \ldots, r$.

Now consider the matrix $\tilde{A} \in \mathbb{R}^{n \times r}$ obtained from $\bar{A}$ by replacing the first two columns of $\bar{A}$ with $\tilde{\alpha}$ and $\tilde{\beta}$, respectively. Because $\text{col}(\tilde{A}) = \text{col}(\bar{A}) = \text{col}(M^{(1)})$, we have

$$\mu_i\left(M^{(1)}\right) = \frac{n}{r}\left\|\mathcal{P}_{\text{col}(\tilde{A})}\left(e_i\right)\right\|^2.$$

But the columns of $\tilde{A}$ have unit norms and are orthogonal to each other. It follows that

$$\mu_i\left(M^{(1)}\right) = \frac{n}{r}\left\|\tilde{A}\tilde{A}^\top e_i\right\|^2 = \frac{n}{r}\left\|\tilde{A}^\top e_i\right\|^2.$$

For $s_1+s_2 < i \le n$, since $\bar{s} \ge s_1$ we have $\left\|\tilde{A}^\top e_i\right\|^2 = \left\|\bar{A}^\top e_i\right\|^2 = \left\|A^\top e_i\right\|^2$ so $\mu_i\left(M^{(1)}\right) = \mu_i\left(M^{(0)}\right)$. For $i \in [s_1+s_2]$, we have

$$\left\|\tilde{A}^\top e_i\right\|^2 = \tilde{\alpha}_i^2 + \tilde{\beta}_i^2 = \begin{cases} \frac{1}{s_1} + \frac{(s_1-1)^2}{s_1(s_1-1)+s_1^2\bar{s}} \le \frac{2}{s_1} = 2\left\|A^\top e_i\right\|^2, & i = 1 \\ \frac{1}{s_1} + \frac{1}{s_1(s_1-1)+s_1^2\bar{s}} \le \frac{2}{s_1} = 2\left\|A^\top e_i\right\|^2, & i = 2, \ldots, s_1 \\ \frac{s_1\bar{s}}{(s_1-1+s_1\bar{s})s_2} \le \frac{1}{s_2} = \left\|A^\top e_i\right\|^2, & i = s_1+1, \ldots, s_1+s_2. \end{cases}$$

This means

$$\mu_i\left(M^{(1)}\right) \le \frac{2n}{r}\left\|A^\top e_i\right\|^2 = 2\mu_i(M^{(0)}), \forall i \in [s_1+s_2],$$

which completes the proof of the lemma. ■

Appendix D. Proof of Theorem 7

Suppose the rank-r SVD of $\bar{M}$ is $\bar{U}\bar{\Sigma}\bar{V}^\top$; so $\bar{U}\bar{\Sigma}\bar{V}^\top = RMC = RU\Sigma V^\top C$. By definition, we have

$$\frac{\bar{\mu}_i r}{n} = \left\|P_{\tilde{U}}(e_i)\right\|_2^2,$$

where $P_{\tilde{U}}(\cdot)$ denotes the projection onto the column space of $\tilde{U}$, which is the same as the column space of RU. This projection has the explicit form

$$P_{\tilde{U}}(e_i) = RU\left(U^\top R^2 U\right)^{-1} U^\top R e_i.$$

It follows that

$$\begin{aligned}\frac{\bar{\mu}_i r}{n} &= \left\| RU\left(U^\top R^2 U\right)^{-1} U^\top R e_i \right\|_2^2 \\ &= R_i^2 e_i^\top U\left(U^\top R^2 U\right)^{-1} U^\top e_i \\ &\le R_i^2 \left[\sigma_r\left(RU\right)\right]^{-2} \left\| U^\top e_i \right\|_2^2 \\ &\le R_i^2 \frac{\mu_0 r}{n} \left[\sigma_r\left(RU\right)\right]^{-2}, \end{aligned} \tag{30}$$

where $\sigma_r(\cdot)$ denotes the r-th singular value and the last inequality follows from the standard incoherence assumption $\max_{i,j}\{\mu_i, \nu_j\} \le \mu_0$. We now bound $\sigma_r\left(RU\right)$. Since RU has rank r, we have

$$\sigma_r^2\left(RU\right) = \min_{\|x\|=1} \|RUx\|_2^2 = \min_{\|x\|=1} \sum_{i=1}^n R_i^2 \left| e_i^\top U x \right|^2. \tag{31}$$

If we let $z_i := \left|e_i^\top U x\right|^2$ for each $i \in [n]$, then z_i satisfies

$$\sum_{i=1}^n z_i = \|Ux\|_2^2 = \|x\|_2^2 = 1$$

and by the standard incoherence assumption,

$$z_i \le \left\| U^\top e_i \right\|_2^2 \|x\|_2^2 \le \frac{\mu_0 r}{n}.$$

Therefore, the value of the minimization (31) is lower-bounded by the optimal value of the following program

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \ & \sum_{i=1}^n R_i^2 z_i \\ \text{s.t.} \ & \sum_{i=1}^n z_i = 1, \quad 0 \le z_i \le \frac{\mu_0 r}{n},\ i = 1, \dots, n. \end{aligned} \tag{32}$$

From the theory of linear programming, we know the minimum is achieved at an extreme point z^* of the feasible set. Such an extreme point z^* satisfies $z_i^* \ge 0, \forall i$ and n linear equalities

$$\begin{aligned} \sum_{i=1}^n z_i^* &= 1, \\ z_i^* &= 0, \quad \text{for } i \in I_1, \\ z_i^* &= \frac{\mu_0 r}{n}, \text{ for } i \in I_2 \end{aligned}$$

for some index sets I_1 and I_2 such that $I_1 \cap I_2 = \phi$, $|I_1| + |I_2| = n - 1$. It is easy to see that we must have $|I_2| = \left\lfloor \frac{n}{\mu_0 r} \right\rfloor$. Since $R_1 \le R_2 \le \ldots \le R_n$, the minimizer z^* has the form

$$z_i^* = \begin{cases} \frac{\mu_0 r}{n}, & i = 1, \ldots, \left\lfloor \frac{n}{\mu_0 r} \right\rfloor, \\ 1 - \left\lfloor \frac{n}{\mu_0 r} \right\rfloor \cdot \frac{\mu_0 r}{n}, & i = \left\lfloor \frac{n}{\mu_0 r} \right\rfloor + 1, \\ 0, & i = \left\lfloor \frac{n}{\mu_0 r} \right\rfloor + 2, \ldots, n, \end{cases}$$

and the value of the minimization (32) is at least

$$\sum_{i=1}^{\lfloor n/(\mu_0 r) \rfloor} R_i^2 \frac{\mu_0 r}{n}.$$

This proves that $\sigma_r^2(RU) \ge \frac{\mu_0 r}{n} \sum_{i=1}^{\lfloor n/(\mu_0 r) \rfloor} R_i^2$. Combining with (30), we obtain that

$$\frac{\bar{\mu}_i r}{n} \le \frac{R_i^2}{\sum_{i'=1}^{\lfloor n/(\mu_0 r) \rfloor} R_i^2}, \quad \frac{\bar{\nu}_j r}{n} \le \frac{C_j^2}{\sum_{j'=1}^{\lfloor n/(\mu_0 r) \rfloor} C_{j'}^2};$$

the proof for $\bar{\nu}_j$ is similar. Applying Theorem 2 to the equivalent problem (9) with the above bounds on $\bar{\mu}_i$ and $\bar{\nu}_j$ proves the theorem.

Appendix E. Weighted vs Unweighted Nuclear Norm Minimization for Non-uniform Sampling

In this section we provide a concrete example of the gain of weighting under the setting of Section 5.1, where the observed entries are given and distributed non-uniformly. Suppose M is an n-by-n matrix with rank r, and its incoherence parameter satisfies $\mu_0 r = c$, where c is a numerical constant. We assume the sampling probabilities have the form $p_i^{\mathrm{r}} = p_i^{\mathrm{c}} = \min\{\gamma \frac{i^{0.15} \log n}{n^{0.65}}, 1\}$ for $i = 1, 2, \ldots, n$; here the minimization ensures $p_i^{\mathrm{r}} p_j^{\mathrm{c}}$ is a probability. Note that the parameter γ determines the expected number of samples $\sum_{i,j} p_i^{\mathrm{r}} p_j^{\mathrm{c}}$. For the condition (11) for the unweighted approach to hold, we need $\gamma^2 \gtrsim n^{0.3}$, and thus the the expected number of samples is at least

$$\sum_{i,j} p_i^{\mathrm{r}} p_j^{\mathrm{c}} \ge \sum_{i,j} \gamma \frac{i^{0.15}}{n^{0.65}} \cdot \gamma \frac{j^{0.15}}{n^{0.65}} = \Omega(n^{1.3}),$$

where we use the estimate $\sum_{i=1}^n i^{0.15} = \Theta(n^{1.15})$. On the other hand, the condition (10) for the weighted approach is satisfied as long as $\gamma^2 \gtrsim n^{0.15}$, so the the expected number of samples satisfies

$$\sum_{i,j} p_i^{\mathrm{r}} p_j^{\mathrm{c}} \le \sum_{i,j} \gamma \frac{i^{0.15}}{n^{0.65}} \cdot \gamma \frac{j^{0.15}}{n^{0.65}} \cdot \log^2 n = O(n^{1.15} \log^2 n)$$

when $\gamma^2 = \Theta(n^{0.15})$. Therefore, the number of samples required by the condition (10) for the weighted approach is *order-wise* smaller than the unweighted counterpart (11). Note that the conditions (10) and (11) are the *best known sufficient* conditions for exact matrix completion using the weighted and unweighted approaches, respectively, so the comparison above suggests a significant gain in sample complexity using the weighted approach.

Appendix F. Matrix Bernstein Inequality

Theorem 16 (Tropp 2012) *Let $X_1, \ldots, X_N \in \mathbb{R}^{n_1 \times n_2}$ be independent zero mean random matrices. Suppose*

$$\max\left\{\left\|\mathbb{E}\sum_{k=1}^{N} X_k X_k^\top\right\|, \left\|\mathbb{E}\sum_{k=1}^{N} X_k^\top X_k\right\|\right\} \leq \sigma^2$$

and $\|X_k\| \leq B$ almost surely for all k. Then we have

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{N} X_k\right\| \geq t\right\} \leq (n_1+n_2)\exp\left(\frac{-t^2/2}{Bt/3+\sigma^2}\right)$$

As a consequence, for any $c > 0$, we have

$$\left\|\sum_{k=1}^{N} X_k\right\| \leq 2\sqrt{c\sigma^2\log(n_1+n_2)} + cB\log(n_1+n_2). \tag{33}$$

with probability at least $1-(n_1+n_2)^{-(c-1)}$.

References

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

D. Achlioptas, Z. Karnin, and E. Liberty. Matrix entry-wise sampling: simple is best. *http://cs-www.cs.yale.edu/homes/el327/papers/matrixSampling.pdf*, 2013.

S. Arora, E. Hazan, and S. Kale. A fast random sampling algorithm for sparsifying matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 272–279. Springer, 2006.

C. Boutsidis, M. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Symposium on Discrete Algorithms*, pages 968–977, 2009.

N. Burq, S. Dyatlov, R. Ward, and M. Zworski. Weighted eigenfunction estimates with applications to compressed sensing. *SIAM Journal on Mathematical Analysis*, 44(5): 3481–3501, 2012.

J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

S. Chatterjee and A. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Information Processing Letters*, 111(8):385–389, 2011.

P. Drineas, M. Magdon-Ismail, M. Mahoney, and D. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13: 3475–3506, 2012.

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

R. Foygel, O. Shamir, N. Srebro, and R. Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems 24*, pages 2133–2141. 2011.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pages 665–674. ACM, 2013.

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

F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *IEEE Transactions on Image Processing*, 23(2):612–622, 2014.

A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems 26*, pages 836–844, 2013.

A. Krishnamurthy and A. Singh. On the power of adaptivity in matrix completion and approximation. *arXiv preprint arXiv:1407.3619*, 2014.

Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215*, 2009.

M. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

H. Rauhut and R. Ward. Sparse Legendre expansions via ℓ_1-minimization. *Journal of Approximation Theory*, 164(5):517–533, 2012.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

D. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

N. Srebro and R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064, 2010.

J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

X. Yang and G. Karniadakis. Reweighted ℓ_1 minimization method for stochastic elliptic differential equations. *Journal of Computational Physics*, 248:87–108, 2013.