Weighted Matrix Completion From Non-Random, Non-Uniform Sampling Patterns

Simon Foucart, Deanna Needell, *Member, IEEE*, Reese Pathak,
Yaniv Plan, and Mary Wootters, *Member, IEEE*

***Abstract*— We study the matrix completion problem when the observation pattern is deterministic and possibly non-uniform. We propose a simple and efficient debiased projection scheme for recovery from noisy observations and analyze the error under a suitable weighted metric. We introduce a simple function of the weight matrix and the sampling pattern that governs the accuracy of the recovered matrix. We derive theoretical guarantees that upper bound the recovery error and nearly matching lower bounds that showcase optimality in several regimes. Our numerical experiments demonstrate the computational efficiency and accuracy of our approach, and show that debiasing is essential when using non-uniform sampling patterns.**

***Index Terms*— Matrix completion, nonuniform sampling.**

I. Introduction

The *matrix completion problem* is to determine a complete $d_1 \times d_2$ matrix $\mathbf{M}$ from a subset $\Omega \subset [d_1] \times [d_2]$ of its entries. A typical assumption that makes such a problem well-posed is that the underlying matrix from which the entries are observed is low-rank (or approximately low-rank). Matrix completion has many applications, including collaborative filtering [27], system identification [46], sensor localization [9], [57], [58], rank aggregation [26], scene recovery in imaging [17], [63], multi-class learning [1]–[3], and more. This is now a well-studied problem, and there are several main approaches to its solution, such as low-rank projection [33], [34] and convex optimization [13], [60], which have rigorous provable recovery guarantees (see e.g. [13]–[15], [25], [29], [33], [34], [36], [37], [40], [41], [48], [53], [54]).

Besides a low-rank assumption on the underlying matrix, one also clearly needs to assume something on the sampling pattern Ω. Theoretical guarantees for matrix completion typically enforce that the sampling pattern is obtained from (most often uniform) random sampling [8], [21], [22], [39], [47], [61]. However, for many applications, the sampling pattern may not be uniformly random, and indeed may not be reasonably modeled as random at all.

In this paper, we study the problem of matrix completion with *deterministic* sampling, that is when Ω is fixed arbitrarily. This version of the problem has been studied before [7], [23], [31], [43], [44], although much less extensively than the case with random sampling.

For some sampling patterns Ω, recovering the entire matrix $\mathbf{M}$ accurately from observations indexed by Ω may not be possible: for example, consider the case where Ω only has support on the left half of the matrix. Thus, we use an appropriate *weighted* error metric of the form $\|\mathbf{H} \circ (\hat{\mathbf{M}} - \mathbf{M})\|_F$, where $\hat{\mathbf{M}}$ is the recovered matrix, and $\mathbf{H}$ is an appropriate weight matrix. Informally, the weight matrix $\mathbf{H}$ allows us to quantify which entries can be recovered accurately and which cannot.

Our work extends a great deal of prior work which assumes uniform sampling of matrix entries and considers an unweighted error metric. This corresponds to taking $\mathbf{H}$ to be the all-ones matrix; note that for uniformly sampled entries, this choice of $\mathbf{H}$ is a rank-1 matrix which approximates the matrix $\mathbf{1}_\Omega \in \{0,1\}^{d_1 \times d_2}$. We generalize these results to (almost) any rank-1 matrix $\mathbf{H}$ which approximates $\mathbf{1}_\Omega$.

More precisely, we show that when $\mathbf{H}$ satisfies certain conditions with respect to the sampling pattern Ω, which can be easily verified, a simple "debiased projection" algorithm performs well. Moreover, this algorithm is extremely efficient. We also establish lower bounds that show that our debiased projection algorithm is nearly optimal in several situations. Finally, we include numerical results that demonstrate the efficacy and advantages of our approach.

A. Background and Motivation

Given a sampling pattern Ω, we write $\mathbf{1}_\Omega$ to denote the matrix whose entries are 1 on Ω and zero elsewhere, so that

Manuscript received October 31, 2019; revised July 27, 2020; accepted November 4, 2020. Date of publication November 19, 2020; date of current version January 21, 2021. The work of Simon Foucart was supported in part by the NSF under Grant DMS-1622134, Grant DMS-1664803, and Grant CCF-1934904. The work of Deanna Needell was supported in part by the NSF CAREER under Grant DMS-1348721, in part by the NSF BIGDATA under Grant 1740325, and in part by the NSF DMS under Grant 2011140. The work of Reese Pathak was supported in part by the Stanford UAR Major Grant. The work of Yaniv Plan was supported in part by the NSERC Discovery under Grant 22R23068, in part by the PIMS CRG 33: High-Dimensional Data Analysis, and in part by the Tier II CRC in Data Science. The work of Mary Wootters was supported in part by the NSF CAREER Award under Grant CCF-1844628. This article was presented at the *Wavelets and Sparsity XVII*, the 2017 International Society for Optics and Photonics. *(Corresponding author: Yaniv Plan.)*

Simon Foucart is with the Department of Mathematics, Texas A&M University, College Station, TX 77843 USA (e-mail: foucart@tamu.edu).

Deanna Needell is with the Department of Mathematics, UCLA, Los Angeles, CA 90095 USA (e-mail: deanna@math.ucla.edu).

Reese Pathak is with the Department of EECS, UC Berkeley, Berkeley, CA 94709 USA (e-mail: pathakr@berkeley.edu).

Yaniv Plan is with the Department of Mathematics, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: yaniv@math.ubc.ca).

Mary Wootters is with the Department of Computer Science, Stanford University, Stanford, CA 94305 USA, and also with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: marykw@stanford.edu).

Communicated by R. Talmon, Associate Editor for Signal Processing.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2020.3039308.

Digital Object Identifier 10.1109/TIT.2020.3039308

the entries of $\mathbf{M}_\Omega = \mathbf{1}_\Omega \circ \mathbf{M}$ are equal to those of $\mathbf{M}$ on Ω, and are equal to 0 on Ω^c. Above, $\circ$ denotes the (entrywise) Hadamard product.

While most work on matrix completion has been in the case where Ω is random (and usually uniform), there has been some work on the deterministic case, summarized in Section IV. Our starting point is the work [7], [31], [44], which shows that when $\mathbf{1}_\Omega \in \{0,1\}^{d_1 \times d_2}$ is "close" to an appropriately scaled version of the all-ones matrix (more precisely, when $\mathbf{1}_\Omega$ is the adjacency matrix of an expander graph) and the matrix $\mathbf{M}$ is sufficiently incoherent, it is possible to efficiently recover an estimate $\hat{\mathbf{M}}$ so that $\|\hat{\mathbf{M}} - \mathbf{M}\|_F$ is small.

Of course, as noted above, there are some sampling patterns so that we cannot hope to recover $\hat{\mathbf{M}}$ with small error $\|\hat{\mathbf{M}} - \mathbf{M}\|_F$. As a simple (and extreme) example consider sampling the entire *left* half of a matrix. We can exactly recover the left half of $\mathbf{M}$, but we learn nothing about the right half of $\mathbf{M}$.

Thus, our goal will be to find a weight matrix $\mathbf{H}$ so that we can recover $\hat{\mathbf{M}}$ so that $\left\|\mathbf{H} \circ (\hat{\mathbf{M}} - \mathbf{M})\right\|_F$ is small, compared to $\|\mathbf{H}\|_F$. In the first example above where $\mathbf{1}_\Omega$ is close to the all-ones matrix, we would choose $\mathbf{H}$ to be the all-ones matrix; in the second example where Ω samples only the left half of the matrix, we would choose $\mathbf{H}$ to be the matrix with ones on the left half and zeros on the right half.

More precisely, we are motivated by the following question.

Question 1: Given a sampling pattern Ω, and noisy observations $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$, for what *rank-one* weight matrices $\mathbf{H}$ can we efficiently find a matrix $\hat{\mathbf{M}}$ so that $\left\|\mathbf{H} \circ (\hat{\mathbf{M}} - \mathbf{M})\right\|_F$ is small compared to $\|\mathbf{H}\|_F$? And how can we efficiently find such weight matrices $\mathbf{H}$, or certify that a fixed $\mathbf{H}$ has this property?

The notion of using a weighted error metric in this context is not new [31], [43]. However, the difference between Question 1 and previous work is that we consider only rank-1 weight matrices. While this sacrifices some generality, as discussed in more detail in Section IV, this allows us to achieve faster algorithms that also tolerate noise, and additionally allows us to prove lower bounds.

B. Our Results

The goal of this paper is to answer Question 1. Our method is a simple weighted (which we refer to as "debiased") projection algorithm that performs well precisely when the quantity $\lambda = \left\|\mathbf{H} - \mathbf{H}^{(-1)} \circ \mathbf{1}_\Omega\right\|$ is small; here and throughout, $\|\cdot\|$ denotes the usual spectral norm and $\mathbf{H}^{(-1)}$ denotes the Hadamard (entry-wise) inverse. The parameter λ is efficient to compute, and moreover our algorithms are efficient. More precisely, we give two algorithms, one for exactly low-rank matrices and one for "approximately" low-rank matrices. Our algorithm for approximately low-rank matrices runs in essentially the time it takes to compute an SVD. Our algorithm for approximately low-rank matrices can be implemented by solving a semidefinite program. In addition, we derive lower bounds that show that our approach is nearly optimal in several situations.

To illustrate our upper and lower bounds, we consider two extreme cases, depending on the magnitude of λ.

Case 1: when λ is small. First, we illustrate our upper and lower bounds in a case where λ is small. More precisely, we consider the following case study: choose a rank-1 matrix $\mathbf{H}$, and let $(i,j) \in \Omega$ with probability H_{ij}^2. In this case, when the error matrix $\mathbf{Z}$ has standard deviation σ of roughly the same order of magnitude as the entries of $\mathbf{M}$, we prove matching upper and lower bounds that show that the error guarantee for our algorithm is nearly optimal, up to logarithmic factors in d and polynomial factors in r.

We stress that in this example above, even though we imagine drawing Ω at random, our results are still uniform: that is, we draw Ω once and fix it, and then prove that our algorithm works deterministically for all $\mathbf{M}$.

In addition to being a good way to showcase our results when λ is small, we believe that this particular case study is interesting for two reasons.

- First, this case study models a natural situation where the sampling distribution is not very uniform. For example, suppose that the matrix $\mathbf{M}$ represents preferences of users for items. There are some prolific users that rate many items, and there are some popular items that are rated by many users. In this case, it is reasonable to expect that the sampling pattern arises as above from the rank-1 matrix $\mathbf{H}$ which is the outer product of the vectors indicating how prolific each user is and how popular each item is. Another similar example is in survey design, where an important batch of questions may be asked more frequently than other questions.
- Second, this setting has applications to proportional sampling. More precisely, we show in Section VII-D how our results for this case study can be used to do matrix completion for incoherent matrices by sampling proportional to leverage scores. We recover results that are qualitatively similar to the results of [18]. Unlike that work, our results hold more generally for approximately-low-rank matrices, although as discussed more below in Section IV, the two works are incomparable.

Case 2: when λ is large. Second, we illustrate our bounds in settings where λ is large; we focus on symmetric sampling patterns Ω for which $\mathbf{1}_\Omega$ has a large gap between the largest and second-largest eigenvalues. In this setting our upper bounds are not very strong, since they depend on the parameter λ. However, we prove lower bounds which show that some dependence on λ is required. While we are not able to get a lower bound that matches our upper bound in the dependence on λ, we are able to show that the error must necessarily increase as λ increases, which suggest that our approach is qualitatively correct. To the best of our knowledge, this is the first lower bound showing that any dependence on a parameter like λ is necessary, despite the fact that previous works present upper bounds with some dependence on a parameter like this.

Finally, we present empirical results using both real and synthetic sampling patterns that show that debiasing is essential when the sampling pattern is non-uniform. We showcase reduction in recovery errors compared with standard (non debiased) approaches as we vary the sampling pattern constructed via regular graphs.

C. Organization

The rest of the paper is organized as follows. In Section II we instantiate notation, formalize the problem setup, and record several useful results that our theory will utilize. We establish our main results in Section III and relate our work to existing work in Section IV. Sections V and VI give generalized upper and lower bounds, respectively, which are specialized to important settings in Sections VII and VIII. We display numerical results in Section IX.

II. Set-up and Preliminaries

In this section we set notation and our formal problem statement.

A. Notation

We begin by setting notation. For an integer d, we use $[d]$ to mean the set $\{1, \ldots, d\}$. Throughout the paper, bold capital letters ($\mathbf{X}$) represent matrices, and bold lowercase letters ($\mathbf{x}$) represent vectors. Entries of a matrix $\mathbf{X}$ or a vector $\mathbf{x}$ are denoted by $X_{i,j}$ or x_i respectively. For a set $\Omega \subseteq [d_1] \times [d_2]$ and a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\mathbf{X}_\Omega$ to denote the matrix which is equal to $\mathbf{X}$ on entries in Ω and 0 otherwise.

We use $\|\mathbf{X}\|_\infty = \max_{i,j} |X_{i,j}|$ to denote the entry-wise ℓ_∞ norm for matrices, $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{i,j}^2}$ to denote the Frobenius norm, and $\|\mathbf{X}\|$ to denote the spectral norm of $\mathbf{X}$. We define the *max-norm* by

$$\|\mathbf{X}\|_{\max} := \min_{\mathbf{X}=\mathbf{U}\mathbf{V}^\top} \|\mathbf{U}\|_{2,\infty} \|\mathbf{V}\|_{2,\infty}.$$

Letting $B_{\max}$ be the max-norm unit ball, $B_{\max} := \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{X}\|_{\max} \le 1\}$, Grothendiek's inequality [32, Chapter 10] shows that $B_{\max}$ is close to a polytope with vertices rank-1 matrices with ± 1-valued entries. Concretely, letting $\mathcal{F}$ be the set of such matrices, $\mathcal{F} := \{uv^T : u \in \{+1, -1\}^{d_1}, v \in \{+1, -1\}^{d_2}\}$, Grothendieck's inequality states that $conv(\mathcal{F}) \subset B_{\max} \subset K_G \cdot conv(\mathcal{F})$, where $K_G \le 1.783$ is Grothendiek's constant. Given a rank-r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \le \gamma$, we have that $\|\mathbf{M}\|_{\max} \le \sqrt{r}\gamma$ (see [52, Corollary 2.2]). In this sense, the max norm serves as a proxy for the rank of a flat matrix that is robust to small perturbations.

We use K_r to denote the cone of rank-r matrices. We use B_∞, $B_{\max}$, B_F respectively to denote the unit balls for the corresponding norm.

B. Formal Set-Up

We now describe our formal set-up. Suppose that $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is a unknown matrix. In this paper, we will assume either that $\mathbf{M}$ is in $K_r \cap \beta B_\infty$—that is, $\mathbf{M}$ is "flat" and has low rank—or that $\mathbf{M}$ is in $\beta\sqrt{r}B_{\max}$—that is, $\mathbf{M}$ is "approximately" low-rank.

Remark 2 (Assumptions of "Flatness"): It is not hard to see that some assumption of "flatness" is required for matrix completion. Indeed, if the matrix $\mathbf{M}$ could be arbitrary, then there could be some arbitrarily large entry $M_{i,j}$ so that $(i,j) \notin \Omega$, and it would be impossible to obtain any nontrivial guarantee on the reconstruction error. There have been several notions of "flatness" introduced in the literature. One is *incoherence,* (as in, e.g., [14], [33], [53]), which says that if $\mathbf{M} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$ for $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times r}$, then the rows $\mathbf{u}$ of $\mathbf{U}$ and $\mathbf{v}$ of $\mathbf{V}$ have $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \le \mu_0\, r/d$ and $\langle \mathbf{u}, \mathbf{v} \rangle \le \mu_1 \sqrt{r}/d$. Another notion, which we use here and is also used in [20], is that $\|\mathbf{M}\|_\infty$ is bounded. Notice that the standard incoherence assumption implies that $\|\mathbf{M}\|_\infty \le \|\mathbf{M}\|(\mu_0\, r/d)^2$. We quantify "approximately low-rank" using the max-norm, which also has some notion of "flatness" built into it by definition. This assumption of flatness has also been used in the context of matrix completion, e.g., in [11].

Fix a sampling pattern $\Omega \subseteq [d_1] \times [d_2]$ and a rank-1 matrix $\mathbf{W}$.[1] Our goal will be to design an algorithm that gives provable guarantees for a worst-case $\mathbf{M}$, even if it is adapted to Ω. Our algorithm will observe $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$, where $Z_{i,j} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian random variables. From these observations, the goal is to learn something about $\mathbf{M}$. Notice that, depending on Ω, it might not be possible to estimate $\mathbf{M}$ well in a standard metric (like the Frobenius norm). Instead, in this paper, we are interested in learning $\mathbf{M}$ with small error in a *weighted* Frobenius norm; that is, we'd like to develop efficient algorithms to find a matrix $\hat{\mathbf{M}}$ so that

$$\left(\sum_{i,j} W_{ij}(M_{ij} - \hat{M}_{ij})^2\right)^{1/2} = \left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F$$

is small for some matrix $\mathbf{W}$ of interest. On the other hand, we will also prove lower bounds that demonstrate for which $(\Omega, \mathbf{W})$ combinations certain error bounds are not possible.

When measuring weighted error, it is important to normalize appropriately in order to understand what the bounds mean. In our setting, we will always report error normalized by $\left\|\mathbf{W}^{(1/2)}\right\|_F$: that is the goal is that

$$\frac{\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} = \left(\sum_{i,j} \frac{W_{ij}}{\sum_{i',j'} W_{i'j'}} (M_{ij} - \hat{M}_{ij})^2\right)^{1/2}$$

is small. Written out this way, it is clear that this gives a weighted average of the per-entry squared error. In light of the discussion above, we formally define our problem below. Ideally, the error function δ will tend to zero as $|\Omega|$ grows.

Remark 3 (Universality): We emphasize that the requirement in the problem above is a universal one. That is,

[1] In the introduction we discussed a rank-1 weight matrix $\mathbf{H}$ which plays the role of $\mathbf{W}^{(1/2)}$. Stating the results that way is easier to parse in an introduction, but it will be more convenient for the proofs to state the formal problem in terms of $\mathbf{W}$.

Problem: Weighted Universal Matrix Completion

Parameters:

- Dimensions d_1, d_2
- A sampling pattern $\Omega \subset [d_1] \times [d_2]$
- Parameters $\sigma, \beta, r > 0$
- A rank-1 weight matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ so that $W_{ij} > 0$ for all i, j.
- A set K (which for us will either be $K_r \cap \beta B_\infty$ or $\beta\sqrt{r}B_{\max}$)

Goal: Design an efficient algorithm $\mathcal{A}$ with the following guarantees:

- $\mathcal{A}$ takes as input entries $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$ so that $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d.
- $\mathcal{A}$ runs in polynomial time
- With high probability over the choice of $\mathbf{Z}$, $\mathcal{A}$ returns an estimate $\hat{\mathbf{M}}$ of $\mathbf{M}$ so that

$$\left(\sum_{i,j}\left(\frac{W_{ij}}{\sum_{i',j'} W_{i'j'}}\right)(M_{ij} - \hat{M}_{ij})^2\right)^{1/2}$$
$$= \frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \leq \delta(d_1, d_2, \Omega, r, \sigma, \beta)$$

for all $\mathbf{M} \in K$, where δ is some function of the problem parameters.

for a fixed sampling pattern, the algorithm $\mathbf{A}$ must work simultaneously for *all* relevant matrices $\mathbf{M}$.

Remark 4 (Strictly Positive $\mathbf{W}$*):* Notice that the requirement that W_{ij} be strictly greater than zero (rather than possibly equal to zero) is without loss of generality. Indeed, if $W_{ij} = 0$ for some (i, j), then either the i'th row or the j'th column of $\mathbf{W}$ are zero, and we can reduce the problem to a smaller one by ignoring that row or column.

III. Results

In this section, we state informal versions of our results. We assume that $d_1 = d_2 = d$ to make the results easier to parse, although most of our results extend to rectangular matrices.[2] We present more detailed statements of our results later in the paper.

A. General Results

Our main upper bounds give two algorithms for weighted universal matrix completion. These are formally stated in Theorems 15 and 16, and we give an informal version below. Our bounds will depend on two parameters of the sampling pattern Ω and the weight matrix $\mathbf{W}$. For a fixed $\mathbf{W}$ and Ω, define

$$\lambda = \left\|\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)} \circ \mathbf{1}_\Omega\right\| \tag{1}$$

$$\mu^2 = \max\{\max_i \left(\sum_j \frac{\mathbf{1}_{(i,j)\in\Omega}}{W_{ij}}\right), \tag{2}$$

$$\max_j \left(\sum_i \frac{\mathbf{1}_{(i,j)\in\Omega}}{W_{ij}}\right)\}. \tag{3}$$

The parameter λ is a measure of how "close" $\mathbf{1}_\Omega$ is to the matrix $\mathbf{W}$. Indeed, if $\mathbf{1}_\Omega$ happens to be rank 1 and $\mathbf{1}_\Omega = \mathbf{W}$, then $\lambda = 0$.

The parameter μ measures how "close" $\mathbf{1}_\Omega$ is to $\mathbf{W}$, as well as capturing how "lopsided" they are. If $\mathbf{1}_\Omega = \mathbf{W}$ then μ^2 is just the max column or row weight of Ω. However, if W is very different from $\mathbf{1}_\Omega$, for example, by putting not very much weight on a row that is heavily sampled by Ω, then μ will be larger.

We study two algorithms. The first, which applies when $\mathbf{M}$ is exactly rank k, is a simple debiased projection-based method. More precisely, we will estimate a rank-r matrix $\mathbf{M}$ from the observations $\mathbf{Y}_\Omega = \mathbf{M}_\Omega + \mathbf{Z}_\Omega$ by

$$\hat{\mathbf{M}}_0 =$$
$$\mathbf{W}^{(-1/2)} \circ \operatorname{argmin}_{\mathrm{rank}(\mathbf{X})=r} \left\|\mathbf{X} - \mathbf{W}^{(-1/2)} \circ (\mathbf{Y}_\Omega)\right\|.$$

Note that computing this solution only requires computing a (truncated) SVD followed by a matrix Hadamard product, so it is quite computationally efficient (cubic or better) [19]. In Theorem 15, we will show the following.

Theorem 5 (General Upper Bound for Rank-k Matrices, Informal): Let $\mathbf{W} \in \mathbb{R}^{d \times d}$ be a rank-one matrix with strictly positive entries, and fix $\Omega \subseteq [d] \times [d]$. Suppose that $\mathbf{M} \in \mathbb{R}^{d \times d}$ has rank r and $\|\mathbf{M}\|_\infty \leq \beta$, and let $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$ where the entries of $\mathbf{Z}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Then with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}_0)\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim \frac{\beta r \lambda + \sigma\mu\sqrt{r\log(d)}}{\left\|\mathbf{W}^{(1/2)}\right\|_F}.$$

The second algorithm applies when $\mathbf{M} \in \beta\sqrt{r}B_{\max}$ is approximately low-rank. Let

$$\hat{\mathbf{M}}_1 = \mathbf{W}^{(-1/2)} \circ \operatorname{argmin}_{\|\mathbf{X}\|_{\max} \leq \beta\sqrt{r}}$$
$$\left\|\mathbf{X} - \mathbf{W}^{(-1/2)} \circ (\mathbf{M}_\Omega + \mathbf{Z}_\Omega)\right\|.$$

We note that this estimator can be computed using semidefinite programming (SDP). Recall that the max-norm of a $d_1 \times d_2$ matrix is SDP-representable [42], [60]:

$$\|\mathbf{X}\|_{\max} = \inf\{t \;:\; \text{there exists } \mathbf{W} \text{ s.t. } \mathbf{W} \succeq 0, \quad \mathbf{W}_{12} = \mathbf{X},\; \operatorname{diag}(\mathbf{W}) \leq t\}, \tag{4}$$

for any $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. Above, the auxillary variable $\mathbf{W}$ is a $d \times d$ matrix, where $d = (d_1 + d_2)$. Further, it has the block decomposition:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{12}^T & \mathbf{W}_{22} \end{pmatrix}.$$

Above, $\mathbf{W}_{11} \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{W}_{22} \in \mathbb{R}^{d_2 \times d_2}$. Define the function $f\colon \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}_+ \to \mathbb{R}^{d_1 \times d_2}$ by

$$f(\mathbf{X}_0, \varepsilon) := \operatorname{argmin}_{\|\mathbf{X}\|_{\max} \leq \varepsilon} \|\mathbf{X} - \mathbf{X}_0\|.$$

[2] The only exception are Theorems 33 and 32 which are in terms of the eigenvalues of $\mathbf{1}_\Omega$. For these bounds we assume that $\mathbf{1}_\Omega$ is square and symmetric.

In view of representation (4), we see that $f(\mathbf{X}_0, \varepsilon) = \mathbf{W}^\star_{12}$, where

$$\mathbf{W}^\star = \underset{\mathbf{W}}{\operatorname{argmin}}\Big\{ \|\mathbf{W}_{12} - \mathbf{X}_0\| \; : \; \mathbf{W} \succeq 0, \;\; \operatorname{diag}(\mathbf{W}) \leq \varepsilon \Big\}.$$

Note that the operator norm satisfies $\|\mathbf{X}\| \leq t$ if and only if $\mathbf{X}^T\mathbf{X} \preceq t^2\mathbf{I}$ and $t \geq 0$. Using Schur complements, we see that $f(\mathbf{X}_0, \varepsilon) = \mathbf{W}^\star_{12}$, where

$$(t^\star, \mathbf{W}^\star) = \underset{t, \mathbf{W}}{\operatorname{argmin}}\Big\{ t \; : \; \begin{pmatrix} t\mathbf{I} & \mathbf{W}_{12} - \mathbf{X}_0 \\ \mathbf{W}_{12}^T - \mathbf{X}_0^T & t\mathbf{I} \end{pmatrix} \succeq 0, \quad \mathbf{W} \succeq 0, \operatorname{diag}(\mathbf{W}) \leq \varepsilon \Big\}. \tag{5}$$

Thus, we see that $f(\mathbf{X}_0, \varepsilon)$ may be computed by solving an SDP over the positive semidefinite cone in $\mathbb{R}^{(d+1)\times(d+1)}$. It is well-known that SDPs are polynomial-time solvable; moreover since our estimator can be expressed as

$$\hat{\mathbf{M}}_1 = \mathbf{W}^{-1/2} \circ f(\mathbf{W}^{-1/2} \circ (\mathbf{M}_\Omega + \mathbf{Z}_\Omega), \beta\sqrt{r}),$$

we see that solving the SDP implied by (5) is the dominant cost. We mention in passing that there are other techniques that permit computing this estimator without using semidefinite programming (see section 4 in [11] and references therein).

Theorem 6 (General Upper Bound for Approximately Rank-r Matrices, Informal): Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be a rank-one matrix with strictly positive entries, and fix $\Omega \subseteq [d] \times [d]$. Suppose that $\mathbf{M} \in \mathbb{R}^{d\times d}$ has $\|\mathbf{M}\|_{\max} \leq \beta\sqrt{r}$ and let $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$ where the entries of $\mathbf{Z}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Then with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}_1)\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim \sqrt{\beta}\left(\frac{\beta r\lambda + \sigma\mu\sqrt{r\log(d)}}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\right)^{1/2}.$$

Notice that the only difference between the two guarantees is that the average weighted per-entry error bound for approximately low-rank matrices is the square root of the error bound for exactly rank-r matrices. As we will see below, this translates into the following fact: if we want the average weighter per-entry error to be at most ε, then the number of samples required for the exactly rank-r case will scale like $1/\varepsilon$, while the number of samples required for the approximately rank-r case will scale like $1/\varepsilon^2$. This quantitative behavior has been observed before (eg, in [20]), and we will also show that this dependence on ε is necessary.

Remark 7 (Computable Parameters): We note that both λ and μ are quite easy to compute. This is valuable because it means that, given a sampling pattern Ω and a desired weight matrix $\mathbf{W}^{(1/2)}$, one can quickly compute the guarantees given by Theorems 5 and 6. In contrast, common deterministic conditions which guarantee accurate recovery under random samples (for example the *restricted eigenvalue condition* [48]) are not in general computationally easy to verify.

Remark 8 (Exact Recovery): One may hope to recover exactly low-rank matrices with zero error when $\sigma = 0$. While this is possible under stronger assumptions, the mild flatness assumption that we make is not sufficient for exact recovery even if $\mathbf{M}$ is exactly low rank (this has been noted previously in the work [48, Section 3.4]). For example, suppose that M = $\beta e_1\, e_1^T$, i.e., it has one non-zero entry with value β in the upper left corner. Then M is not recoverable if the top left entry is not sampled, and the proposed recovery methods would return the zero matrix, giving a small error.

The bounds above are a bit difficult to parse: how should we think of λ and μ? As we will see below, it depends on the setting, and in particular on whether or not $\mathbf{1}_\Omega$ is "close" to $\mathbf{W}$. In order to understand the bounds below, we specialize them to two cases. In the first, we consider Ω which by construction are "close" to $\mathbf{W}$, so that λ is small. In the second, we consider sampling patterns Ω that are "far" from $\mathbf{W}$.

B. Case Study: When λ Is Small

First, we study the case when λ is small (that is, when $\mathbf{W}$ is close to $\mathbf{1}_\Omega$). Suppose that $\mathbf{W}$ has entries in $(0, 1]$. We'd like to consider a "typical" Ω that is close to $\mathbf{W}$, so we study a random sampling pattern Ω so that $(i, j) \in \Omega$ with probability W_{ij}, independently for each (i, j). Below, we will use the shorthand "$\Omega \sim \mathbf{W}$" to describe Ω that is sampled in this way.

We emphasize that even though Ω is drawn at random in this thought experiment, the goal is to understand our bounds for *deterministic* sampling matrices Ω. That is, the upper bounds are still uniform (they hold simultaneously for all appropriate matrices $\mathbf{M}$), and this model is just a way to generate matrices Ω so that λ is small, on which to test our uniform bounds. We will show that for most Ω that are close to $\mathbf{W}$ (in the above sense), the upper bound above is nearly tight. In this random setting, an essential difference between our results and much of the prior art is computability of parameters. A key sufficient condition for good matrix completion is the *restricted eigenvalue condition* [48], which holds with high probability in the random setting. However, we believe there are no known polynomial time methods to compute the parameters involved in the restricted eigenvalue condition, and so in general it cannot be verified under model uncertainties. In contrast, the parameters proposed in this paper are easily computable.

In order to make sure that an Ω drawn from this ensemble is actually close to $\mathbf{W}$, we also need to assume that the entries of $\mathbf{W}$ are not too small; in particular, that they are not smaller than $1/d$; otherwise, it is not hard to see that the parameters λ and μ can become large. In this model, this means we are assuming that there are at least $\sqrt{d}$ observations in Ω per row or column.

Our results in this setting show that, under these assumptions, our upper bounds above are nearly tight in the setting when $\beta \approx \sigma$ (that is, when the noise is on the same order as the entries of $\mathbf{M}$). The formal results are given in Theorems 25 and 26 (upper bounds for rank k and approximately rank k respectively) and in Theorems 28 and 29 (lower bounds). We summarize these informally below.

We begin with our results for exactly rank-r matrices.

Theorem 9 (Results for Rank-r Matrices When $\Omega \sim \mathbf{W}$, Informal): Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be a rank-1 matrix so that for all i, j, $1/d \leq W_{ij} \leq 1$. Choose $\Omega \sim \mathbf{W}$ as described above.

Upper bound: With probability at least $1 - O(1/d)$ over the choice of Ω, the following holds. There is an algorithm $\mathcal{A}$ so that for any rank-r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \leq \beta$, $\mathcal{A}$ returns $\hat{\mathbf{M}} = \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)$ so that with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim \sigma\sqrt{\frac{rd}{|\Omega|}}\log(d) + \beta\sqrt{\frac{r^2\ d}{|\Omega|}}\log(d).$$

Lower bound: On the other hand, with probability at least $1 - e^{-O(\|\mathbf{W}^{(1/2)}\|_F^2)}$ over the choice of Ω, for any algorithm that only sees the values $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$ and returns $\hat{\mathbf{M}}$, there is some rank r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \leq \beta$ so that with probability at least $1/2$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \gtrsim \min\left\{\sigma\sqrt{\frac{rd}{|\Omega|\log(d)}}, \beta\sqrt{\frac{d}{|\Omega|\log^3(d)}}\right\}.$$

If additionally we assume that $\mathbf{W}$ is "flat" in the sense that the largest entry is no larger than a constant times the smallest entry, then we may conclude the stronger result that

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ \left(\mathbf{M} - \hat{\mathbf{M}}\right)\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \gtrsim \min\left\{\sigma\sqrt{\frac{rd}{|\Omega|}}, \frac{\beta}{\sqrt{\log(d)}}\right\}$$

In particular, is σ is on the order of β, the upper and lower bounds are approximately the same (up to logarithmic factors and factors of r).

Next, we state our results for approximately rank-r matrices.

Theorem 10 (Results for Approximately Rank-r Matrices When $\Omega \sim \mathbf{W}$, Informal): Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be a rank-1 matrix so that for all i, j, $1/d \leq W_{ij} \leq 1$. Choose $\Omega \sim \mathbf{W}$ as described above.

Upper bound: With probability at least $1 - O(1/d)$ over the choice of Ω, the following holds. There is an algorithm $\mathcal{A}$ so that for any $d \times d$ matrix $\mathbf{M} \in \beta\sqrt{r}B_{\max}$, $\mathcal{A}$ returns $\hat{\mathbf{M}} = \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)$ so that with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim \beta\left(\frac{r^2\ d}{|\Omega|}\right)^{1/4}\log^{1/2}(d) + \sqrt{\beta\sigma}\left(\frac{rd}{|\Omega|}\right)^{1/4}\log^{1/4}(d).$$

Lower bound: On the other hand, suppose additionally that $\mathbf{W}$ is "flat," in the sense that the sense that the largest entry is no larger than a constant times the smallest entry. Then with probability at least $1 - e^{-O(\|\mathbf{W}^{(1/2)}\|_F^2)}$ over the choice of Ω, for any algorithm that only sees the values $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$ and returns $\hat{M}$, there is some $\mathbf{M} \in \beta\sqrt{r}B_{\max}$ so that with probability at least $1/2$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ \left(\hat{\mathbf{M}} - \mathbf{M}\right)\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \gtrsim \sqrt{\beta\sigma}\left(\frac{rd}{|\Omega|}\right)^{1/4}.$$

Again, this lower bound is tight up to logarithmic factors and factors of r in the case that $\sigma \approx \beta$ and $\mathbf{W}$ is reasonably "flat."

C. Case Study: When λ Is Large

Next we focus on the case when λ is large. We assume the sampling pattern is symmetric so we may consider real eigenvalues. In order to prove lower bounds here, we make a few assumptions, in particular that the top *two* eigenvectors of $\mathbf{1}_\Omega$ are "flat," in the sense that the largest element is no larger than a constant times the smallest.

Example 1: Our running example is the following extreme sampling pattern:

$$\mathbf{1}_{\Omega_t} := \begin{bmatrix} 1 & 1 & 1 & 0 & \cdots & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 1 & 1 & 1 \end{bmatrix} \in \{0,1\}^{d\times t} \tag{6}$$

where there are t ones per row and t is odd. That is, $\mathbf{1}_{\Omega_t}$ is the symmetric circulant matrix whose first row is

$$\mathbf{z} = (\underbrace{1, 1, \ldots, 1}_{(t+1)/2}, 0, 0, \ldots, 0, 0, \underbrace{1, 1, \ldots, 1}_{(t-1)/2}).$$

It is well-known that the eigenvectors of $\mathbf{1}_{\Omega_t}$ are given by the rows of the discrete cosine transform (in particular, they satisfy the flatness condition above) and that the eigenvalues are given by the elements of $\mathbf{Fz}$ where $\mathbf{F}$ is the discrete Fourier transform. In particular, the largest eigenvalue of $\mathbf{1}_{\Omega_t}$ is t, and the second largest is

$$\lambda_2(\mathbf{1}_{\Omega_t}) = \sum_{\ell=-(t-1)/2}^{(t+1)/2} \omega^\ell = \omega^{(1-t)/2}\left(\frac{\omega^{t+1} - 1}{\omega - 1}\right),$$

where

$$\omega = e^{-2\pi i/d}$$

is a primitive d'th root of unity. Now, we may compute

$$|\lambda_2(\mathbf{1}_{\Omega_t})| = \left|\frac{\omega^{t+1} - 1}{\omega - 1}\right| = \sqrt{\frac{1 - \cos(2t\pi/d)}{1 - \cos(2\pi/d)}},$$

which is at least

$$\sqrt{\frac{1 - \cos(2t\pi/d)}{1 - \cos(2\pi/d)}} \geq t\left(1 - c(t/d)^2\right)$$

for some constant c (using the Taylor expansion for cosine to bound both terms). In particular, it is quite close to $\lambda_1 = t$. This means that even if we choose $\mathbf{W}$ to be the rank-1 matrix that is as close as possible to $\mathbf{1}_{\Omega_t}$ (which in this case would be $\mathbf{W} = \frac{t}{d}\mathbf{1}\mathbf{1}^T$), we still have

$$\begin{aligned}\lambda &= \left\|\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)} \circ \mathbf{1}_\Omega\right\| \\ &= \left\|\sqrt{\frac{t}{d}}\mathbf{1}\mathbf{1}^T - \sqrt{\frac{d}{t}}\mathbf{1}_{\Omega_t}\right\| \\ &= \sqrt{\frac{d}{t}}\lambda_2 = \Theta(\sqrt{dt}).\end{aligned}$$

Thus, λ is quite large.

Now, returning the the general case (provided that the top eigenvectors of $\mathbf{1}_\Omega$ are flat), suppose that we do choose $\mathbf{W}$ to be the best rank-1 approximation to $\mathbf{1}_\Omega$. This is a reasonable choice because it is an easy-to-compute matrix which intuitively makes λ small. Under these assumptions, it is not hard to work out what happens to the upper bound, which we do in Theorem 32. We are also able to prove a lower bound in Theorem 33. We informally record these results below.

Theorem 11 (Bounds for Rank-k Matrices so That $\mathbf{1}_\Omega$ Is Balanced and Has a Big Spectral Gap, Informal): Fix $\Omega \in [d] \times [d]$. Suppose that $\mathbf{W}$ is the best rank-1 approximation to $\mathbf{1}_\Omega$ and suppose that $\mathbf{W}$ is flat in the sense that the largest entry is no larger than a constant times the smallest entry. Suppose also that the second eigenvector of $\mathbf{1}_\Omega$ is flat in the same sense. Suppose that the entries of $\mathbf{Z}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Suppose that $\mathbf{M} \in \mathbb{R}^{d\times d}$ has rank r and $\|\mathbf{M}\|_\infty \leq \beta$, and let $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$ where the entries of $\mathbf{Z}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Upper bound: There is an algorithm $\mathcal{A}$ so that for any rank-r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \leq \beta$, $\mathcal{A}$ returns $\hat{\mathbf{M}} = \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)$ so that with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim r\beta\left(\frac{\lambda_1}{\lambda_2}\right) + \sigma\sqrt{\frac{r\log(d)}{\lambda_1}}.$$

Lower bound: On the other hand, for any such algorithm $\mathcal{A}$ that only sees the values $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$ and returns $\hat{\mathbf{M}}$, there is some rank r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \leq \beta$ so that with probability at least $1/2$ over the choice of $\mathbf{Z}$,

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \gtrsim \min\left\{\frac{\beta}{\sqrt{r\log(d)}}, \sigma\sqrt{\frac{r}{\lambda_1 - \lambda_2}}\right\}.$$

We note that the lower bound and the upper bound do not match. However, the lower bound does capture *some* dependence on the gap between λ_1 and λ_2, and in particular the bounds match when this gap is very small. In particular, if $\lambda_2 = \lambda_1 - O(1)$ and $\sigma \approx \beta$ then the upper bound essentially reads

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \lesssim r\beta,$$

which is trivial given that estimating $\hat{\mathbf{M}} = \mathbf{0}$ will result in a weighted per-entry error bound of at most β. However, the lower bound shows that in this case a non-trivial guarantee is impossible: it essentially reads

$$\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \gtrsim \frac{\beta}{\sqrt{r\log(d)}}.$$

Thus, in this extreme case, the upper and lower bounds match up to polynomial factors in r and $\log(d)$.

Example 2: Returning to our example of Ω_t above, we see that

$$\lambda_1 - \lambda_2 = O\left(\frac{t^3}{d^2}\right)$$

for some constant t. This is $O(1)$ when $t = d^{2/3}$. Thus we conclude from the analysis above that for this particular sampling pattern $\Omega_{d^{2/3}}$, one cannot recover even a rank-1 matrix $\mathbf{M}$ in the presence of Gaussian noise significantly better than by just guessing $\hat{\mathbf{M}} = \mathbf{0}$. We note that in this case, the number of observations is $d^{5/3}$, which in the uniform sampling case would be more than enough to recover a rank-1 matrix.

IV. Related Work

There are two lines of work that are related to ours. The first is a line of work on *deterministic* or *universal* matrix completion. In this line of work, one asks: what guarantees can one get for a sampling pattern Ω that are simultaneously valid on *all* matrices $\mathbf{M}$. (Notice that this question is interesting even if Ω is random to begin with: there is a big difference between the universal guarantee that "with high probability, Ω is good for all $\mathbf{M}$," and the randomized guarantee that "for all $\mathbf{M}$, Ω is good with high probability.")

The second is a line of work where Ω is sampled randomly, but from biased distributions. Our first case study (when Ω is drawn according to a weight matrix $\mathbf{H}$) does give universal guarantees, but our results are also interesting from the perspective of sampling from biased distributions.

We briefly review both of these areas in more detail below.

A. Deterministic/Universal Matrix Completion

The works of Heiman *et al.* [31], Bhojanapalli and Jain [7] and Li *et al.* [44] relate the sampling pattern Ω to a graph whose adjacency matrix is given by $\mathbf{1}_\Omega$. Those works show that as long as this pattern is suitably close to an expander graph—in particular, if the deterministic sampling pattern is sufficiently uniform—then efficient recovery is possible, provided that the matrix $\mathbf{M}$ is sufficiently incoherent.

There are also works which aim to understand when there is a unique (or only finitely many) low-rank matrices $\mathbf{M}$ that can complete $\mathbf{M}_\Omega$ as a function of the sampling pattern Ω. For example, [49] gives conditions on Ω under which there are only finitely many low-rank matrices that agree with $\mathbf{M}_\Omega$, [56] give a condition under which the matrix can be locally uniquely completed. The works [4] generalizes these results to the setting where there is sparse noise added to the matrix. The works [6], [50] study when rank estimation is possible as a function of a deterministic Ω, and [5] studies when a low-rank tensor can be uniquely completed. Recently,

[16] gave a combinatorial condition on Ω which characterizes when a low-rank matrix can be recovered up to a small error in the Frobenius norm from observations in Ω and showed that nuclear norm minimization will approximately recover $\mathbf{M}$ whenever it is possible.

So far, all the works mentioned are interested in when recovery of the entire matrix (as measured by un-weighted Frobenius norm) is possible. In our work, we are also interested in the case when such recovery is *not* possible. To that end, we introduce a weighting matrix $\mathbf{H}$ to capture this. See [35] for an interesting alternative, that uses an algebraic approach to answer when an entry can be completed or not.

The notion of weights has been studied before. The work [31] of Heiman et al. shows that for any weighting matrix $\mathbf{H}$, there is a deterministic sampling pattern Ω and an algorithm that observes $\mathbf{M}_\Omega$ and returns $\hat{\mathbf{M}}$ so that $\left\|\mathbf{H} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F$ is small. Their algorithm can be informally described as finding the matrix with the smallest γ_2-norm that is correct on the observed entries. Lee and Shraibman [43] start with this framework, and study the more general class of algorithms that find the "simplest" matrix that is correct on the observed entries, where "simplest" can mean of smallest norm, smallest rank, or a broad class of definitions. We note that this algorithm is not efficient in general (e.g. if "simple" is taken to mean low-rank), although again if "simple" is taken to mean, e.g., the γ_2 norm, then this algorithm takes polynomial time by solving a semidefinite program. That work gives a way of measuring which deterministic sampling patterns Ω are good with respect to a weight matrix $\mathbf{H}$. Similar to our work, they introduce a parameter which measures the distance between the weight matrix $\mathbf{H}$ and the sampling pattern Ω. They show that if this parameter is small, then the entries of the matrix can be recovered with appropriate weights. Moreover, they show, given Ω, how to efficiently compute a weight matrix $\mathbf{H}$ so that the performance of the algorithm is optimal. Unlike our work, their parameter is a bit more complicated, and is obtained by solving a semidefinite program involving Ω.

To summarize, there are a few main differences between our work and previous work:

- We are interested in cases where it may not be easy to estimate $\mathbf{M}$ in Frobenius norm from the noisy samples $\mathbf{M}_\Omega + \mathbf{Z}_\Omega$, which was the goal in [7], [44]. (And certainly we may not be able to uniquely recover $\mathbf{M}$, as is the goal in [6], [50]).
- Our algorithms are extremely simple, compared to, say, solving multiple semidefinite programs as in [43]. In particular, we focus on debiased projection-based algorithms. These algorithms are extremely simple (computationally and intuitively) and are provably optimal in some cases when there is noise or when $\mathbf{M}$ need not be exactly low-rank. However, in the non-noisy exactly low-rank case, our algorithm need not recover the matrix exactly; the algorithm of [7] is able to do this.
- We are interested in *rank-1* weighting patterns. This is more restrictive than the works [31], [43], but allows us both to obtain more efficient algorithms and to prove lower bounds.

B. Weighted Matrix Completion and Matrix Completion From Biased Samples

Weighted matrix completion has appeared in several works (see e.g. [22], [31], [43], [48]) under the assumption of random biased sampling. The connection between weighting and biased sampling is most easily expressed in the supervised learning setup. Indeed, let $\mathbf{D} \in \mathbb{R}^{d\times d}$ encode a random distribution over matrix indices so that

$$0 \le D_{i,j} \le 1, \qquad \text{and} \qquad \sum_{i,j} D_{i,j} = 1.$$

Let one observation take the form $Y_{i,j} = M_{i,j} + Z_{i,j}$ where (i, j) is sampled randomly according to $\mathbf{D}$ and suppose you are given m independent observations of this form (allowing repetition of matrix entries). Considering squared *loss function*, the *excess risk* of an estimator $\hat{\mathbf{M}}$ is

$$\begin{aligned}\mathbb{E}&\left[(\hat{M}_{i,j} - Y_{i,j})^2 - (M_{i,j} - Y_{i,j})^2\right]\\ &= \sum_{i,j} D_{i,j}(\hat{M}_{i,j} - M_{i,j})^2\\ &= \left\|\mathbf{D}^{1/2} \circ (\hat{\mathbf{M}} - \mathbf{M})\right\|_F^2.\end{aligned}$$

Bounding the excess risk then gives a weighted error bound for the estimator.

In [48], the authors consider the case when rank$(\mathbf{D})$ = 1, which almost corresponds with the random model given in our Section III-B. They identify a certain *restricted eigenvalue condition* which holds with high probability under the random model. In the random model, there error bounds are similar to the ones in our paper. However, the *restricted eigenvalue condition* is not known to be verifiable in polynomial time. They also give a lower bound, essentially matching ours, but under the assumption of uniformly random sampling.

Several other papers consider random sampling without making the assumption that rank$(\mathbf{D})$ = 1. In [38], the authors consider nuclear-norm minimization in the case when the sampling distribution is not uniform. They give unweighted error bounds which degrade (as they should) as the sampling distribution becomes less uniform. In [11], the authors allow general sampling distribution and consider the least squares estimator under a max-norm constraint. By bounding the *Rademacher complexity* of the max-norm ball, they bound the excess risk. They are also able to extend this to the *binary* setup [12]. Much of this analysis was based on previous works analyzing the max norm [24], [62].

Although our focus is not on random sampling patterns, as we discuss below in Section III-B, our analysis does imply some interesting consequences for that setting as well. In particular, one may consider the random model $\Omega \sim \mathbf{W}$ by which we mean that the pattern Ω is obtained by sampling the (i,j)th entry with probability W_{ij}. In [48], the authors consider this random model for a rank-1 matrix $\mathbf{W}$ and show that when $\mathbf{W} \circ \mathbf{M}$ is (nearly) low rank and not too spiky, the solution $\hat{\mathbf{M}}$ of a semidefinite program (SDP) yields a small error $\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F$. They also provide a theoretical lower bound for this setting, although unlike our results below, theirs becomes trivial when the spectral gap is large. In our work however, we focus on efficient projective algorithms for recovery rather than SDPs. In addition, we are able to provide uniform results in the sense that they hold with high probability for all matrices.

Other works that consider the random model $\Omega \sim \mathbf{W}$ include [61], where in fact like our work, the authors also consider the loss function $\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F$, but without assuming $\mathbf{W}$ is low-rank. There, the authors consider random but non-uniform sampling pattern distributions (rather than fixed patterns as in our work), and establish lower bounds for this model using trace-norm minimization. In fact, [59] shows that the trace norm is a good proxy for the rank in terms of sampling guarantees in the sense that if there does happen to be a matrix $\mathbf{M}*$ that is close to $\mathbf{M}$ in this weighted Frobenius norm, then (computationally intractable) minimization with rank constraints can recover $\hat{\mathbf{M}}$ approximately as close to $\mathbf{M}$ as $\mathbf{M}*$ is, using on the order of rn samples via $\Omega \sim \mathbf{W}$. In some sense, our results can be viewed as a generalization to those of [61], in that our lower bounds hold for any algorithm and the random model can be viewed as generating a special case of our result.

Our results can also be viewed as generalizations of those that use alternative sampling strategies adapted for e.g. coherence matrices. In [18], the authors show that by sampling according to *leverage scores*, one can recover coherent matrices using nuclear norm minimization. Our setting can be cast in this framework, which discuss in detail in Section VII-D. However, this again isn't our main focus, as we are focusing on efficient algorithms and establishing universal and uniform lower bounds. Nonetheless, one can construct appropriate weight matrices $\mathbf{W}$ that yield sampling distributions related to the leverage scores. See Section VII-D for comparison and more discussion.

Other works that incorporate non-uniform sampling include [47], which proposes a graph-theoretic algorithm for matrix completion when the entries are power-law distributed. Lastly, [45] proposes a so-called *isomeric* property for sampling patterns, viewed as a generalization of low-rankness, that guarantees (exact) matrix completion. There, the authors show that uniform sampling implies the isomeric condition, but this condition is a weaker assumption than uniformity, and propose a Schatten quasi-norm induced method for recovery. We again take a different approach than these works, focusing on simple recovery and universal bounds.

V. General Upper Bounds

In this section we prove general upper bounds for weighted recovery of low-rank, or approximately low-rank, matrices from deterministic sampling patterns. We first record a few theorems that we will use in our analysis.

A. Useful Theorems

We will use the Matrix-Bernstein Inequality (Theorem 1.4 in [64]).

Theorem 12: Let $\mathbf{X}_i \in \mathbb{R}^{d \times d}$ for $i = 1, \ldots, n$ be independent, random, symmetric matrices, so that

$$\mathbb{E}\mathbf{X}_i = 0 \quad \text{and} \quad \|\mathbf{X}_i\| \leq R \quad \text{almost surely.}$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{ \left\| \sum_i \mathbf{X}_i \right\| \geq t \right\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right) \quad \text{where}$$

$$\sigma^2 = \left\| \sum_i \mathbb{E}(\mathbf{X}_i^2) \right\|.$$

We will also use the following bound about sums of random matrices with Gaussian coefficients (Theorem 1.5 in [64])

Theorem 13: Let $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ for $i = 1, \ldots, n$ be any fixed matrices, and let $g_1, \ldots, g_n$ be independent standard normal random variables. Define

$$\sigma^2 = \max\left\{ \left\| \sum_i \mathbf{X}_i \mathbf{X}_i^T \right\|, \left\| \sum_i \mathbf{X}_i^T \mathbf{X}_i \right\| \right\}.$$

Then for all $t > 0$,

$$\mathbb{P}\left\{ \left\| \sum_i g_i \mathbf{X}_i \right\| \geq t \right\} \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2}{2\sigma^2} \right).$$

We will also need the Hanson-Wright Inequality (see, e.g., [55]).

Theorem 14 (Hanson-Wright Inequality): There is some constant $c > 0$ so that the following holds. Let $\boldsymbol{\xi} \in \{0, \pm 1\}^d$ be a vector with mean-zero, independent entries, and let $\mathbf{F}$ be any matrix which has zero diagonal. Then

$$\mathbb{P}\left\{ |\boldsymbol{\xi}^T \mathbf{F} \boldsymbol{\xi}| > t \right\} \leq 2 \exp\left(-c \cdot \min\left\{ \frac{t^2}{\|\mathbf{F}\|_F^2}, \frac{t}{\|\mathbf{F}\|} \right\} \right).$$

B. Bounds for Rank-r Matrices

Theorem 15 (General Upper Bound for Rank-r Matrices): Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ have strictly positive entries, and fix $\Omega \subseteq [d_1] \times [d_2]$. Suppose that $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ has rank r. Suppose that $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$ and let

$$\hat{\mathbf{M}} = \mathbf{W}^{(-1/2)} \circ \operatorname{argmin}_{\text{rank}(\mathbf{X})=r} \left\| \mathbf{X} - \mathbf{W}^{(-1/2)} \circ (\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \right\|.$$

Then with probability at least $1 - 1/(d_1 + d_2)$ over the choice of $\mathbf{Z}$,

$$\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F \leq 2\sqrt{2}r\lambda \|\mathbf{M}\|_\infty + 4\sqrt{2}\sigma\mu\sqrt{r \log(d_1 + d_2)},$$

where λ and μ are as in (1) and (2), respectively.

Proof: Let $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$. Observe that $\mathbf{M}, \hat{\mathbf{M}}$ are both rank r and hence $\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})$ is at most rank $2r$. Thus,

$$\begin{aligned} \left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F &\leq \sqrt{2r} \left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\| \\ &\leq \sqrt{2r} \left\| \mathbf{W}^{(1/2)} \circ \hat{\mathbf{M}} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\| \\ &\quad + \sqrt{2r} \left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\| \\ &\leq 2\sqrt{2r} \left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\|, \end{aligned}$$

using the definition of $\hat{\mathbf{M}}$ in the final line. Then we bound

$$\left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\|$$

$$\begin{aligned}
&\leq \left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{M}_\Omega \right\| \\
&+ \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \\
&= \left\| \mathbf{M} \circ \left(\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)} \circ \mathbf{1}_\Omega \right) \right\| \\
&+ \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \\
&\leq \|\mathbf{M}\|_{\max} \cdot \lambda + \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \\
&\leq \sqrt{r} \, \|\mathbf{M}\|_\infty \, \lambda + \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\|,
\end{aligned}$$

using the fact that $\mathbf{M}$ is rank r and hence $\|\mathbf{M}\|_{\max} \leq \sqrt{r} \, \|\mathbf{M}\|_\infty$. Thus we conclude that

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F \qquad (7) \\
&\leq 2\sqrt{2} r \lambda \, \|\mathbf{M}\|_\infty + 2\sqrt{2r} \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\|,
\end{aligned}$$

and it remains to bound the second term. We have

$$\mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{\mathbf{1}_{(i,j)\in\Omega} Z_{ij}}{\sqrt{W_{ij}}} \mathbf{e}_i \mathbf{e}_j^T,$$

where $\mathbf{e}_i$ is the i'th standard basis vector. We may apply Theorem 13 with

$$\mathbf{X}_{ij} = \frac{\mathbf{1}_{(i,j)\in\Omega}}{\sqrt{W_{ij}}} \mathbf{e}_i \mathbf{e}_j^T.$$

We have

$$\begin{aligned}
\left\| \sum_{i,j} \mathbf{X}_{ij} \mathbf{X}_{ij}^T \right\| &= \left\| \sum_i \left(\sum_j \frac{\mathbf{1}_{(i,j)\in\Omega}}{W_{ij}} \right) \mathbf{e}_i \mathbf{e}_i^T \right\| \\
&= \max_i \sum_j \frac{\mathbf{1}_{(i,j)\in\Omega}}{W_{ij}} \leq \mu^2
\end{aligned}$$

and similarly

$$\left\| \sum_{i,j} \mathbf{X}_{ij}^T \mathbf{X}_{ij} \right\| = \max_j \sum_i \frac{\mathbf{1}_{(i,j)\in\Omega}}{W_{ij}} \leq \mu^2$$

and by Theorem 13, for any $t > 0$,

$$\mathbb{P}\left\{ \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \geq t \right\} \leq 2(d_1 + d_2) \exp\left(\frac{-t^2}{2\sigma^2 \mu^2} \right).$$

We conclude that with probability at least $1 - \frac{1}{d_1+d_2}$, we have

$$\left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \leq 2\sigma\mu\sqrt{\log(d_1 + d_2)}.$$

Plugging this into (7) proves the theorem. □

C. Bounds for Approximately Rank-r Matrices

In this section we prove a bound analogous to Theorem 15 for the case when $\mathbf{M} \in \beta\sqrt{r} B_{\max}$ is only approximately low rank. We use the same simple projection algorithm, except this time we project onto the max norm ball instead of onto the cone of rank r matrices.

Theorem 16 (General Upper Bound for Approximately Rank-r Matrices): There is a constant C so that the following holds. Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ have strictly positive entries, and fix $\Omega \subseteq [d_1] \times [d_2]$. Suppose that $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ has $\|\mathbf{M}\|_{\max} \leq \beta\sqrt{r}$. Suppose that $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$ and let

$$\begin{aligned}
\hat{\mathbf{M}} = \mathbf{W}^{(-1/2)} \circ \operatorname{argmin}_{\|\mathbf{X}\|_{\max} \leq \beta\sqrt{r}} \\
\left\| \mathbf{X} - \mathbf{W}^{(-1/2)} \circ (\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \right\|.
\end{aligned}$$

Then with probability at least $1 - 1/(d_1 + d_2)$ over the choice of $\mathbf{Z}$,

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F \\
&\leq C \cdot \left\| \mathbf{W}^{(1/2)} \right\|_F^{1/2} \left(\beta\sqrt{r\lambda} + \sqrt{\beta\sigma} \left(\mu^2 \; r \log(d_1 + d_2) \right)^{1/4} \right)
\end{aligned}$$

where λ and μ are as in (1) and (2), respectively.

Proof: Let $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$, and let $\mathbf{Q} = \mathbf{M} - \hat{\mathbf{M}}$. Then we have

$$\begin{aligned}
\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F^2 &= \langle \mathbf{W} \circ \mathbf{Q}, \mathbf{Q} \rangle \qquad (8) \\
&\leq \|\mathbf{Q}\|_{\max} \, \|\mathbf{W} \circ \mathbf{Q}\|_{\max^*}.
\end{aligned}$$

The first factor we can bound by

$$\|\mathbf{Q}\|_{\max} \leq 2\beta\sqrt{r},$$

by the assumption on $\mathbf{M}$ and the definition of $\hat{\mathbf{M}}$. For the second factor, we have

$$\begin{aligned}
&\|\mathbf{W} \circ \mathbf{Q}\|_{\max^*} \leq K_G \max_{\mathbf{a}\in\{\pm 1\}^{d_1}, \mathbf{b}\in\{\pm 1\}^{d_2}} \left\langle \mathbf{a}\mathbf{b}^T, \mathbf{W} \circ \mathbf{Q} \right\rangle \\
&= K_G \max_{\mathbf{a}\in\{\pm 1\}^{d_1}, \mathbf{b}\in\{\pm 1\}^{d_2}} \left\langle \mathbf{a}\mathbf{b}^T \circ \mathbf{W}^{(1/2)}, \mathbf{W}^{(1/2)} \circ \mathbf{Q} \right\rangle \\
&\leq K_G \max_{\mathbf{a}\in\{\pm 1\}^{d_1}, \mathbf{b}\in\{\pm 1\}^{d_2}} \left\| \mathbf{a}\mathbf{b}^T \circ \mathbf{W}^{(1/2)} \right\|_* \left\| \mathbf{W}^{(1/2)} \circ \mathbf{Q} \right\| \\
&= K_G \left\| \mathbf{W}^{(1/2)} \right\|_* \left\| \mathbf{W}^{(1/2)} \circ \mathbf{Q} \right\| \\
&= K_G \left\| \mathbf{W}^{(1/2)} \right\|_F \left\| \mathbf{W}^{(1/2)} \circ \mathbf{Q} \right\|
\end{aligned}$$

where in the last line we have used the fact that $\mathbf{W}^{(1/2)}$ is rank 1 and so the nuclear norm is equal to the Frobenius norm. Then we bound

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ \mathbf{Q} \right\| = \left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\| \\
&\leq \left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\| \\
&+ \left\| \mathbf{W}^{(1/2)} \circ \hat{\mathbf{M}} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\| \\
&\leq 2 \left\| \mathbf{W}^{(1/2)} \circ \mathbf{M} - \mathbf{W}^{(-1/2)} \circ \mathbf{Y}_\Omega \right\| \\
&\leq 2 \left\| \left(\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)} \circ \mathbf{1}_\Omega \right) \circ \mathbf{M} \right\| \\
&+ 2 \left\| \mathbf{W}^{(-1/2)} \circ \mathbf{Z}_\Omega \right\| \\
&\leq 2 \left(\|\mathbf{M}\|_{\max} \lambda + 2\sigma\mu\sqrt{\log(d_1 + d_2)} \right),
\end{aligned}$$

using in the last line the analysis from the proof of Theorem 15. The putting it together with (8), we have

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}}) \right\|_F^2 \leq \|\mathbf{Q}\|_{\max} \|\mathbf{W} \circ \mathbf{Q}\|_{\max^*} \\
&\leq 2\beta\sqrt{r} K_G \left\| \mathbf{W}^{(1/2)} \right\|_F \\
&\cdot 2 \left(\beta\sqrt{r}\lambda + 2\sigma\mu\sqrt{\log(d_1 + d_2)} \right).
\end{aligned}$$

Taking the square root and choosing C appropriately completes the proof. □

VI. General Lower Bounds

As we will see in our case studies in Sections VII and VIII, the upper bounds from Section V are tight in some situations. In order to prove lower bounds in those specific settings, in this section we give general lower bounds which can be specialized to both the exactly rank-r and the approximately-rank-r settings. Our lower bounds all rest on Fano's Inequality, which we recall below.

Theorem 17 (Fano's Inequality): Let $\mathcal{F} = \{f_0, \dots, f_n\}$ be a collection of densities on $\mathcal{X}$, and suppose that $\mathcal{A} : \mathcal{X} \to \{0, \dots, n\}$. Suppose there is some $\beta > 0$ so that for any $i \neq j$, $D_{KL}(f_i \,\|\, f_j) \leq \beta$. Then

$$\max_i \mathbb{P}_{X \sim f_i} \{\mathcal{A}(X) \neq i\} \geq 1 - \frac{\beta + \log(2)}{\log(n)}.$$

The following lemma specializes Fano's inequality to our setting.

Lemma 18: Let $K \subset \mathbb{R}^{d_1 \times d_2}$, and let $\mathcal{X} \subset K$ be a finite subset of K so that $|\mathcal{X}| > 16$. Let $\Omega \subseteq [d_1] \times [d_2]$ be a sampling pattern. Let $\sigma > 0$ and choose

$$\kappa \leq \frac{\sigma\sqrt{\log|\mathcal{X}|}}{4 \max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F},$$

and suppose that

$$\kappa\mathcal{X} \subseteq K.$$

Let $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ be a matrix whose entries $Z_{i,j}$ are i.i.d., $Z_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Let $\mathbf{H} \subseteq \mathbb{R}^{d_1 \times d_2}$ be any weight matrix.

Then for any algorithm $\mathcal{A} : \mathbb{R}^\Omega \to \mathbb{R}^{d_1 \times d_2}$ that takes as input $\mathbf{X}_\Omega + \mathbf{Z}_\Omega$ for $\mathbf{X} \in K$ and outputs an estimate $\widehat{\mathbf{X}}$ to $\mathbf{X}$, there is some $\mathbf{M} \in K$ so that

$$\begin{aligned}&\|\mathbf{H} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M}\|_F \\ &\geq \frac{\kappa}{2} \min_{\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}} \|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F\end{aligned}$$

with probability at least 1/2.

Proof: Consider the net

$$\mathcal{X}' = \{\kappa\mathbf{X} \,:\, \mathbf{X} \in \mathcal{X}\}$$

which is a scaled version of $\mathcal{X}$. By assumption, $\mathcal{X}' \subseteq K$.

Recall that the KL divergence between two multivariate Gaussians is given by

$$\begin{aligned}&D_{KL}\left(\mathcal{N}(\mu_1, \Sigma_1) \,\|\, \mathcal{N}(\mu_2, \Sigma_2)\right) \\ &= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) \\ &+ \frac{1}{2}\left\langle \Sigma_2^{-1}(\mu_2 - \mu_1), \mu_2 - \mu_1 \right\rangle.\end{aligned}$$

Specializing to $\mathbf{U}, \mathbf{V} \in \mathcal{X}'$, with $\mathbf{I} = \mathbf{I}_{\Omega \times \Omega}$,

$$\begin{aligned}&D_{KL}\left(\mathbf{U}_\Omega + \mathbf{Z}_\Omega \,\|\, \mathbf{V}_\Omega + \mathbf{Z}_\Omega\right) \\ &= D_{KL}\left(\mathcal{N}(\mathbf{U}_\Omega, \sigma^2\mathbf{I}) \,\|\, \mathcal{N}(\mathbf{V}_\Omega, \sigma^2\mathbf{I})\right) \\ &= \frac{\|\mathbf{U}_\Omega - \mathbf{V}_\Omega\|_F^2}{2\sigma^2} \\ &\leq \max_{\mathbf{X}' \in \mathcal{X}'} \frac{\|\mathbf{X}'\|_F}{\sigma^2} \\ &= \frac{\kappa^2 \max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}\|_F^2}{\sigma^2}.\end{aligned}$$

Suppose that $\mathcal{A}$ is as in the statement of the lemma. Define an algorithm $\overline{\mathcal{A}} : \mathbb{R}^\Omega \to \mathbb{R}^{d_1 \times d_2}$ so that $\overline{\mathcal{A}}(\mathbf{Y}) = \mathbf{X}$ for the unique $\mathbf{X} \in \mathcal{X}'$ so that

$$\begin{aligned}&\|\mathbf{H} \circ (\mathbf{X} - \mathcal{A}(\mathbf{Y}))\|_F \\ &< \frac{1}{2} \min_{\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}'} \|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F := \rho/2\end{aligned}$$

if it exists, and $\overline{\mathcal{A}}(\mathbf{Y}) = \mathcal{A}(\mathbf{Y})$ otherwise.

Then by Fano's inequality (Theorem 17), there is some $\mathbf{M} \in \mathcal{X}'$ so that

$$\begin{aligned}&\mathbb{P}\left\{\overline{\mathcal{A}}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \neq \mathbf{M}\right\} \\ &\geq 1 - \frac{\max_{\mathbf{X} \in \mathcal{X}'} \|\mathbf{X}_\Omega\|_F^2}{\sigma^2 \log(|\mathcal{X}| - 1)} - \frac{\log(2)}{\log(|\mathcal{X}| - 1)} \\ &= 1 - \frac{\kappa^2 \max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F^2}{\sigma^2 \log(|\mathcal{X}| - 1)} - \frac{\log(2)}{\log(|\mathcal{X}| - 1)} \\ &\geq 1 - \frac{1}{4} - \frac{\log(2)}{\log(|\mathcal{X}| - 1)} \\ &\geq 1/2,\end{aligned}$$

using the assumption that $|\mathcal{X}| \geq 16$ as well as the fact that

$$\kappa \leq \frac{\sigma\sqrt{\log|\mathcal{X}|}}{4 \max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F} \leq \frac{\sigma\sqrt{\log(|\mathcal{X}| - 1)}}{2 \max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F}.$$

If $\overline{\mathcal{A}}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \neq \mathbf{M}$, then $\|\mathbf{H} \circ \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)\|_F > \rho/2$, and so

$$\begin{aligned}&\mathbb{P}\left\{\|\mathbf{H} \circ \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M}\|_F \geq \rho/2\right\} \\ &\geq \mathbb{P}\left\{\overline{\mathcal{A}}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \neq \mathbf{M}\right\} \geq 1/2.\end{aligned}$$

Finally, we observe that

$$\begin{aligned}\frac{\rho}{2} &= \frac{1}{2} \min_{\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}'} \|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F \\ &= \frac{\kappa}{2} \min_{\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}} \|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F,\end{aligned}$$

which completes the proof. □

Our lower bounds in Sections VII and VIII will follow from Lemma 18 by choosing an appropriate net. Below we prove a general lemma about picking a net, which we will use multiple times in subsequent proofs.

Lemma 19: There is some constant c so that the following holds. Let $r, d_1, d_2 > 0$ be sufficiently large, and suppose that $d_1 \geq d_2$. Let K be the cone of rank-r matrices. Let $\mathbf{H}$ be any rank-1 weight matrix, and let $\mathbf{A}$ be any rank-1 matrix with $\|\mathbf{A}\|_\infty \leq 1$. Write $\mathbf{H} = \mathbf{h}\mathbf{g}^T$ and $\mathbf{A} = \mathbf{a}\mathbf{b}^T$, and let

$$\mathbf{z} = (\mathbf{g} \circ \mathbf{b})^{(2)} \qquad \mathbf{v} = (\mathbf{h} \circ \mathbf{a})^{(2)}.$$

Let

$$\gamma = c\sqrt{r\log(d_1 d_2)}$$

There is a net $\mathcal{X} \subseteq K \cap \gamma B_\infty \cap rB_{\max}$ so that:

1) The net has size $|\mathcal{X}| \geq N$, for

$$N = 2e\exp\left(c \cdot \mathrm{MIN}\right),$$

where MIN is the minimum of the quantities:

$$\begin{aligned}&\frac{\|\mathbf{v}\|_1^2\,\|\mathbf{z}\|_1^2}{\|\mathbf{v}\|_2^2\,\|\mathbf{z}\|_2^2},\\ &\frac{\|\mathbf{v}\|_1\,\|\mathbf{z}\|_1}{\|\mathbf{v}\|_\infty\,\|\mathbf{z}\|_2\,\sqrt{r\log(r)}},\\ &\frac{\|\mathbf{v}\|_1\,\|\mathbf{z}\|_1}{\|\mathbf{v}\|_\infty\,\|\mathbf{z}\|_\infty\, r\log(r)},\\ &\frac{r^2\,\|\mathbf{v}\|_1^2}{r\,\|\mathbf{v}\|_2^2}.\end{aligned} \tag{9}$$

2) $\|\mathbf{X}_\Omega\|_F \le \sqrt{c\cdot r}\,\|\mathbf{A}_\Omega\|_F$ for all $\mathbf{X}\in\mathcal{X}$.
3) $\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\|_F \ge \sqrt{r}\,\|\mathbf{A}\circ\mathbf{H}\|_F$ for all $\mathbf{X}\neq\mathbf{X}'\in\mathcal{X}$.

Remark 20: We do not need the assumption that $d_1\ge d_2$ for the statement of Lemma 18 to be true; however, the result is stronger if $d_1\ge d_2$, because in the cases we consider below (where $\mathbf{A},\mathbf{H}$ are "flat enough"), then the term in the minimum is dominated by $\frac{r^2\|\mathbf{v}\|_1^2}{r\|\mathbf{v}\|_2^2}\approx rd_1$. If $d_2\ge d_1$, then we may switch the roles of d_1 and d_2 in the proof below and obtain a bound that depends on d_2.

Proof: Let $\mathcal{L}\subset\{\pm1\}^{d_1\times r}$ be a set of random ±1-valued matrices chosen uniformly at random with replacement, of size $4N$.

Choose $\mathbf{R}\in\{\pm1\}^{d_2\times r}$ to be determined below. Let $\boldsymbol{\ell}_i$, for $i\in[d_1]$, denote the rows of $\mathbf{L}$, and similarly let $\mathbf{r}_i$ for $i\in[d_2]$ denote the rows of $\mathbf{R}$. Let

$$\mathcal{X}=\left\{\mathbf{A}\circ\mathbf{L}\mathbf{R}^T\,:\,\mathbf{L}\in\mathcal{L}\right\}.$$

(We note that if one wishes to prove a similar lemma for $d_2>d_1$, then we should make the net by choosing $\mathbf{R}$ at random and fixing $\mathbf{L}$.)

We begin by estimating the first requirement on $\|\mathbf{X}_\Omega\|_F$, and also the requirement that $\|\mathbf{X}\|_\infty\le\gamma$ and $\|\mathbf{X}\|_{\max}\le r$ for all $\mathbf{X}\in\mathcal{X}$. We have

$$\mathbb{E}\,\|\mathbf{X}_\Omega\|_F^2=\mathbb{E}\sum_{i,j\in\Omega}A_{ij}\left\langle\boldsymbol{\ell}_i,\mathbf{r}_j\right\rangle^2=r\,\|\mathbf{A}_\Omega\|_F^2,$$

where the expectation is over the random choice of $\mathbf{L}$. By Markov's inequality, $\mathbb{P}\left\{\|\mathbf{X}_\Omega\|_F^2>4r\,\|\mathbf{A}_\Omega\|_F^2\right\}\le 1/4$. We also have

$$\|\mathbf{X}\|_\infty=\max_{i,j\in[d_1]\times[d_2]}|A_{ij}|\,|\left\langle\boldsymbol{\ell}_i,\mathbf{r}_j\right\rangle|.$$

Now, for each i,j, $\langle\boldsymbol{\ell}_i,\mathbf{r}_j\rangle$ satisfies

$$\mathbb{P}\left\{|\left\langle\boldsymbol{\ell}_i,\mathbf{r}_j\right\rangle|\ge t\right\}\le\exp\left(\frac{-2t^2}{r}\right)$$

by Hoeffding's inequality. Using the fact that $|A_{ij}|\le1$ by assumption and a union bound over all d_1d_2 values of i,j, we conclude that

$$\mathbb{P}\left\{\|\mathbf{X}\|_\infty>\sqrt{r\log(4d_1d_2)/2}\right\}\le1/4.$$

Finally, by definition the matrices $\mathbf{X}\in\mathcal{X}$ satisfy $\|\mathbf{X}\|_{\max}\le r$, by writing

$$\mathbf{X}=(\mathbf{D_a}\mathbf{L})(\mathbf{D_b}\mathbf{R})^T$$

and observing that each row of $\mathbf{D_a}\mathbf{L}$ has ℓ_2 norm at most $\|\mathbf{a}\|_\infty\sqrt{r}\le\sqrt{r}$ and similarly for each row of $\mathbf{D_a}\mathbf{R}$.

By a union bound, for one matrix $\mathbf{X}\in\mathcal{X}$, the probability that all of $\|\mathbf{X}\|_{\max}\le r$, $\|\mathbf{X}\|_\infty\le\sqrt{r\log(4d_1d_2)/2}$ and $\|\mathbf{X}_\Omega\|_F^2\le4r\,\|\mathbf{A}_\Omega\|_F^2$ is at most $1/2$. Thus, by a Chernoff bound it follows that with high probability, at least $1-\exp(-CN)$ for some constant C, there are at least $|\mathcal{X}|/4$ matrices $\mathbf{X}\in\mathcal{X}$ so that all of these hold. Let $\tilde{\mathcal{X}}\subset\mathcal{X}$ be the set of such $\mathbf{X}$'s. The net guaranteed in the statement of the theorem will be $\tilde{\mathbf{X}}$, which in the favorable case satisfies both items 1 and 2 in the lemma, and also is contained in $K\cap\gamma B_\infty$.

Thus, we turn our attention to item 3: we will show that this holds for $\mathcal{X}$ with high probability, and so in particular it will hold for $\tilde{\mathcal{X}}$, and this will complete the proof of the lemma.

Fix $\mathbf{X}\neq\mathbf{X}'\in\mathcal{X}$, and write

$$\begin{aligned}\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\|_F^2&=\|\mathbf{H}\circ\mathbf{A}\circ(\mathbf{L}-\mathbf{L}')\mathbf{R}\|_F^2\\&=\sum_{i,j\in[d_1]\times[d_2]}H_{ij}^2\,A_{ij}^2\left\langle\boldsymbol{\ell}_i-\boldsymbol{\ell}_i',\mathbf{r}_j\right\rangle^2\\&=4\sum_{i,j\in[d_1]\times[d_2]}H_{ij}^2\,A_{ij}^2\left\langle\boldsymbol{\xi}_i,\mathbf{r}_j\right\rangle^2\end{aligned}$$

where we define $\boldsymbol{\xi}_i=\frac{1}{2}(\boldsymbol{\ell}_i-\boldsymbol{\ell}_i')$. Thus, each entry of $\boldsymbol{\xi}_i$ is independently 0 with probability $1/2$ or ±1 with probability $1/4$ each. Rearranging the terms and recalling the definitions of $\mathbf{v}$ and $\mathbf{z}$ above, we have

$$\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\|_F^2=4\sum_{i=1}^{d_1}v_i\boldsymbol{\xi}_i^T\mathbf{R}^T\mathbf{D_z}\mathbf{R}\boldsymbol{\xi}_i, \tag{10}$$

where $\mathbf{D_z}$ denotes the $d_2\times d_2$ diagonal matrix with $\mathbf{z}$ on the diagonal.

In order to understand (10), we need to understand the matrix $\mathbf{R}^T\mathbf{D_z}\mathbf{R}\in\mathbb{R}^{r\times r}$. The diagonal of this matrix is $\|\mathbf{z}\|_1\mathbf{I}$. We will choose the matrix $\mathbf{R}$ so that the off-diagonal terms are small. More precisely, we will choose $\mathbf{R}$ according to the following claim.

Claim 21: There is a matrix $\mathbf{R}\in\{\pm1\}^{d_2\times r}$ so that:

(a) $\left\|\mathbf{R}^T\mathbf{D_z}\mathbf{R}-\|\mathbf{z}\|_1\mathbf{I}\right\|_F^2\le2\,r^2\,\|\mathbf{z}\|_2^2$ and
(b) $\left\|\mathbf{R}^T\mathbf{D_z}\mathbf{R}-\|\mathbf{z}\|_1\mathbf{I}\right\|\le 2\left(\|\mathbf{z}\|_2\sqrt{r\log(r)}+\|\mathbf{z}\|_\infty r\log(r)\right)$.

Proof: Choose $\mathbf{R}$ at random. We will show that both (a) and (b) above happen with probability strictly greater than $1/2$, so by a union bound there exists a choice for $\mathbf{R}$ which satisfies both.

First, for (a), we compute

$$\begin{aligned}\mathbb{E}\left\|\mathbf{R}^T\mathbf{D_z}\mathbf{R}-\|\mathbf{z}\|_1\mathbf{I}\right\|_F^2&=\sum_{i\neq j}\mathbb{E}(\mathbf{e}_i\mathbf{R}^T\mathbf{D}_z\mathbf{R}\mathbf{e}_j)^2\\&=r(r-1)\,\|z\|_2^2,\end{aligned}$$

which implies by Markov's inequality that

$$\mathbb{P}\left\{\left\|\mathbf{R}^T\mathbf{D_z}\mathbf{R}-\|\mathbf{z}\|_1\mathbf{I}\right\|_F^2>2r^2\,\|\mathbf{z}\|_2^2\right\}<\frac{1}{2}.$$

For (b), we write

$$\mathbf{R}^T\mathbf{D_z}\mathbf{R}-\|\mathbf{z}\|_1\mathbf{I}=\sum_{i=1}^{d_2}z_i(\mathbf{r}_i\mathbf{r}_i^T-\mathbf{I}),$$

which is a sum of mean-zero independent random matrices, so we apply the matrix Bernstein Inequality (Theorem 12). We have for any $t > 0$,

$$\mathbb{P}\left\{\left\|\sum_i z_i(\mathbf{r}_i\mathbf{r}_i^T - \mathbf{I})\right\| > t\right\}$$
$$\leq r\exp\left(\frac{-t^2/2}{r\left\|\mathbf{z}\right\|_2^2 + rt\left\|\mathbf{z}\right\|_\infty/3}\right),$$

using the fact that $\left\|z_i(\mathbf{r}_i\mathbf{r}_i^T - \mathbf{I})\right\| \leq \left\|\mathbf{z}\right\|_\infty (r-1) \leq \left\|\mathbf{z}\right\|_\infty r$ for all i, and that

$$\left\|\mathbb{E}\sum_i z_i^2(\mathbf{r}_i\mathbf{r}_i^T - \mathbf{I})^2\right\| = \left\|\sum_i z_i^2(r-1)\mathbf{I}\right\| \leq r\left\|\mathbf{z}\right\|_2^2.$$

Choosing t as in (b) in the statement of the claim finishes the proof. □

Having chosen this matrix $\mathbf{R}$, we can now analyze the expression (10).

Claim 22: There are constants c, c' and MIN defined as in (9) so that with probability at least

$$1 - 2\exp\left(-c\cdot\text{MIN}\right) - e\cdot\exp\left(\frac{-c'r\left\|\mathbf{v}\right\|_1^2}{\left\|\mathbf{v}\right\|_2^2}\right),$$

we have

$$\left\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\right\|_F^2 \geq r\left\|\mathbf{v}\right\|_1\left\|\mathbf{z}\right\|_1.$$

Proof: We break the left hand side up into two terms:

$$\begin{aligned}\left\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\right\|_F^2 &= 4\sum_i v_i\boldsymbol{\xi}_i^T\mathbf{R}^T\mathbf{D}_\mathbf{z}\mathbf{R}\boldsymbol{\xi}_i\\ &= 4\sum_i v_i\boldsymbol{\xi}_i^T(\mathbf{R}^T\mathbf{D}_\mathbf{z}\mathbf{R} - \left\|\mathbf{z}\right\|_1\mathbf{I})\boldsymbol{\xi}_i\\ &+ 4\left\|\mathbf{z}\right\|_1\sum_i v_i\left\|\boldsymbol{\xi}_i\right\|_2^2\\ &:= (I)\ + (II)\end{aligned}$$

For the first term (I), we will use the Hanson-Wright Inequality (Theorem 14). In our case, the matrix $\mathbf{F}$ is a block-diagonal matrix consisting of d_1 blocks which are $r\times r$, where the i'th block is equal to $4\ v_i(\mathbf{R}^T\mathbf{D}_\mathbf{z}\mathbf{D} - \left\|\mathbf{z}\right\|_1\mathbf{I})$. The Frobenius norm of this matrix is bounded by

$$\begin{aligned}\left\|\mathbf{F}\right\|_F^2 &= 16\sum_i v_i^2\left\|\mathbf{R}^T\mathbf{D}_\mathbf{z}\mathbf{D} - \left\|\mathbf{z}\right\|_1\mathbf{I}\right\|_F^2\\ &\leq 32\ r^2\left\|\mathbf{v}\right\|_2^2\left\|\mathbf{z}\right\|_2^2.\end{aligned}$$

The operator norm of $\mathbf{F}$ is bounded by

$$\begin{aligned}\left\|\mathbf{F}\right\| &= \left\|\mathbf{v}\right\|_\infty\left\|\mathbf{R}^T\mathbf{D}_\mathbf{z}\mathbf{D} - \left\|\mathbf{z}\right\|_1\mathbf{I}\right\|_\infty\\ &\leq 2\left\|\mathbf{v}\right\|_\infty\left(\left\|\mathbf{z}\right\|_2\sqrt{r\log(r)} + \left\|\mathbf{z}\right\|_\infty r\log(r)\right).\end{aligned}$$

Thus, the Hanson-Wright inequality implies that

$$\mathbb{P}\left\{(I)\ > t\right\} \leq 2\exp\left(-c\cdot\text{MIN'}\right),$$

where MIN' is the minimum of the quantities

$$\frac{t^2}{32\ r^2\left\|\mathbf{z}\right\|_2^2\left\|\mathbf{v}\right\|_2^2},$$

$$\frac{t}{\left\|\mathbf{v}\right\|_\infty\left(\left\|\mathbf{z}\right\|_2\sqrt{r\log(r)} + \left\|\mathbf{z}\right\|_\infty r\log(r)\right)}.$$

Plugging in $t = \frac{r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1}{2}$, and replacing the constant c with a different constant c', we have

$$\mathbb{P}\left\{(I)\ > \frac{r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1}{2}\right\} \leq 2\exp\left(-c'\cdot\text{MIN}\right) \tag{11}$$

Next we turn to the second term (II). We write

$$(II) = 4\left\|\mathbf{z}\right\|_1\sum_i v_i\left(\left\|\boldsymbol{\xi}_i\right\|_2^2 - \frac{r}{2}\right) + 2r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1$$

and bound the error term $4\left\|\mathbf{z}\right\|_1\sum_i v_i\left(\left\|\boldsymbol{\xi}_i\right\|_2^2 - r/2\right)$ with high probability. Observe that for each i, $\left\|\boldsymbol{\xi}_i\right\|_2^2 - r/2$ is a mean-zero subgaussian random variable, which satisfies for all $t > 0$ that

$$\mathbb{P}\left\{\left|\ \left\|\boldsymbol{\xi}_i\right\|_2^2 - r/2\right| > t\right\} \leq \exp\left(\frac{-c''\cdot t^2}{r}\right)$$

for some constant c''. Thus by a version of Hoeffding's inequality (e.g., Proposition 5.10 in [65]), for any $t > 0$ we have

$$\mathbb{P}\left\{\left|\sum_i v_i\left\|\boldsymbol{\xi}_i\right\|_2^2 - \frac{\left\|\mathbf{v}\right\|_1 r}{2}\right| > t\right\} \leq e\cdot\exp\left(\frac{-c'''\cdot t^2}{r\left\|\mathbf{v}\right\|_2^2}\right)$$

for some other constant c'''. Thus,

$$\begin{aligned}&\mathbb{P}\left\{\left|(II) - 2r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1\right| > \frac{r\left\|\mathbf{v}\right\|_1\left\|\mathbf{z}\right\|_1}{2}\right\}\\ &= \mathbb{P}\left\{4\left\|\mathbf{z}\right\|_1\left|\sum_i v_i\left(\left\|\boldsymbol{\xi}_i\right\|_2^2 - \frac{r}{2}\right)\right| > \frac{r\left\|\mathbf{v}\right\|_1\left\|\mathbf{z}\right\|_1}{2}\right\}\\ &= \mathbb{P}\left\{\left|\sum_i v_i\left(\left\|\boldsymbol{\xi}_i\right\|_2^2 - \frac{r}{2}\right)\right| > \frac{r\left\|\mathbf{v}\right\|_1}{8}\right\}\\ &\leq e\cdot\exp\left(\frac{-c'''r^2\left\|\mathbf{v}\right\|_1^2}{8r\left\|\mathbf{v}\right\|_2^2}\right).\end{aligned} \tag{12}$$

In the favorable case of both (11) and (12), we conclude that

$$\begin{aligned}&\left\|\mathbf{H}\circ(\mathbf{X}-\mathbf{X}')\right\|_F^2 = (I)\ + (II)\\ &\geq 2r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1 - \left|(II) - 2r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1\right| - \left|(I)\ \right|\\ &\geq r\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1,\end{aligned}$$

as desired. □

Now a union bound over all of the points in $\mathcal{X}$ establishes items 1 and 3 of the lemma, along with the observation that $\left\|\mathbf{z}\right\|_1\left\|\mathbf{v}\right\|_1 = \left\|\mathbf{H}\circ\mathbf{A}\right\|_F^2$. □

VII. Case Study: When λ Is Small

The point of this section is to examine our general bounds from Sections V and VI in the case when the parameter $\lambda = \left\|\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)}\circ\mathbf{1}_\Omega\right\|$ is small. One case where this happens is the following.

Let $\mathbf{W}\in\mathbb{R}^{d_1\times d_2}$ be a rank-1 matrix so so that every entry of $\mathbf{W}$ satisfies

$$W_{ij} \in \left[\frac{1}{\sqrt{d_1 d_2}}, 1\right].$$

Now we'd like to consider some Ω that is "close" to $\mathbf{W}$ in the sense that λ is small. One way to get such an Ω is to draw it randomly so that $(i,j) \in \Omega$ with probability W_{ij}. As we will show below, in this case $\mathbf{W} \approx \mathbf{1}_\Omega$, and in particular λ will be small.[3]

We emphasize that even though Ω is drawn at random in this thought experiment, the goal is to understand our bounds for *deterministic* sampling matrices Ω. That is, the upper bounds are still uniform (they hold simultaneously for all appropriate matrices $\mathbf{M}$), and this model is just a way to generate matrices Ω so that λ is small, to test our uniform bounds on. We will show that for most Ω that are close to $\mathbf{W}$ (in the above sense), our upper and lower bounds from Sections V and VI nearly match.

A. Upper Bounds

In this section we specialize Theorem 15 to the case where Ω is drawn randomly proportional to $\mathbf{W}$, as discussed above.

We begin with some bounds on the parameters λ and μ in this case.

Lemma 23: Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-1 matrix so that for all $i, j \in [d_1] \times [d_2]$, $W_{ij} \in [1/\sqrt{d_1 d_2}, 1]$. Suppose that $\Omega \subseteq [d_1] \times [d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i,j) \in \Omega$ with probability W_{ij}, independently for each (i,j). Then with probability at least $1 - 3/(d_1 + d_2)$ over the choice of Ω, we have

$$\lambda \le 2\sqrt{d_1 + d_2}\log(d_1 + d_2)$$

and

$$\mu \le 2\sqrt{(d_1 + d_2)\log(d_1 + d_2)},$$

where λ and μ are as in (1) and (2).

Proof: Fix $i \in [d_1]$. Bernstein's inequality yields

$$\mathbb{P}\left\{\sum_{j=1}^{d_2} \frac{\mathbf{1}_{(i,j)\in\Omega}}{w_i u_j} - d_2 > 2\sqrt{2}(d_1 + d_2)\log(d_1 + d_2)\right\} \le \frac{1}{(d_1 + d_2)^2}.$$

Hence, by taking a union bound,

$$\mathbb{P}\left\{\max_{i\in[d_1]} \sum_{j=1}^{d_2} \frac{\mathbf{1}_{(i,j)\in\Omega}}{w_i u_j} > 4(d_1 + d_2)\log(d_1 + d_2)\right\} \le \frac{d_1}{(d_1 + d_2)^2} \le \frac{1}{d_1 + d_2}.$$

A similar argument gives the bound

$$\mathbb{P}\left\{\max_{j\in[d_2]} \sum_{i=1}^{d_1} \frac{\mathbf{1}_{(i,j)\in\Omega}}{w_i u_j} > 4(d_1 + d_2)\log(d_1 + d_2)\right\} S \le \frac{1}{d_1 + d_2}.$$

Combining these two inequalities we have

$$\mu \le 2\sqrt{(d_1 + d_2)\log(d_1 + d_2)} \tag{13}$$

with probability at least $1 - 2/(d_1 + d_2)$.

To bound $\lambda = \|\mathbf{W}^{(-1/2)} \circ (\mathbf{W} - \mathbf{1}_\Omega)\|$, put $\gamma_{ij} = (1/\sqrt{w_i u_j})(w_i u_j - \mathbf{1}_{(i,j)\in\Omega})$, $\mathbf{X}_{ij} = \gamma_{ij}\mathbf{e}_i\mathbf{e}_j^T$, and write

$$\mathbf{S} := \mathbf{W}^{(-1/2)} \circ (\mathbf{W} - \mathbf{1}_\Omega) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \mathbf{X}_{ij}.$$

Set $\nu := \max(\|\mathbb{E}\mathbf{S}\mathbf{S}^T\|, \|\mathbb{E}\mathbf{S}^T\mathbf{S}\|)$. Note that

$$\mathbb{E}\mathbf{S}\mathbf{S}^T = \sum_{i=1}^{d_1}\left(\sum_{j=1}^{d_2} \mathbb{E}\gamma_{ij}^2\right)\mathbf{e}_i\mathbf{e}_i^T.$$

Since $\mathbb{E}\gamma_{ij}^2 = 1 - w_i u_j \le 1$, the display above gives $\|\mathbb{E}\mathbf{S}\mathbf{S}^T\| \le d_2$. Similarly, $\|\mathbb{E}\mathbf{S}^T\mathbf{S}\| \le d_1$, and so $\nu \le d_1 + d_2$. Furthermore, with probability 1, $|\gamma_{ij}| \le 2(d_1 \; d_2)^{1/4} \le \sqrt{d_1 + d_2}$ so $\|\mathbf{X}_{ij}\| \le \sqrt{d_1 + d_2}$ almost surely. Then, the matrix Bernstein Inequality (Theorem 12) gives

$$\mathbb{P}\left\{\lambda \ge 2\sqrt{d_1 + d_2}\log(d_1 + d_2)\right\} \le (d_1 + d_2)\exp\left(-\frac{2(d_1 + d_2)\log^2(d_1 + d_2)}{\nu + 2(d_1 + d_2)\log(d_1 + d_2)/3}\right) \le \frac{1}{d_1 + d_2}.$$

Thus,

$$\lambda \le 2\sqrt{d_1 + d_2}\log(d_1 + d_2), \tag{14}$$

with probability at least $1 - 1/(d_1 + d_2)$. □

Let $m = \left\|\mathbf{W}^{(1/2)}\right\|_F$. It is easy to see that $m = \mathbb{E}|\Omega|$ as well, and below we show that $|\Omega|$ is very close to m with high probability.

Lemma 24: Let $m = \left\|\mathbf{W}^{(1/2)}\right\|_F$. There is some constant C so that with probability at least $1 - 2\exp(-C \cdot m)$,

$$||\Omega| - m| \le m/4.$$

Proof: We have

$$||\Omega| - m| = \left|\sum_{i,j}\left(\mathbf{1}_{(i,j)\in\Omega} - W_{ij}\right)\right| = \left|\sum_{i,j}\left(\mathbf{1}_{(i,j)\in\Omega} - \mathbb{E}\mathbf{1}_{(i,j)\in\Omega}\right)\right|,$$

which is the sum of mean-zero independent random varables. By Bernstein's inequality,

$$\mathbb{P}\left\{||\Omega| - m| \ge m/4\right\} \le 2\exp\left(\frac{-m^2/32}{\|\mathbf{w}\|_2^2\|\mathbf{u}\|_2^2 + m/12}\right).$$

Now using the assumption that $\mathbf{w}$ and $\mathbf{u}$ are flat, we have

$$\|\mathbf{w}\|_2^2\|\mathbf{u}\|_2^2 \le \frac{(C')^4}{d_1 d_2}\|\mathbf{w}\|_1^2\|\mathbf{u}\|_1^2 = \frac{(C')^4\, m^2}{d_1 d_2} \le (C')^4\, m,$$

which proves the claim after choosing $C = \frac{1}{32((C')^4 + 12}$. □

With these computations out of the way, we may apply Theorems 15 and 16 in this setting. Theorem 25 follows immediately from Theorem 15 and Lemmas 23 and 24.

[3] The reason that we make the assumption that the entries of $\mathbf{W}$ are not too small is so that λ will be small with high probability. Otherwise, this distribution is not a good case study for the "small λ" case.

Theorem 25: Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-1 matrix so that for all $i, j \in [d_1] \times [d_2]$, $W_{ij} \in [1/\sqrt{d_1 d_2}, 1]$. Suppose that $\Omega \subseteq [d_1] \times [d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i, j) \in \Omega$ with probability W_{ij}, independently for each (i, j). Then with probability at least $1 - 4/(d_1 + d_2)$ over the choice of Ω, the following holds.

There is an algorithm $\mathcal{A}$ so that for any rank-r matrix $\mathbf{M}$ with $\|\mathbf{M}\|_\infty \le \beta$, $\mathcal{A}$ returns $\hat{M} = \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)$ so that with probability at least $1 - 1/d$ over the choice of $\mathbf{Z}$,

$$\begin{aligned}\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} &\le 8\beta\sqrt{\frac{r^2(d_1+d_2)}{|\Omega|}} \log(d_1+d_2) \\ &+ 16\sigma\sqrt{\frac{r(d_1+d_2)}{|\Omega|}} \log(d_1+d_2).\end{aligned}$$

Proof: Plugging in Lemma 23 to Theorem 15 shows that

$$\begin{aligned}&\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F \\ &\le 4r\sqrt{d_1+d_2}\log(d_1+d_2)\,\|\mathbf{M}\|_\infty \\ &+ 8\sigma\sqrt{r(d_1+d_2)}\log(d_1+d_2).\end{aligned}$$

The result follows by using $\|\mathbf{M}\|_\infty \le \beta$, and by normalizing and using Lemma 24 to replace $\left\|\mathbf{W}^{(1/2)}\right\|_F$ with $|\Omega|$. □

Similarly, Theorem 26 below follows immediately from Theorem 16 and Lemmas 23 and Lemma 24.

Theorem 26: There is some constant C so that the following holds. Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-1 matrix so that for all $i, j \in [d_1] \times [d_2]$, $W_{ij} \in [1/\sqrt{d_1 d_2}, 1]$. Suppose that $\Omega \subseteq [d_1] \times [d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i, j) \in \Omega$ with probability W_{ij}, independently for each (i, j). Then with probability at least $1 - 4/(d_1 + d_2)$ over the choice of Ω, the following holds.

There is an algorithm $\mathcal{A}$ so that for $d_1 \times d_2$ matrix $\mathbf{M} \in \beta\sqrt{r}B_{\max}$, $\mathcal{A}$ returns $\hat{M} = \mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)$ so that with probability at least $1 - 1/(d_1 + d_2)$ over the choice of $\mathbf{Z}$,

$$\begin{aligned}\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} &\le C\beta\left(\frac{r^2(d_1+d_2)}{m}\right)^{1/4} \log^{1/2}(d_1+d_2) \\ &+ C\sqrt{\beta\sigma}\left(\frac{r(d_1+d_2)}{m}\right)^{1/4} \log^{1/4}(d_1+d_2).\end{aligned}$$

B. Lower Bound for Exactly Rank r Matrices

In this section, we will prove a lower bound in the case where Ω is drawn proportionally to $\mathbf{W}$. We begin with a warm-up result for "flat" weight matrices $\mathbf{W}$.

Lemma 27 (Lower Bound for Low-Rank Matrices When $\mathbf{W}$ *is Flat and* $\Omega \sim \mathbf{W}$*):* Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-1 matrix with strictly positive entries and with $\|\mathbf{W}\|_\infty \le 1$. Suppose that there is some constant C' so that

$$\max_i |w_i| \le C' \min_i |w_i| \text{ and } \max_i |u_i| \le C' \min_i |u_i|.$$

Suppose that $\Omega \subseteq [d_1] \times [d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i, j) \in \Omega$ with probability W_{ij}, independently for each (i, j). Then with probability at least $1 - \exp(-C \cdot m)$ over the choice of Ω, the following holds:

Let $\sigma > 0$, let $0 < r < \left(\frac{\min\{d_1,d_2\}}{\log(d_1 d_2)}\right)^{1/3}$, and let $K \subset \mathbb{R}^{d_1 \times d_2}$ be the cone of rank r matrices. For any algorithm $\mathcal{A} : \mathbb{R}^\Omega \to \mathbb{R}^{d_1 \times d_2}$ that takes as input $\mathbf{X}_\Omega + \mathbf{Z}_\Omega$ and outputs a guess $\hat{\mathbf{X}}$ for $\mathbf{X}$, for $\mathbf{X} \in K \cap \beta B_\infty$ and $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$ there is some $\mathbf{M} \in K \cap \beta B_\infty$ so that

$$\begin{aligned}&\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \cdot \left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F \\ &\ge c\min\left\{\sigma\sqrt{\frac{r\max\{d_1,d_2\}}{|\Omega|}}, \frac{\beta}{\sqrt{\log(d_1 d_2)}}\right\}\end{aligned}$$

with probability at least $1/2$ over the randomness of $\mathcal{A}$ and the choice of $\mathbf{Z}$. Above, c, C are constants which depend only on C'.

Proof: Suppose without loss of generality that $d_1 \ge d_2$, and that $\log d_1$ and $\log d_2$ are integers (if not, replace them by their floors.) Let $m = \left\|\mathbf{W}^{(1/2)}\right\|_F = \|\mathbf{w}\|_1 \|\mathbf{u}\|_1$, so that $\mathbb{E}|\Omega| = m$.

Next, we instantiate Lemma 19 with $\mathbf{H} = \mathbf{W}^{(1/2)}$ and $\mathbf{A} = \mathbf{1}\mathbf{1}^T$. In the language of that lemma, we have

$$\begin{aligned}\mathbf{v} &= (\mathbf{h} \circ \mathbf{a})^{(2)} = \mathbf{h}^{(2)} = \mathbf{w} \\ \mathbf{z} &= (\mathbf{g} \circ \mathbf{b})^{(2)} = \mathbf{g}^{(2)} = \mathbf{u}.\end{aligned}$$

Let $\mathcal{X}$ be the net guaranteed by Lemma 19. We have

$$\max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F \le \sqrt{cr}\,\|\mathbf{A}_\Omega\|_F = \sqrt{cr|\Omega|} = \sqrt{c'rm} \tag{15}$$

for some constant c', using Lemma 24. We also have

$$\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{X} - \mathbf{X}')\right\|_F \ge \sqrt{r}\left\|\mathbf{W}^{(2)}\right\|_F = \sqrt{rm} \tag{16}$$

for all $\mathbf{X} \ne \mathbf{X}' \in \mathcal{X}$, using the definition of m. We have

$$\max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}\|_\infty \le c\sqrt{r\log(d_1 d_2)}. \tag{17}$$

And finally, again using the assumption that $\mathbf{w}$ and $\mathbf{u}$ are flat, the size of the net is (as in (9)),

$$\begin{aligned}N &= 2e\exp(c \cdot \text{MIN}) \\ &\ge 2e\exp\left(\frac{c}{(C')^4}\text{MIN''}\right)\end{aligned}$$

where MIN" is the minimum over the quantities

$$d_1 d_2, \frac{d_1\sqrt{d_2}}{\sqrt{r\log(r)}}, \frac{d_1 d_2}{r\log(r)}, r\ d_1$$

which yields $N \ge \exp(C'' r d_1)$.

In the last line we have used the assumption that r is not too large compared to d_2, and a suitable choice of C'' which depends on c and C'.

Now we can use this net in Lemma 18. We choose

$$\kappa = \min\left\{c''\sigma\sqrt{\frac{d_1}{m}}, \frac{\beta}{c\sqrt{r\log(d_1 d_2)}}\right\},$$

where $c'' = \frac{1}{4}\sqrt{C''/c'}$ depends on previous constants.

Observe that by (17) we have

$$\frac{\sigma\sqrt{\log|\mathcal{X}|}}{4\max_{\mathbf{X}\in\mathcal{X}}\|\mathbf{X}_\Omega\|_F} \geq \frac{\sigma\sqrt{C''rd_1}}{4\sqrt{c'rm}} \geq \kappa,$$

so this is a legitimate choice of κ in Lemma 18.

Next, we verify that $\kappa\mathcal{X} \subseteq K_r \cap \beta B_\infty$. Indeed, we have

$$\kappa \max_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X}\|_\infty \leq \kappa c\sqrt{r\log(d_1d_2)} \leq \beta,$$

so $\kappa\mathcal{X} \subseteq \beta B_\infty$, and every element of $\mathcal{X}$ has rank r by construction.

Then Lemma 18 concludes that if $\mathcal{A}$ must work on $K_r \cap \beta B_\infty$, then there is a matrix $\mathbf{M} \in K_r \cap \beta B_\infty$ so that

$$\begin{aligned}&\left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F \\ &\geq \frac{\kappa}{2}\min_{\mathbf{X}\neq\mathbf{X}'\in\mathcal{X}} \left\|\mathbf{W}^{(1/2)} \circ (\mathbf{X} - \mathbf{X}')\right\|_F \\ &\geq \frac{1}{2}\min\left\{c''\sigma\sqrt{\frac{d_1}{m}}, \frac{\beta}{c\sqrt{r\log(d_1d_2)}}\right\}\sqrt{rm} \\ &= \frac{1}{2}\min\left\{c''\sigma\sqrt{rd_1}, \frac{\beta\sqrt{m}}{c\sqrt{\log(d_1d_2)}}\right\},\end{aligned}$$

using (16). Normalizing by $\left\|\mathbf{W}^{(1/2)}\right\|_F$ and applying Lemma 24 completes the proof. □

Next, we use Lemma 27 to prove a bound that does not require that the $\mathbf{W}$ be flat; that is, Theorem 28 below is similar to Lemma 27, but we will not require the "flatness" assumption that $\max_i |w_i| \leq C'\min_i |w_i|$ or $\max_i |u_i| \leq C'\min_i |u_i|$. However, we do have to impose an additional restriction that the entries of $\mathbf{u}$ and $\mathbf{w}$ are not smaller than $1/\sqrt{d_1}$ and $1/\sqrt{d_2}$ respectively; this is the same restriction we have for the upper bounds.

Our final lower bound for the case when λ is small is the following.

Theorem 28 (Lower Bound for Rank-k Matrices When $\Omega \sim \mathbf{W}$): Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1\times d_2}$ be a rank-1 matrix with $\|\mathbf{W}\|_\infty \leq 1$, so that

$$\left\|\mathbf{w}^{(-1)}\right\|_\infty \leq \sqrt{d_1} \quad \text{and} \quad \left\|\mathbf{u}^{(-1)}\right\|_\infty \leq \sqrt{d_2}.$$

Let $d = \sqrt{d_1d_2}$. Suppose that $\Omega \subseteq [d_1]\times[d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i,j) \in \Omega$ with probability W_{ij}, independently for each (i,j). Let $m = \|\mathbf{w}\|_1\|\mathbf{u}\|_1$, so that $\mathbb{E}|\Omega| = m$. Then with probability at least $1 - \exp(-C\cdot m)$ over the choice of Ω, the following holds:

Let $\sigma, \beta > 0$, let $0 < r < \left(\frac{\min\{d_1,d_2\}}{\log^2(d)}\right)^{1/3}$, and let $K_r \subset \mathbb{R}^{d_1\times d_2}$ be the cone of rank r matrices. For any algorithm $\mathcal{A}: \mathbb{R}^\Omega \to \mathbb{R}^{d_1\times d_2}$ that takes as input $\mathbf{X}_\Omega + \mathbf{Z}_\Omega$ and outputs a guess $\hat{\mathbf{X}}$ for $\mathbf{X}$, for $\mathbf{X} \in K \cap \beta B_\infty$ and $Z_{ij} \sim \mathcal{N}(0,\sigma^2)$ there is some $\mathbf{M} \in K_r \cap \beta B_\infty$ so that

$$\begin{aligned}&\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\|\mathbf{W}\circ(\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega))\|_F \\ &\geq c\min\left\{\sigma\sqrt{\frac{r\max\{d_1,d_2\}}{m\log(d)}}, \beta\sqrt{\frac{d}{m\log^3(d)}}\right\}.\end{aligned}$$

with probability at least $1/2$ over the randomness of $\mathcal{A}$ and the choice of $\mathbf{Z}$. Above, c, C are constants which depend only on C'.

Proof: Suppose without loss of generality that $d_1 \geq d_2$. Write

$$\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{\log(d_1)/2}) \quad \text{where} \quad \mathbf{w}_i \in \mathbb{R}^{s_i},$$

so that all entries of $\mathbf{w}_i$ are in $[2^{-i}, 2^{1-i}]$ for all i. (Here, we assume without loss of generality that the coordinates of $\mathbf{w}$ are arranged in decreasing order). Notice that this is possible because the $\max_i |w_i| \leq 1$ and $\min_i |w_i| \geq \frac{1}{\sqrt{d_1}}$ by assumption. Similarly write

$$\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{\log(d_2)/2}) \quad \text{where} \quad \mathbf{u}_i \in \mathbb{R}^{t_i},$$

so that all entries of $\mathbf{u}_i$ are in $[2^{-i}, 2^{1-i}]$.

Now there is some i and j so that

$$s_i \geq \frac{2d_1}{\log(d_1)} \qquad t_j \geq \frac{2d_2}{\log(d_2)}.$$

Now consider $\tilde{\mathbf{W}} = \mathbf{w}_i\mathbf{u}_j^T \in \mathbb{R}^{s_i\times t_j}$, and let $\tilde{\Omega}$ be the restriction of Ω the the s_i rows and t_j columns corresponding to $\tilde{W}$. Notice that $\|\mathbf{w}_i\|_\infty \leq 2\left\|\mathbf{w}_i^{(-1)}\right\|_\infty^{-1}$ and $\|\mathbf{u}_j\|_\infty \leq 2\left\|\mathbf{u}_j^{(-1)}\right\|_\infty^{-1}$ by definition. Now we may apply Lemma 27 to the problem of recovering an $s_i \times t_j$ matrix, and conclude that there is some matrix $\mathbf{M} \in \mathbb{R}^{s_i\times t_j}$ so that

$$\begin{aligned}&\left\|\tilde{\mathbf{W}}\circ(\mathcal{A}(\mathbf{M}_{\tilde{\Omega}} + \mathbf{Z}_{\tilde{\Omega}}))\right\|_F \geq \\ &c\min\left\{\sigma\sqrt{\frac{r\max\{d_1,d_2\}}{\log(d_1d_2)}}, \frac{\left\|\tilde{\mathbf{W}}^{(1/2)}\right\|_F\beta}{\sqrt{\log(d_1d_2)}}\right\}.\end{aligned}$$

Now, we observe that

$$\left\|\tilde{\mathbf{W}}^{(1/2)}\right\|_F = \sqrt{\|\mathbf{w}_i\|_1\|\mathbf{u}_j\|_1} \geq \sqrt{\frac{4\sqrt{d_1d_2}}{\log(d_1)\log(d_2)}},$$

using the fact that both $\mathbf{w}_i$ and $\mathbf{u}_j$ have entries no smaller than $1/\sqrt{d_1}$ and $1/\sqrt{d_2}$ respectively. Thus, normalizing appropriately, we conclude

$$\begin{aligned}&\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\|\mathbf{W}\circ(\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega))\|_F \\ &\geq \frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\left\|\tilde{\mathbf{W}}\circ(\mathcal{A}(\mathbf{M}_{\tilde{\Omega}} + \mathbf{Z}_{\tilde{\Omega}}))\right\|_F \\ &\geq c\min\Bigg\{\sigma\sqrt{\frac{r\max\{d_1,d_2\}}{m\log(d_1d_2)}}, \\ &\beta\sqrt{\frac{\sqrt{d_1d_2}}{m\log(d_1)\log(d_2)\log(d_1d_2)}}\Bigg\}.\end{aligned}$$

The theorem follows after simplifying with the definition $d = \sqrt{d_1d_2}$. □

C. Lower Bound for Approximately Rank r Matrices

In Theorem 25 that for rank r matrices it was sufficient for $m = \left\|\mathbf{W}^{(1/2)}\right\|_F = \mathbb{E}|\Omega|$ to grow like $1/\delta^2$ in order to guarantee that the per-entry normalized error is at most δ. However, our upper bound for approximately rank-r matrices (Theorem 26) requires m to grow like $1/\delta^4$. In this section, we show that this dependence is necessary.

As with the previous section, we begin by focusing on flat matrices. Unfortunately, our lower bounds do not seem to extend in the same way to non-flat matrices without losing the correct dependence on the error. Thus, we state a result in the approximately low-rank setting only for flat matrices.

Theorem 29 (Lower Bound Approximately Low-Rank Matrices When $\mathbf{W}$ Is Flat and $\Omega \sim W$): Let $\mathbf{W} = \mathbf{w}\mathbf{u}^T \in \mathbb{R}^{d_1 \times d_2}$ be a rank-1 matrix with $\|\mathbf{W}\|_\infty \leq 1$. Suppose that there is some constant C' so that

$$\max_i |w_i| \leq C' \min_i |w_i| \text{ and } \max_i |u_i| \leq C' \min_i |u_i|.$$

Suppose that $\Omega \subseteq [d_1] \times [d_2]$ so that for each $i \in [d_1], j \in [d_2]$, $(i,j) \in \Omega$ with probability W_{ij}, independently for each (i,j). Let $m = \mathbb{E}|\Omega|$. Then with probability at least $1 - \exp(-C \cdot m)$ over the choice of Ω, the following holds:

Let $\sigma, \beta > 0$, and suppose that

$$\frac{\beta}{\sigma} \leq \frac{\min\{d_1,d_2\}^{1/3} \cdot \max\{d_1,d_2\}^{1/2}}{\sqrt{rm}\log^{2/3}(d_1 d_2)}.$$

Let $K = \beta\sqrt{r}B_{\max}$ be the max-norm ball of radius $\beta\sqrt{r}$.

For any algorithm $\mathcal{A} : \mathbb{R}^\Omega \to \mathbb{R}^{d_1 \times d_2}$ that takes as input $\mathbf{X}_\Omega + \mathbf{Z}_\Omega$ and outputs a guess $\hat{\mathbf{X}}$ for $\mathbf{X}$, for $\mathbf{X} \in K$ and $Z_{ij} \sim \mathcal{N}(0,\sigma^2)$ there is some $\mathbf{M} \in K$ so that

$$\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \cdot \left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F \geq c\sqrt{\sigma\beta}\left(\frac{d}{m}\right)^{1/4},$$

with probability at least $1/2$ over the randomness of $\mathcal{A}$ and the choice of $\mathbf{Z}$. Above, c, C are constants which depend only on C'.

Proof: The proof proceeds similarly to that of Lemma 27, except we will use Lemma 19 with a different choice of r: intuitively, we will choose a net consisting of higher-rank matrices that have smaller ℓ_∞ norm, but still have bounded max-norm.

Choose a parameter

$$s = \left\lfloor \frac{\beta}{\sigma}\sqrt{\frac{mr}{d_1}} \right\rfloor.$$

Without loss of generality, suppose that $d_1 \geq d_2$. Let $m = \left\|\mathbf{W}^{(1/2)}\right\|_F^2 = \|\mathbf{w}\|_1 \|\mathbf{u}\|_1 = \mathbb{E}|\Omega|$. We will instantiate Lemma 19 with $\mathbf{H} = \mathbf{W}^{(1/2)}$, $\mathbf{A} = \mathbf{1}\mathbf{1}^T$, and as in Lemma 27, this choice implies that $\mathbf{v} = \mathbf{w}$, $\mathbf{z} = \mathbf{u}$ in the language of Lemma 19. However, unlike in the Lemma 27, we will instantiate Lemma 19 with s in the place of the "r" parameter. Following the same analysis as before, this yields a net of size at least $N \geq \exp(C'' s d_1)$, with

$$\max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F \leq \sqrt{c' s m},$$

$$\left\|\mathbf{W}^{(1/2)} \circ (\mathbf{X} - \mathbf{X}')\right\|_F \geq \sqrt{sm},$$

$$\max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}\|_{\max} \leq s.$$

Above, we have used our condition on β/σ to ensure that $s \leq \left(\frac{\min\{d_1,d_2\}}{\log(d_1 d_2)}\right)^{1/3}$, the analog of our condition on r in Lemma 27. Now choose

$$\kappa = \min\left\{c''\sigma\sqrt{\frac{d_1}{m}}, \frac{\beta\sqrt{r}}{s}\right\}.$$

As before, we have

$$\frac{\sigma\sqrt{\log|\mathcal{X}|}}{4\max_{\mathbf{X} \in \mathcal{X}} \|\mathbf{X}_\Omega\|_F} \geq \frac{\sigma\sqrt{C'' s d_1}}{4\sqrt{c' s m}} \geq \kappa,$$

so this is a legitimate choice for κ. We also have

$$\kappa\mathcal{X} \subseteq \kappa s B_{\max} \subseteq \beta\sqrt{r}B_{\max},$$

and so this net does indeed live in K.

Thus, Lemma 18 concludes that if $\mathcal{A}$ must work on K, then there is a matrix $\mathbf{M} \in K$ so that

$$\begin{aligned}
&\left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F \\
&\geq \frac{\kappa}{2} \min_{\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}} \left\|\mathbf{W}^{(1/2)} \circ (\mathbf{X} - \mathbf{X}')\right\|_F \\
&\geq \frac{1}{2}\min\left\{c''\sigma\sqrt{\frac{d_1}{m}}, \frac{\beta\sqrt{r}}{s}\right\}\sqrt{sm} \\
&= \frac{1}{2}\min\left\{c''\sigma\sqrt{d_1\ s}, \beta\sqrt{\frac{rm}{s}}\right\}.
\end{aligned}$$

Now normalizing by $\left\|\mathbf{W}^{(1/2)}\right\|_F = \sqrt{m}$ and plugging in our choice of s, we have

$$\begin{aligned}
&\frac{\left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F}{\left\|\mathbf{W}^{(1/2)}\right\|_F} \\
&\geq c''' \cdot \min\left\{\sigma\sqrt{\frac{d_1\ s}{m}}, \beta\sqrt{\frac{r}{s}}\right\} \\
&= c'''\sqrt{\sigma\beta}\left(\frac{d_1\ r}{m}\right)^{1/4},
\end{aligned}$$

as desired. □

D. Application: Proportional Sampling

In this section, we show how to apply Theorem 26 to recover results similar to some of those in [18]. In that work, the authors show how to do matrix completion on *coherent* low-rank matrices, assuming that the observations are drawn from an appropriately biased distribution: more precisely, they show that if entries are sampled with a probability that is based on the *leverage scores* of the matrix,[4] then matrix completion is possible. That work proposes a two-stage matrix completion scheme, which first samples from uniformly random entries to estimate the SVD $\mathbf{M} \approx \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$, and then uses this to

[4] The leverage scores of a matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ with SVD $\mathbf{U\Sigma V}^T$ are the values $\|\mathbf{e}_i^T\mathbf{U}\|_2$ and $\|\mathbf{e}_j^T\mathbf{V}\|_2$ for $i \in [d_1], j \in [d_2]$

approximate the leverage scores and sample according to this approximation.

Even though the focus of this work is *deterministic* matrix completion, Theorem 26 does apply to the randomized setting as well, and this allows us to recover results similar to those of [18]. More precisely, we show below that given *any* matrix $\mathbf{X}$, if we define $\mathbf{W}$ appropriately, then we can ensure that $\mathbf{M} = \mathbf{W}^{(-1/2)} \circ \mathbf{X}$ has small max norm, and so by sampling proportional to the entries of $\mathbf{W}$, we can use Theorem 26 to obtain error bounds on

$$\|\mathbf{W}^{(1/2)} \circ (\mathbf{M} - \hat{\mathbf{M}})\|_F = \|\mathbf{X} - \mathbf{W}^{(1/2)} \circ \hat{\mathbf{M}}\|_F.$$

Thus, if we have a good enough estimate of $\mathbf{X}$ to do the sampling, we may use our algorithm from Theorem 26 to obtain $\hat{\mathbf{M}}$ and then set $\hat{\mathbf{X}} = \mathbf{W}^{(1/2)} \circ \hat{\mathbf{M}}$.

The sampling procedure that results ends up being slightly different than the leverage scores, but it can still be approximated in a two-stage algorithm using an approximate SVD $\mathbf{M} \approx \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$. The main difference between this and the corresponding theorem in [18] is that their work applies to exactly rank-r matrices, while the result below applies to any matrix, and is stated in terms of $(\|\mathbf{X}\|_* \cdot \|\mathbf{X}\|_F^{-1})^2$, a proxy for the rank.

Corollary 30: For a $d_1 \times d_2$ matrix $\mathbf{X}$ with SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and an integer m, define

$$W_{i,j} = \frac{\|\mathbf{e}_i^T \mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_2^2 \|\mathbf{e}_j^T \mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_2^2}{\|\mathbf{X}\|_*^2} m.$$

There is a randomized algorithm $\mathcal{A}$ with query access to a $d_1 \times d_2$ matrix $\mathbf{X}$ so that the following holds.

Suppose that $\mathbf{X}$ is any matrix so that $W_{i,j} \in [1/\sqrt{d_1 d_2}, 1]$. Then with probability at least $1 - \frac{4}{d_1+d_2}$, $\mathcal{A}$ queries at most

$$m \geq \frac{C}{\varepsilon^4} \left(\frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|_F} \right)^4 (d_1 + d_2) \log^2(d_1 + d_2),$$

entries of $\mathbf{X}$, and returns $\hat{\mathbf{X}}$ so that

$$\left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F \leq \varepsilon \|\mathbf{X}\|_F,$$

where C is some absolute constant. Moreover, $\mathcal{A}$ queries entries of $\mathbf{X}$ independently, querying $X_{i,j}$ with probability proportional to $W_{i,j}$.

Remark 31: We remark that Corollary 30 suggests one potential application of our work that may be an interesting future direction to analyze and develop rigorously. Indeed, a natural approach to take advantage of this result would be to truncate the $W_{i,j}$ entries if they happen to be larger than 1, leading to a possible "two-stage" scheme within this framework. Also, we note that our results are incomparable to those in [18], as discussed in Section IV, the focus of [18] is instead to recover coherent matrices using adapted sampling patterns.

Before we prove the corollary, we interpret it. We observe that the ratio $(\|\mathbf{X}\|_* \cdot \|\mathbf{X}\|_F^{-1})^2$ is a proxy for the rank of $\mathbf{X}$, in the sense that if $\mathbf{X}$ is actually rank r, then this quantity is at most r by the Cauchy-Schwarz inequality. Thus, Corollary 30 says that if $\mathbf{X}$ is not *too* coherent (in the sense that the $W_{i,j}$ are between $1/\sqrt{d_1 d_2}$ and 1), there is some way to sample $m \approx r^2 \cdot (d_1 + d_2)$ entries of matrix $\mathbf{X}$ with $(\|\mathbf{X}\|_* \|\mathbf{X}\|_F^{-1})^2 \leq r$ and use these entries to accurately reconstruct $\mathbf{X}$.

Proof: Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of $\mathbf{X}$. Then let m be as in the statement of the corollary, and let $\mathbf{W}$ be the matrix so that

$$W_{i,j} = \frac{\|\mathbf{e}_i^T \mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_2^2 \|\mathbf{e}_j^T \mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_2^2}{\|\mathbf{X}\|_*^2} m.$$

Notice that $\mathbf{W}$ is rank 1. Let

$$\mathbf{M} = \mathbf{W}^{(-1/2)} \circ \mathbf{X}.$$

Then we may write

$$\mathbf{M} = \mathbf{W}^{(-1/2)} \circ \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \frac{\|\mathbf{X}\|_*}{\sqrt{m}} \mathbf{D}_1 \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \mathbf{D}_2,$$

where $\mathbf{D}_1$ is a $d_1 \times d_1$ diagonal matrix with (i,i)-th entry $\frac{1}{\|\mathbf{e}_i^T \mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_2}$ and $\mathbf{D}_2$ is a $d_2 \times D_2$ diagonal matrix with (j,j)-th entry $\frac{1}{\|\mathbf{e}_j^T \mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_2}$. By construction, every row of $\mathbf{D}_1 \mathbf{U}\mathbf{\Sigma}^{(1/2)}$ has ℓ_2-norm at most 1, and every row of $\mathbf{D}_2 \mathbf{V}\mathbf{\Sigma}^{(1/2)}$ has ℓ_2 norm at most 1, and so we conclude that

$$\|\mathbf{M}\|_{\max} \leq \|\mathbf{W}^{(-1/2)} \circ \mathbf{X}\|_{\max} = \frac{\|\mathbf{X}\|_*}{\sqrt{m}}.$$

Now, we define the algorithm $\mathcal{A}$ as follows: sample $(i,j) \in [d_1] \times [d_2]$ with probability $W_{i,j}$, independently for each (i,j). Then estimate $\hat{\mathbf{X}}$ is given by

$$\hat{\mathbf{X}} = \mathbf{W}^{(1/2)} \circ \hat{\mathbf{M}},$$

where $\hat{\mathbf{M}}$ is the estimate guaranteed by Theorem 26.[5]

First, we observe that the expected number of samples taken by $\mathcal{A}$ is

$$\begin{aligned}\sum_{i,j} W_{i,j} &= \sum_{i,j} \frac{\|\mathbf{e}_i^T \mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_2^2 \|\mathbf{e}_j^T \mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_2^2}{\|\mathbf{X}\|_*^2} m \\ &= \frac{m \cdot \|\mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_F^2 \|\mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_F^2}{\|\mathbf{X}\|_*^2} \\ &\leq m,\end{aligned}$$

where above we used the fact that

$$\|\mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_F^2 \|\mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_F^2 = \|\mathbf{X}\|_*^2.$$

To see this, notice that

$$\begin{aligned}&\|\mathbf{U}\mathbf{\Sigma}^{(1/2)}\|_F^2 \|\mathbf{V}\mathbf{\Sigma}^{(1/2)}\|_F^2 \\ &= \left(\sum_\ell |\sigma_\ell| \, \|\mathbf{U}\mathbf{e}_\ell\|_2^2 \right) \left(\sum_r |\sigma_r| \, \|\mathbf{V}\mathbf{e}_r\|_2^2 \right) \\ &= \sum_{\ell,r} |\sigma_\ell||\sigma_r| \\ &= \left(\sum_\ell |\sigma_\ell| \right)^2 \\ &= \|\mathbf{X}\|_*^2,\end{aligned}$$

[5] Looking into the proof of that theorem, we see that we should take

$$\hat{\mathbf{X}} = \operatorname{argmin}_{\|\mathbf{Z}\|_{\max} \leq \frac{\|\mathbf{X}\|_*}{\sqrt{m}}} \|\mathbf{Z} - \mathbf{W}^{(-1)} \circ \mathbf{X}_\Omega\|.$$

using the fact that $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and hence have columns of unit norm.

Next, we observe that by Theorem 26 (with $\sigma = 0$), we have

$$\begin{aligned}\|\mathbf{X}-\hat{\mathbf{X}}\|_F &= \|\mathbf{W}^{(1/2)}\circ(\mathbf{M}-\hat{\mathbf{M}})\|_F \\ &\le C\|\mathbf{W}^{(1/2)}\|_F\|\mathbf{M}\|_{\max}\left(\frac{d_1+d_2}{m}\right)^{1/4}\sqrt{\log(d_1+d_2)} \\ &\le C\sqrt{m}\cdot\frac{\|\mathbf{X}\|_*}{\sqrt{m}}\cdot\left(\frac{d_1+d_2}{m}\right)^{1/2}\sqrt{\log(d_1+d_2)} \\ &= C\cdot\|\mathbf{X}\|_F\left(\frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|_F}\right)\left(\frac{d_1+d_2}{m}\right)^{1/4}\sqrt{\log(d_1+d_2)}.\end{aligned}$$

In particular, if

$$m \ge \frac{C^4\left(\frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|_F}\right)^4(d_1+d_2)\log^2(d_1+d_2)}{\varepsilon^4},$$

then

$$\|\mathbf{X}-\hat{\mathbf{X}}\|_F \le \varepsilon\cdot\|\mathbf{X}\|_F\,,$$

as claimed. After adjusting the constant C this implies the corollary. □

VIII. Case Study: When λ Is Large

The point of this section is to examine our general bounds from Sections V and VI in the case when the parameter $\lambda = \left\|\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)}\circ\mathbf{1}_\Omega\right\|$ is large, and to show that some dependence on this parameter is necessary. We will not be able to obtain tight bounds on the dependence on λ, but we will be able to obtain upper and lower bounds that have similar qualitative dependence on λ in some parameter regimes. We leave it as an interesting open problem to understand the "correct" dependence on λ.

While this seems like a difficult challenge in general, we are able to make progress when $\mathbf{1}_\Omega$ happens to be the adjacency matrix of a connected graph. In this case, λ is directly related to the spectral gap of the underlying graph, and there are many tools available to study it. Thus, our results below apply to this special case.

A. Upper Bound

In this section we specialize our upper bound, Theorem 15, to the case where λ is large.

Theorem 32 (Upper Bound for Rank-r Matrices in Terms of λ_1 and λ_2): Let $\Omega \subseteq [d]\times[d]$ be such that $\mathbf{1}_\Omega$ is the adjacency matrix of a connected, undirected graph on d vertices. Let $\mathbf{v}_1, \mathbf{v}_2$ be the eigenvectors of $\mathbf{1}_\Omega$ corresponding to the top two largest eigenvalues (by magntitude), λ_1, λ_2. Suppose that there is some constant C so that

$$\max_i |(v_1)_i| \le C\min_i |(v_1)_i|,$$
$$\max_i |(v_2)_i| \le C\min_i |(v_2)_i|$$

Let $\sigma > 0$ and let $\mathbf{W}$ denote the best rank-1 approximation to $\mathbf{1}_\Omega$.

Suppose that $\mathbf{M} \in \mathbb{R}^{d\times d}$ has rank r. Suppose that $Z_{ij} \sim \mathcal{N}(0,\sigma^2)$ and let

$$\hat{\mathbf{M}} = \mathbf{W}^{(-1/2)}\circ \operatorname{argmin}_{\text{rank}(\mathbf{X})=r}\left\{\left\|\mathbf{X} - \mathbf{W}^{(-1/2)}\circ(\mathbf{M}_\Omega + \mathbf{Z}_\Omega)\right\|\right\}.$$

Then with probability at least $1 - 1/2d$ over the choice of $\mathbf{Z}$,

$$\begin{aligned}&\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\left\|\mathbf{W}^{(1/2)}\circ(\mathbf{M}-\hat{\mathbf{M}})\right\|_F \\ &\quad\le c\left(r\left(\frac{\lambda_2}{\lambda_1}\right) + \sigma\sqrt{\frac{r\log(d)}{\lambda_1}}\right),\end{aligned}$$

where c is a constant that depends only on C.

Proof: To apply Theorem 15, we must compute λ, μ and $\left\|\mathbf{W}^{(1/2)}\right\|_F$ in terms of λ_1. We can explicitly write $\mathbf{W} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T$, since λ_1 is the largest eigenvalue of $\mathbf{1}_\Omega$. Of course, since $\mathbf{1}_\Omega$ defines the adjacency matrix of a connected graph, $\mathbf{v}_1$ has strictly positive entries (by the Perron-Frobenius Theorem) and hence $W_{ij} > 0$ for all i, j. Let $\bar{W} = \min_{ij} W_{ij}$, so that, using our assumption $W_{ij} \in [\bar{W}, C\bar{W}]$ for all i, j.

Since $\bar{W}$ is rank-1,

$$\lambda_1 = \|\mathbf{W}\| = \|\mathbf{W}\|_F = \sqrt{\sum_{ij} W_{ij}^2} \in [\bar{W}d, C\bar{W}d].$$

Rearranging, this means

$$\bar{W} \in \left[\frac{\lambda_1}{Cd}, \frac{\lambda_1}{d}\right].$$

Then we compute

$$\begin{aligned}\lambda &= \left\|\mathbf{W}^{(1/2)} - \mathbf{W}^{(-1/2)}\circ\mathbf{1}_\Omega\right\| \\ &\le \frac{1}{\sqrt{\bar{W}}}|\lambda_2| \le \sqrt{\frac{Cd}{\lambda_1}}|\lambda_2|\end{aligned}$$

and

$$\begin{aligned}\mu^2 &= \max\left\{\max_i\sum_j\frac{\Omega_{ij}}{W_{ij}}, \max_j\sum_i\frac{\Omega_{ij}}{W_{ij}}\right\} \\ &\le \frac{1}{\bar{W}}\max_i\|\mathbf{r}_i\|_0 \\ &\le \frac{Cd}{\lambda_1}\max_i\|\mathbf{r}_i\|_0\end{aligned}$$

where $\mathbf{r}_i$ is the i'th row of $\mathbf{1}_\Omega$, and $\|\cdot\|_0$ denotes the number of nonzero entries. Now we have, for all i,

$$\langle\mathbf{r}_i, \mathbf{v}_1\rangle = \lambda_1(v_1)_i \in [\lambda_1\bar{v}, C\lambda_1\bar{v}],$$

where $\bar{v} = \min_i(v_1)_i > 0$. We also have

$$\langle\mathbf{r}_i, \mathbf{v}_1\rangle = \sum_{j=1}^d (r_i)_j(v_1)_j \in [\bar{v}\|\mathbf{r}_i\|_0, C\bar{v}\|\mathbf{r}_i\|_0],$$

and together these imply that

$$\frac{\lambda_1}{C} \le \|\mathbf{r}_i\|_0 \le C\lambda_1$$

for all i. Thus, we can simplify the bound on μ^2 to

$$\mu^2 \le \frac{Cd}{\lambda_1}\max_i\|\mathbf{r}_i\|_0 \le C^2\ d.$$

Finally, we bound (as $C \geq 1$)

$$\left\|\mathbf{W}^{(1/2)}\right\|_F = \sqrt{\sum_{i,j} W_{ij}} \in \left[d\sqrt{\bar{W}}, Cd\ \sqrt{\bar{W}}\right] \subseteq \left[\frac{1}{C}\sqrt{\lambda_1\ d}, C\sqrt{\lambda_1\ d}\right].$$

Plugging all of these bounds into Theorem 15 proves the theorem (and $c = 4\sqrt{2}C^2$ suffices). □

B. Lower Bound

Theorem 33: Let $\Omega \subseteq [d] \times [d]$ be such that $\mathbf{1}_\Omega$ is symmetric and corresponds to the adjacency matrix of a connected undirected graph on d vertices. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be the first and second eigenvectors of $\mathbf{1}_\Omega$, (normalized so that $\|\mathbf{v}_1\|_2 = \|\mathbf{v}_2\|_2 = 1$) with corresponding eigenvalues λ_1 and λ_2. Suppose that

$$\max_i |(v_1)_i| \leq C \min_i |(v_1)_i|,$$
$$\max_i |(v_2)_i| \leq C \min_i |(v_2)_i|.$$

Let $\mathbf{W} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T$ be the best rank-1 approximation to $\mathbf{1}_\Omega$.

Let $\sigma > 0$, let $0 < r < d^{1/3}/\log^{2/3}(d)$, and let $K \subset \mathbb{R}^{d\times d}$ be the cone of rank r matrices. For any algorithm $\mathcal{A} : \mathbb{R}^\Omega \to \mathbb{R}^{d\times d}$ that takes as input $\mathbf{X}_\Omega + \mathbf{Z}_\Omega$ and outputs a guess $\hat{\mathbf{X}}$ for $\mathbf{X}$, for $\mathbf{X} \in K \cap \beta B_\infty$ and $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$, there is some $\mathbf{M} \in K \cap \beta B_\infty$ so that

$$\frac{1}{\left\|\mathbf{W}^{(1/2)}\right\|_F}\left\|\mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M})\right\|_F \geq c \min\left\{\sqrt{\frac{1}{\lambda_1 - \lambda_2}} \cdot \sigma\sqrt{r},\ \frac{\beta}{\sqrt{r\log(d)}}\right\}$$

Proof: By the Perron-Frobenius theorem, $\mathbf{v}_1 \succ \mathbf{0}$, aka, it has all strictly positive entries. Since $\mathbf{v}_1 \perp \mathbf{v}_2$, $\mathbf{v}_2$ must have some negative entries. Without loss of generality, suppose that the coordinates are ordered so that the entries of $\mathbf{v}_2$ are decreasing, and write

$$\mathbf{v}_2 = (\mathbf{h}, -\mathbf{g})$$

where $\mathbf{h}, \mathbf{g} \succeq \mathbf{0}$, $\mathbf{h} \in \mathbb{R}^s$ and $\mathbf{g} \in \mathbb{R}^t$. (This defines s, t so that $s + t = d$). Write $\mathbf{v}_1 = (\mathbf{a}, \mathbf{b})$ for $\mathbf{a} \in \mathbb{R}^s$, $\mathbf{b} \in \mathbb{R}^t$ according to the same partition of coordinates, and write

$$\mathbf{1}_\Omega = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B} \\ \mathbf{B}^T & \mathbf{B}_1 \end{pmatrix}$$

so that $\mathbf{B}_0 \in \mathbb{R}^{s\times s}$, $\mathbf{B}_1 \in \mathbb{R}^{t\times t}$ and $\mathbf{B} \in \mathbb{R}^{s\times t}$. Notice that by orthogonality,

$$0 = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{a}, \mathbf{h} \rangle - \langle \mathbf{b}, \mathbf{g} \rangle$$

and so

$$\langle \mathbf{a}, \mathbf{h} \rangle = \langle \mathbf{b}, \mathbf{g} \rangle.$$

Let

$$\alpha := 2\langle \mathbf{a}, \mathbf{h} \rangle = 2\langle \mathbf{b}, \mathbf{g} \rangle = \langle \mathbf{a}, \mathbf{h} \rangle + \langle \mathbf{b}, \mathbf{g} \rangle.$$

Claim 34:

$$\mathbf{h}^T\mathbf{B}\mathbf{g} \leq \frac{\alpha(\lambda_1 - \lambda_2)}{4}.$$

Proof: Let $\mathbf{x} = (\mathbf{h}, \mathbf{g})$, so that $\mathbf{x} \succeq \mathbf{0}$. Then

$$\mathbf{x}^T\mathbf{1}_\Omega\mathbf{x} = \mathbf{h}^T\mathbf{B}_0\mathbf{h} + \mathbf{g}^T\mathbf{B}_1\mathbf{g} + 2\mathbf{h}^T\mathbf{B}\mathbf{g}$$

and

$$\lambda_2 = \mathbf{v}_2^T\mathbf{1}_\Omega\mathbf{v}_2 = \mathbf{h}^T\mathbf{B}_0\mathbf{h} + \mathbf{g}^T\mathbf{B}_1\mathbf{g} - 2\mathbf{h}^T\mathbf{B}\mathbf{g},$$

so

$$\mathbf{x}^T\mathbf{1}_\Omega\mathbf{x} = \lambda_2 + 4\mathbf{h}^T\mathbf{B}\mathbf{g}.$$

Observing that $\alpha = \langle \mathbf{x}, \mathbf{v}_1 \rangle$, we have

$$\mathbf{x} = \alpha\mathbf{v}_1 + \left(\sqrt{1 - \alpha^2}\right)\mathbf{z},$$

for some vector $\mathbf{z}$ that satisfies $\mathbf{z} \perp \mathbf{v}_1$. Then

$$\begin{aligned} \lambda_2 + 4\mathbf{h}^T\mathbf{B}\mathbf{g} &= \mathbf{x}^T\mathbf{1}_\Omega\mathbf{x} \\ &= \alpha^2\mathbf{v}_1^T\mathbf{1}_\Omega\mathbf{v}_1 + (1 - \alpha^2)\mathbf{z}^T\mathbf{1}_\Omega\mathbf{z} \\ &\quad + 2\alpha\sqrt{1 - \alpha^2}\mathbf{v}_1^T\mathbf{1}_\Omega\mathbf{z} \\ &= \alpha^2\mathbf{v}_1^T\mathbf{1}_\Omega\mathbf{v}_1 + (1 - \alpha^2)\mathbf{z}^T\mathbf{1}_\Omega\mathbf{z} \\ &\leq \alpha^2\lambda_1 + (1 - \alpha^2)\lambda_2, \end{aligned}$$

from which we conclude that

$$4\mathbf{h}^T\mathbf{B}\mathbf{g} \leq \alpha^2(\lambda_1 - \lambda_2),$$

as desired. Above, we have used the fact that $\mathbf{v}_1^T\mathbf{1}_\Omega\mathbf{z} = \lambda_1\langle \mathbf{v}_1, \mathbf{z} \rangle = 0$ since $\mathbf{v}_1 \perp \mathbf{z}$. □

Now define

$$\mathbf{W} = \lambda_1\mathbf{v}_1\mathbf{v}_1^T$$

to be the best rank-1 approximation to $\mathbf{1}_\Omega$. Choose

$$\mathbf{A} = \frac{1}{\sqrt{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}}\left((\mathbf{h}, \mathbf{0})(\mathbf{0}, \mathbf{g})^T\right)^{(1/2)}$$

and

$$\mathbf{H} = \mathbf{W}^{(1/2)} = \sqrt{\lambda_1}\left((\mathbf{a}, \mathbf{b})(\mathbf{a}, \mathbf{b})^T\right)^{(1/2)}.$$

Then we may compute

$$\begin{aligned} \|\mathbf{H} \circ \mathbf{A}\|_F^2 &= \frac{1}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}\lambda_1\langle \mathbf{a}, \mathbf{h} \rangle \langle \mathbf{b}, \mathbf{g} \rangle \\ &= \frac{\lambda_1\alpha^2}{4\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty} \end{aligned}$$

and

$$\begin{aligned} \|\mathbf{A}_\Omega\|_F^2 &= \frac{1}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}\sum_{i=1}^{s}\sum_{j=1}^{t} B_{ij}h_ig_j \\ &= \frac{\mathbf{h}^T\mathbf{B}\mathbf{g}}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty} \\ &\leq \frac{\alpha^2(\lambda_1 - \lambda_2)}{4\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty} \end{aligned}$$

using the claim. Now we apply Lemma 19 with this choice of $\mathbf{A}$. In the language of that lemma, we have

$$\mathbf{z} = \left(\frac{\lambda_1}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}\right)(\mathbf{h} \circ \mathbf{a}, \mathbf{0}),$$
$$\mathbf{v} = \left(\frac{\lambda_1}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}\right)(\mathbf{0}, \mathbf{g} \circ \mathbf{b}).$$

By our assumption, there is some constant C so that $\max_i |h_i| \le C \min_i |h_i|$, and the same for $\mathbf{g}, \mathbf{a}, \mathbf{b}$. Thus, as in the proof of Lemma 27, Lemma 19 guarantees a net $\mathcal{X}$ so that

$$|\mathcal{X}| \ge 2e \exp(C'rd),$$

for some constant C' (which depends on C) and so that, for some constant c, for all $\mathbf{X} \in \mathcal{X}$ we have

$$\|\mathbf{X}_\Omega\|_F \le \frac{c\alpha\sqrt{r(\lambda_1 - \lambda_2)}}{\sqrt{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}},$$

for all $\mathbf{X} \neq \mathbf{X}' \in \mathcal{X}$ we have

$$\|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F \ge \frac{\sqrt{r\lambda_1}\alpha}{\sqrt{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}},$$

and finally for all $\mathbf{X} \in \mathcal{X}$ we have

$$\|\mathbf{X}\|_\infty \le c\sqrt{r\log(d)}.$$

Now we want to apply Lemma 18, and we choose

$$\kappa = \min\left\{ c'' \left(\frac{\sigma}{\alpha}\right) \sqrt{\frac{rd\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}{\lambda_1 - \lambda_2}}, \frac{\beta}{c\sqrt{r\log(d)}} \right\}.$$

for some constant c'' to be chosen below. Observe that

$$\frac{\sigma\sqrt{\log|\mathcal{X}|}}{4\max_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X}\|_F} = \frac{\sigma\sqrt{C'rd}\sqrt{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}}{2\alpha\sqrt{\lambda_1 - \lambda_2}} \ge \kappa$$

for an appropriate choice of $c'' = \sqrt{C'}/2$, so this is a legitimate choice of κ for Lemma 18. We conclude that for any algorithm $\mathcal{A}$ that works on $K \cap \beta B_\infty$, there is some matrix $\mathbf{M} \in K \cap \beta B_\infty$ so that

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M}) \right\|_F \\
&\ge \frac{\kappa}{2} \min_{\mathbf{X}\neq\mathbf{X}'\in\mathcal{X}} \|\mathbf{H} \circ (\mathbf{X} - \mathbf{X}')\|_F \\
&\ge \frac{1}{2} \min\left\{ c'' \left(\frac{\sigma}{\alpha}\right) \sqrt{rd\frac{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}{\lambda_1 - \lambda_2}}, \frac{\beta}{c\sqrt{r\log(d)}} \right\} \\
&\cdot \sqrt{\frac{\lambda_1\alpha^2}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty}} \\
&\ge c''' \min\left\{ \sqrt{\frac{\lambda_1}{\lambda_1 - \lambda_2}}\sigma\sqrt{rd}, \beta\sqrt{\frac{\lambda_1\alpha^2}{\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty r\log(d)}} \right\}
\end{aligned}$$

for some constant c'''.

Claim 35: Under the assumptions on $\mathbf{v}_1, \mathbf{v}_2$, we have

$$\frac{1}{C^2\ d} \le \|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty \le \frac{C^2}{d},$$

and

$$\alpha \ge \frac{1}{C^2}.$$

Proof: First, we observe that since $\mathbf{v}_1, \mathbf{v}_2$ have ℓ_2-norm 1 and have entries that are all about the same magnitude, up to a factor of C, we must have every entry of $\mathbf{v}_1$ and $\mathbf{v}_2$ in the interval $\left[\frac{1}{C\sqrt{d}}, \frac{C}{\sqrt{d}}\right]$. (Indeed, if one entry of $\mathbf{v}_1$ is larger than $C/\sqrt{d}$, then all entries are larger than $1/\sqrt{d}$, which contradicts the requirement on the 2-norm, and similarly if the smallest entry is smaller than $1/(C\sqrt{d})$; the same holds for $\mathbf{v}_2$). This immediately implies the claim about $\|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty$ since $\mathbf{v}_2 = (\mathbf{h}, -\mathbf{g})$.

Next, we recall that

$$\alpha = \langle \mathbf{h}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle = 2\langle \mathbf{h}, \mathbf{a} \rangle = 2\langle \mathbf{g}, \mathbf{b} \rangle,$$

and using the observation above about the size of the entries of $\mathbf{v}_1, \mathbf{v}_2$, we have

$$\alpha \ge \frac{s}{C^2\ d} + \frac{t}{C^2\ d} = \frac{1}{C^2}.$$

□

Using the claim, we bound $\alpha, \|\mathbf{h}\|_\infty \|\mathbf{g}\|_\infty$ in the second term above, and conclude that there is some $\mathbf{M}$ so that

$$\begin{aligned}
&\left\| \mathbf{W}^{(1/2)} \circ (\mathcal{A}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{M}) \right\|_F \\
&\ge c'''' \min\left\{ \sqrt{\frac{\lambda_1}{\lambda_1 - \lambda_2}} \cdot \sigma\sqrt{rd},\ \beta\sqrt{\frac{d\lambda_1}{r\log(d)}} \right\}
\end{aligned}$$

for some other constant c''''. This completes the proof, after normalizing by

$$\left\| \mathbf{W}^{(1/2)} \right\|_F = \sqrt{\lambda_1 \|\mathbf{v}_1\|_1^2} \le C\sqrt{d\lambda_1}.$$

□

IX. Experiments

In this section, we illustrate the results of numerical experiments for our debiased projection method,

$$\begin{aligned}
&\hat{\mathbf{M}}_{\text{debias}} := \\
&\mathbf{W}^{(-1/2)} \circ \operatorname{argmin}_{\operatorname{rank}(\mathbf{X})=r} \|\mathbf{X} - \mathbf{W}^{(-1/2)} \circ \mathbf{M}\|_F. \quad (18)
\end{aligned}$$

In this section, we refer to this algorithm as a *debiased*, low-rank projection. This is in contrast to a standard (see section 2.5 of [51]) low-rank projection,

$$\hat{\mathbf{M}}_{\text{standard}} := \operatorname{argmin}_{\operatorname{rank}(\mathbf{Y})=r} \left\| \mathbf{Y} - p^{-1}\mathbf{M} \right\|_F. \quad (19)$$

Above, $p := |\Omega|/d_1 d_2$. We report the performance of these two procedures for various synthetic and data-derived sampling patterns.

The purpose of these experiments is to highlight the effect of debiasing. To make this comparison, we mainly compare our debiased low-rank projection algorithm to the standard low-rank projection algorithm described above. In section IX-C, we compare our debiased max-norm projection algorithm against other potentially biased algorithms based on convex minimization.

A. Data-Derived Sampling Patterns

In our first experiments, we use sampling patterns taken from the Jester joke corpus [28], and the Movielens 100k dataset [30]. In the first dataset, users rate jokes; in the second, users rate movies. We take d_1 to be the number of users enrolled in the dataset, and d_2 to be the number of rated items (*e.g.*, jokes or films). For the Jester dataset, we have $d_1 = 73,421$ and $d_2 = 100$, For the Movielens dataset, we have $d_1 = 997$ and $d_2 = 538$. We take $\Omega \subseteq [d_1] \times [d_2]$ to be

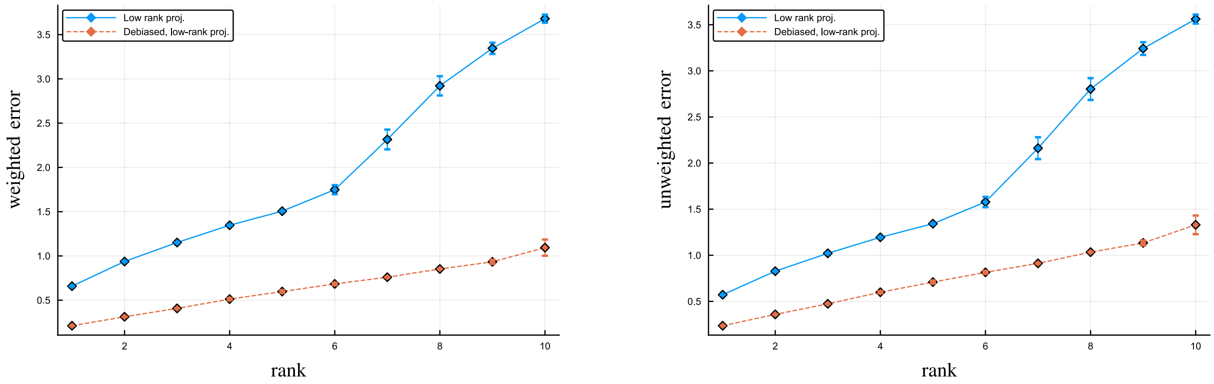

Fig. 1. Average errors of our debiased algorithm, versus standard low-projection procedure on Jester sampling pattern and standard Gaussian data. Error bars denote standard deviations over 30 samples of the noise; each line (and error bar) is averaged over 30 trials.

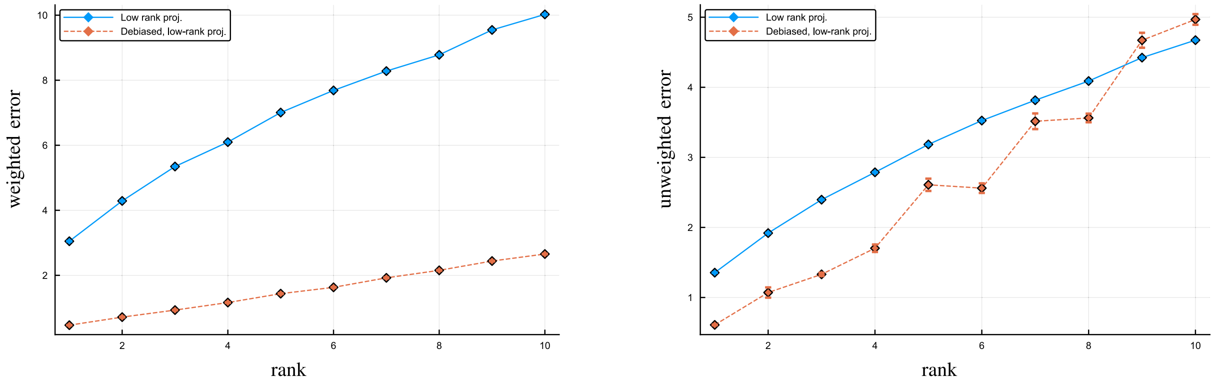

Fig. 2. Average errors of our debiased algorithm, versus standard low-rank projection procedure on Movielens sampling pattern and standard Gaussian data. Error bars denote standard deviations over 30 samples of the noise; each line (and error bar) is averaged over 30 trials.

the observed indices in each dataset. In particular, $(i, j) \in \Omega$ whenever user i rates item j. With $\mathbf{1}_\Omega = \left(\mathbf{1}_{(i,j)\in\Omega}\right)_{\substack{1\le i\le d_1,\\ 1\le j\le d_2}}$, in the following experiments we take $\mathbf{W}$ to be the best rank-1 approximation to $\mathbf{1}_\Omega$, $\mathbf{W} := \operatorname{argmin}_{\operatorname{rank}(U)=1}\|\mathbf{1}_\Omega - U\|$. In the examples we present, $\mathbf{W}_{ij} \ge 0$ for all i, j.

To produce Figures 1 and 2 below, we generate synthetic low-rank matrices, and use the sampling patterns from the real-life data described above. We consider ranks r between 1 and 10. For $N = 50$ trials, we construct random rank r matrices $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{d_1\times d_2}$, with independent standard normal entries. For each matrix $\mathbf{X}_i$, we average over $T = 25$ tests, testing our algorithm versus the the standard projection algorithm (*e.g.*, truncated SVD) on $\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,T}$, where $\mathbf{Y}_{i,j} = \mathbf{1}_\Omega \circ (\mathbf{X}_i + \mathbf{Z}_{i,j})$, where $\mathbf{Z}_{i,j} \in \mathbb{R}^{d_1\times d_2}$ has independent standard normal entries. We measure error using the weighted Frobenius norm. In particular, with $\hat{\mathbf{X}}$ an estimate of $\mathbf{X}$, we report

$$\frac{\|\mathbf{W}^{(1/2)}\circ(\mathbf{X}-\hat{\mathbf{X}})\|_F}{\|\mathbf{W}^{(1/2)}\|_F} \quad\text{and}\quad \frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_F}{\sqrt{d_1 d_2}}.$$

In the sequel, we refer to these as the *weighted error* and *unweighted error*, respectively.

We remark that although we only provide guarantees on the performance in the weighted Frobenius norm, our procedures exhibit good empirical performance (relative to standard projection procedures) even in the usual Frobenius norm.

B. Synthetic Sampling Patterns

For both of the experiments below, we focus on various synthetic constructions of Ω in order to demonstrate how the performance of debiased projection depends on the parameters of the sampling pattern.

For d and r specified below, we always construct random data as follows. For $N = 15$, we pick matrices $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{d\times d}$ with for $n = 1, \dots, N$,

$$\mathbf{X}_n = \mathbf{U}_n\mathbf{V}_n^T, \quad \mathbf{U}_n, \mathbf{V}_n \in \mathbb{R}^{d\times r},$$
$$(\mathbf{U}_n)_{ij}, (\mathbf{V}_n)_{ij} \overset{iid}{\sim} \mathcal{N}(0,1).$$

We describe two experiments below. The first one is motivated by the first case study, when $\Omega \sim \mathbf{W}$, and the second is motivated by the second case study when the spectral gap is large.

1) Sampling $\Omega \sim \mathbf{W}$: In the following experiment, we simulate our first case study, sampling $\Omega \sim \mathbf{W}$ for a rank-1 matrix $\mathbf{W}$. For simplicity, we take the weight matrix $\mathbf{W} = \mathbf{w}\mathbf{w}^T$ to be symmetric. We also take $d = 1000$ and $r = 10$.

We choose several different $\mathbf{W}$'s with different levels of "flatness," to show how the performance of our algorithm depends on the flatness of $\mathbf{W}$. More precisely, let

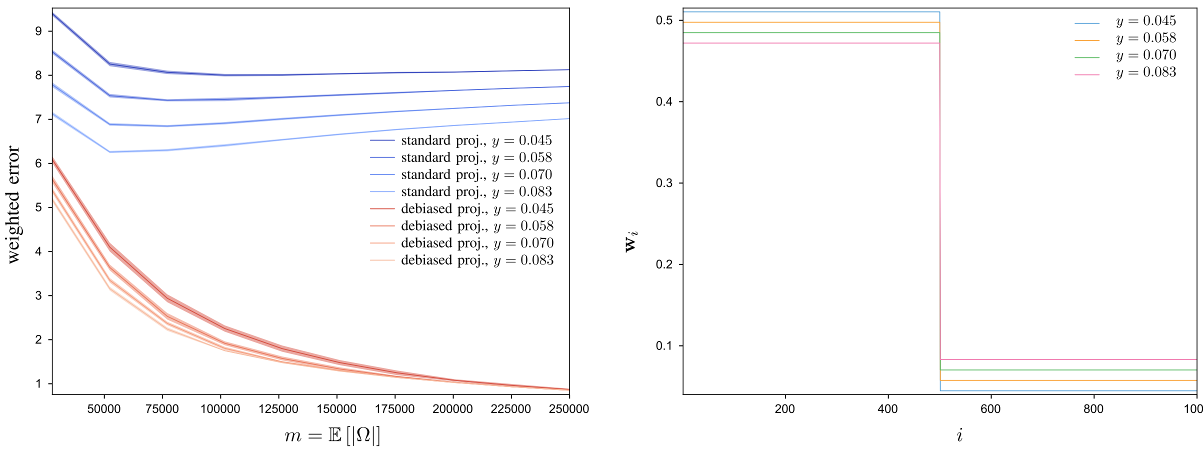

Fig. 3. (Left) Weighted error for various $\Omega \sim \mathbf{W}$ as described above. Shading indicates a single standard deviation, across the draws $\Omega_t \sim \mathbf{W}$. (Right) Entries of $\mathbf{w}$ for various choices of y and $m = 77,045$.

$m \in [4\ d\log d, d^2/4]$ and $y \in [\sqrt{2/d}, \sqrt{\log d/d}]$. For each m and y, we construct sampling patterns, $\Omega_1, \ldots, \Omega_T$, with $T = 15$ in the following manner. We select $\mathbf{W}$ with $\mathbf{w}$ given by

$$\mathbf{w} = \left(f(y,m,d)\mathbf{1}_{d/2}, y\mathbf{1}_{d/2}\right),$$
$$\text{where} \quad f(y,m,d) = \frac{2\sqrt{m}}{d} - y. \tag{20}$$

Above, $f(y,m,d)$ is chosen such that $\|\mathbf{w}\|_1 = \sqrt{m}$, and hence $\mathbb{E}[|\Omega|] = m$, when $\Omega \sim \mathbf{W}$. Thus, a larger value of y corresponds to a "flatter" matrix $\mathbf{W}$.

With such choices of $\mathbf{w}$, $\mathbf{W}_{ij} \in (0,1]$, and we draw $\Omega_t \sim \mathbf{W}$ for $t = 1, \ldots T$. For each t, we run the standard truncated SVD algorithm and our debiased projection procedure on $Y_{n,t} = \mathbf{1}_{\Omega_t} \circ (\mathbf{X}_n + \mathbf{Z}_{n,t})$, for $n = 1, \ldots, N$ and $t = 1, \ldots, T$. Here, $(\mathbf{Z}_{n,t})_{ij} \sim \mathcal{N}(0,1)$ for $1 \leq i,j \leq d$. We repeat this experiment for various m and y in the intervals above. The range of y is chosen to ensure that the plots of weighted error above are over the same range of m, while still respecting the constraints on $\mathbf{W}$ (*i.e.* that it have nonnegative entries in $[1/\sqrt{d}, 1]$).

Figure 3 indicates that under the experimental conditions given above, the SVD procedure has worse performance as $\mathbf{w}$ becomes flatter, while as expected our debiased procedure has improved performance as $\mathbf{w}$ becomes flatter. The debiased projection algorithm out-performed the standard projection in all cases.

2) Dependence on Spectral Gap: The following experiment demonstrates how the performance of the debiased projection algorithm depends on the spectral gap of $\mathbf{1}_\Omega$.

To construct sampling patterns of various spectral gaps, we consider graph products on $k = 50$ vertices. For regularities $\rho \in [2, 20]$, we construct two graphs $G_\rho, \widetilde{G}_\rho$ on k vertices that are ρ-regular.

The graph G_ρ is constructed such that each vertex $v \in \{0, \ldots, k-1\}$ is adjacent to the vertices $v' \equiv v + t \mod k$, $t = 0, 1, \ldots, \rho$. Notice that G_ρ is the same graph as is considered in Example 1, and the spectral gap is quite small. The second graph will be a random ρ-regular graph, so the spectral graph is with high probability large. More precisely, for $T = 15$ and trials $t = 1, \ldots, T$, the graph $\widetilde{G}_\rho^{(t)}$ is constructed as a random ρ-regular graph [10].

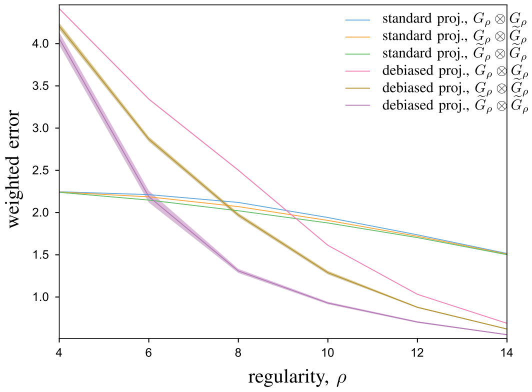

Fig. 4. Weighted error for sampling patterns taken from the adjacency matrices of graph products of low- and high-spectral gap, regular graphs. Shading indicates a single standard deviation over the draws $\widetilde{G}_\rho$.

In order to consider graphs with a range of λ, we consider three distributions on graphs: $G_\rho \otimes G_\rho$ (which has a small spectral gap), $G_\rho \otimes \widetilde{G}_\rho^{(t)}$ (which has an intermediate spectral gap) and $\widetilde{G}_\rho^{(t)} \otimes \widetilde{G}_\rho^{(t)}$ (which has a large spectral gap).

For $t = 1, \ldots, T$, we let Ω_t be the sampling pattern induced by each of these three graphs, and we draw observations $\mathbf{Y}_{n,t} = \mathbf{1}_{\Omega_t} \circ (\mathbf{X}_{n,t} + \mathbf{Z}_{n,t})$ as in the previous set of experiments. As before, we carry out both our debiased projection and truncated SVD on the $\mathbf{Y}_t$. In these experiments, we take the rank of the data to be $r = 10$. The results are shown in Figure 4. Figure 4 shows that, as expected, the debiased projection algorithm performs better when the spectral gap is smaller. This is also true of standard projection, although the effect is less pronounced. As ρ increases (that is, as Ω becomes denser), the debiased projection algorithm out-performs the standard projection algorithm.

C. Comparison With Max-Norm Projection

In this section, we continue our empirical evaluation of our matrix completion procedures, by comparing debiased

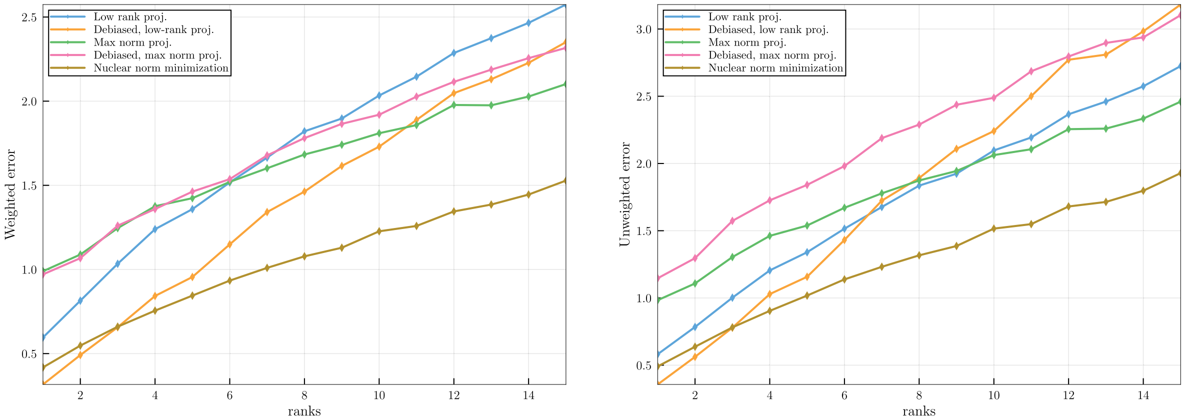

Fig. 5. Average errors of debiased and standard SVD and max-norm projection procedures and nuclear norm minimization on Jester sampling pattern and standard Gaussian data.

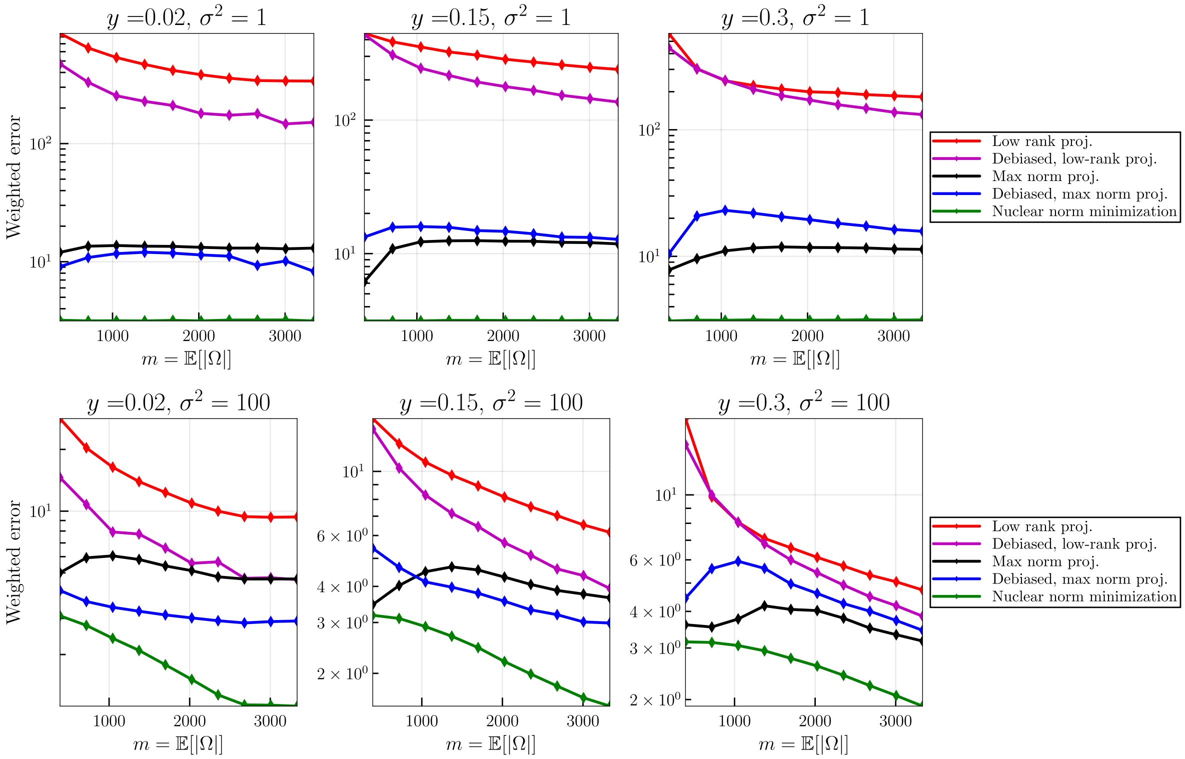

Fig. 6. Average errors of debiased and standard SVD and max-norm projection procedures and nuclear norm minimization when $\Omega \sim \mathbf{W}$ for spiky sampling patterns. Larger values of y denote less spiky sampling patterns. The top row indicates when $\mathbf{Z}$ has $N(0,1)$ entries, the bottom row indicates when $\mathbf{Z}$ has $N(0,100)$ entries.

SVD (18), standard SVD (19), as well as our debiased max-norm projection procedure (21) and a max-norm projection without debiasing (23), given below.

$$\hat{\mathbf{M}}_{\text{debiased}-\text{MNP}} = \mathbf{W}^{(-1/2)} \circ \text{argmin}_{\|\mathbf{X}\|_{\max} \le B}$$
$$\left\| \mathbf{X} - \mathbf{W}^{(-1/2)} \circ (\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \right\|, \tag{21}$$
$$\hat{\mathbf{M}}_{\text{standard}-\text{MNP}} = \text{argmin}_{\|\mathbf{X}\|_{\max} \le B} \tag{22}$$
$$\left\| \mathbf{X} - p^{-1}(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) \right\|. \tag{23}$$

Above, $p := |\Omega|/d_1 d_2$. Just as we use the true rank of $\mathbf{X}$ for our SVD experiments, we use the true max-norm of $\mathbf{X}$ as our choice of B for our implementations of algorithms (21) and (23). Finally, we also implement for comparison a nuclear norm minimization procedure:

$$\hat{\mathbf{M}}_{\text{NNM}} = \text{argmin}_{\|(\mathbf{M}_\Omega + \mathbf{Z}_\Omega) - \mathbf{X}_\Omega\|_F \le \delta} \|\mathbf{X}\|_*.$$

Above, $\|\cdot\|_*$ denotes the nuclear norm. Following [13], we take $\delta = \sigma\sqrt{m + \sqrt{8m}}$, when $\mathbf{Z}$ has iid $N(0, \sigma^2)$ entries.

1) Jester Sampling Pattern: We repeat our experiment as described in section IX-A, with the following modifications. First, we set $d = d_1 = d_2 = 100$, and instead of using the entire sampling pattern, we sample d columns and d rows (uniformly at random) to construct a random $d \times d$-submatrix of the original Jester sampling pattern for these experiments. We do this to keep the amount of computation manageable, as the max-norm projection and nuclear norm minimization require solving a semidefinite program, whereas the SVD-based procedures as used in section IX-A only require a single SVD computation. We present our experimental results in Figure 5.

Consistent with our previous experimental results presented in section IX-A, we see in Figure 5 that our debiased SVD procedure tends to outperform standard SVD in the weighted error metric and when the rank is low enough, in the unweighted error metric. When the rank is quite low, our procedure even outperforms the significantly more computationally expensive estimate given by nuclear norm minimization. We remark that nuclear norm minimization is a fundamentally different approach than our projection procedures. Specifically, nuclear norm minimization has a parameter δ used to bound the fidelity of the approximation on Ω. Moreover, it does not explicitly constrain rank. On the other hand, the SVD approaches and max-norm projection are constrained to matrices of low rank or and approximately low rank.

2) Spiky Patterns: Sampling $\Omega \sim \mathbf{W}$: We repeat our experiment as described in section IX-B.1, with the following modifications. First, we set $d = d_1 = d_2 = 100$, and instead of using the entire sampling pattern, we sample a random $d \times d$-submatrix of the original Jester sampling pattern for these experiments. We do this to keep the amount of computation manageable, as the max-norm projection and nuclear norm minimization require solving an semidefinite program, whereas the SVD-based procedures as used in section IX-A only require a single SVD computation. Secondly, we select a slightly broader range of spikiness parameters (denoted by y as in (20)). Finally, we also compare to multiple noise levels. Specifically we run the experiment with noise matrices $\mathbf{Z}$ with iid $N(0, 1)$ and $N(0, 100)$ entries. We present our experimental results in Figure 6.

Figure 6 shows that our debiasing procedures generally lead to lower error. Specifically, debiased SVD always outperforms standard SVD in the weighted error metric, and debiased max-norm projection outperforms its non-debiased counterpart when the sampling patterns are very spiky or the noise level is high. Finally, we see that nuclear norm minimization performs well, though is more competitive with our projection based procedures in the high-noise setting. As mentioned previously, nuclear norm minimization is a fundamentally different approach than the other rank-based projection algorithms we present, in addition to being more computationally expensive than the SVD-based approaches presented above.

Acknowledgment

The authors would like to thank the AIM SQuaRE program, where this work was begun. They would also like to thank Rachel Ward for helpful conversations and for pointing out the potential application in Corollary 30.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "Low-rank matrix factorization with attributes," 2006, *arXiv:cs/0611124*. [Online]. Available: https://arxiv.org/abs/cs/0611124

[2] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 17–24.

[3] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.

[4] M. Ashraphijuo, V. Aggarwal, and X. Wang, "On deterministic sampling patterns for robust low-rank matrix completion," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 343–347, Dec. 2018.

[5] M. Ashraphijuo and X. Wang, "Fundamental conditions for low-CP-rank tensor completion," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 2116–2145, 2017.

[6] M. Ashraphijuo, X. Wang, and V. Aggarwal, "Rank determination for low-rank data completion," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3422–3450, 2017.

[7] S. Bhojanapalli and P. Jain, "Universal matrix completion," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1881–1889.

[8] S. Bhojanapalli, P. Jain, and S. Sanghavi, "Tighter low-rank approximation via sampling the leveraged element," in *Proc. 26th Annu. ACM-SIAM Symp. Discrete Algorithms*, Oct. 2015, pp. 902–920.

[9] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Trans. Sensor Netw.*, vol. 2, no. 2, pp. 188–220, May 2006.

[10] B. Bollobás, *Random graphs, volume 73 of Cambridge Studies in Advanced Mathematics*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[11] T. T. Cai and W.-X. Zhou, "Matrix completion via max-norm constrained optimization," *Electron. J. Statist.*, vol. 10, no. 1, pp. 1493–1525, 2016.

[12] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, pp. 3619–3647, Dec. 2013.

[13] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[14] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[15] E. J. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[16] S. Chatterjee, "A deterministic theory of low rank matrix completion," 2019, *arXiv:1910.01079*. [Online]. Available: http://arxiv.org/abs/1910.01079

[17] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to SFM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1051–1063, Aug. 2004.

[18] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Completing any low-rank matrix, provably," *J. Mach. Learn. Res.*, vol. 16, pp. 2999–3034, Dec. 2015.

[19] A. K. Cline and I. S. Dhillon, "Computation of the singular value decomposition," in *Handbook of Linear Algebra*. Boca Raton, FL, USA: CRC Press, Jan. 2006, pp. 45-1–45-13.

[20] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, no. 3, pp. 189–223, Sep. 2014.

[21] A. Eftekhari, M. B. Wakin, and R. A. Ward, "MC^2: A two-phase algorithm for leveraged matrix completion," 2016, *arXiv:1609.01795*. [Online]. Available: http://arxiv.org/abs/1609.01795

[22] A. Eftekhari, D. Yang, and M. B. Wakin, "Weighted matrix completion and recovery with prior subspace information," 2016, *arXiv:1612.01720*. [Online]. Available: http://arxiv.org/abs/1612.01720

[23] S. Foucart, D. Needell, Y. Plan, and M. Wootters, "De-biasing low-rank projection for matrix completion," *Proc. SPIE Opt. Photon.*, vol. 10394, Aug. 2017, Art. no. 1039417.

[24] R. Foygel and N. Srebro, "Concentration-based guarantees for low-rank matrix reconstruction," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 315–340.

[25] S. Gaïffas and G. Lecué, "Sharp oracle inequalities for the prediction of a high-dimensional matrix," 2010, *arXiv:1008.4886*. [Online]. Available: http://arxiv.org/abs/1008.4886

[26] D. F. Gleich and L.-H. Lim, "Rank aggregation via nuclear norm minimization," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 60–68.

[27] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[28] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: A constant time collaborative filtering algorithm,” *Inf. Retr.*, vol. 4, no. 2, pp. 133–151, 2001.

[29] D. Gross, “Recovering low-rank matrices from few coefficients in any basis,” *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[30] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Trans. Interact. Intell. Syst.*, vol. 5, pp. 1–19, Dec. 2015.

[31] E. Heiman, G. Schechtman, and A. Shraibman, “Deterministic algorithms for matrix completion,” *Random Struct. Algorithms*, vol. 45, no. 2, pp. 306–317, Sep. 2014.

[32] G. J. O. Jameson, *Summing and Nuclear Norms in Banach Space Theory*, vol. 8. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[33] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[34] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from noisy entries,” *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Mar. 2010.

[35] F. J. Király, L. Theran, and R. Tomioka, “The algebraic combinatorial approach for low-rank matrix completion,” *Mach. Learn.*, vol. 16, pp. 1391–1436, Jan. 2015.

[36] O. Klopp and S. Gaiffas, “High dimensional matrix estimation with unknown variance of the noise,” 2011, *arXiv:1112.3055*. [Online]. Available: http://arxiv.org/abs/1112.3055

[37] O. Klopp, “Rank penalized estimators for high-dimensional matrices,” *Electron. J. Statist.*, vol. 5, no. 0, pp. 1161–1183, 2011.

[38] O. Klopp, “Noisy low-rank matrix completion with general sampling distribution,” 2012, *arXiv:1203.0108*. [Online]. Available: http://arxiv.org/abs/1203.0108

[39] O. Klopp, “Noisy low-rank matrix completion with general sampling distribution,” *Bernoulli*, vol. 20, no. 1, pp. 282–303, Feb. 2014.

[40] V. Koltchinskii, “Von neumann entropy penalization and low-rank matrix estimation,” *Ann. Statist.*, vol. 39, no. 6, pp. 2936–2973, Dec. 2011.

[41] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion,” *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, Oct. 2011.

[42] D. Jason Lee, B. Recht, R. Salakhutdinov, N. Srebro, and A. Joel Tropp, “Practical large-scale optimization for max-norm regularization,” in *Proc. Adv. Neural Inf. Process. Syst* Vancouver, BC, USA: Curran Associates, Dec. 2010, pp. 1297–1305.

[43] T. Lee and A. Shraibman, “Matrix completion from any given set of observations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1781–1787.

[44] Y. Li, Y. Liang, and A. Risteski, “Recovery guarantee of weighted low-rank approximation via alternating minimization,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2358–2367.

[45] G. Liu, Q. Liu, X.-T. Yuan, and M. Wang, “Matrix completion with deterministic sampling: Theories and methods,” 2018, *arXiv:1805.02313*. [Online]. Available: http://arxiv.org/abs/1805.02313

[46] Z. Liu and L. Vandenberghe, “Interior-point method for nuclear norm approximation with application to system identification,” *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, Jan. 2010.

[47] R. Meka, P. Jain, and I. S. Dhillon, “Matrix completion from power-law distributed samples,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1258–1266.

[48] S. Negahban and M. J. Wainwright, “Restricted strong convexity and weighted matrix completion: Optimal bounds with noise,” *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1665–1697, 2012.

[49] D. L. Pimentel-Alarcon, N. Boston, and R. D. Nowak, “A characterization of deterministic sampling patterns for low-rank matrix completion,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 623–636, Jun. 2016.

[50] D. L. Pimentel-Alarcon and R. D. Nowak, “A converse to low-rank matrix completion,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 96–100.

[51] Y. Plan, R. Vershynin, and E. Yudovina, “High-dimensional estimation with geometric constraints,” *Inf. Inference*, vol. 6, no. 1, pp. 1–40, 2017.

[52] C. Rashtchian, “Bounded matrix rigidity and john’s theorem,” *Electron. Colloq. Comput. Complex.*, vol. 23, p. 93, Dec. 2016.

[53] B. Recht, “A simpler approach to matrix completion,” *J. Mach. Learn. Res.*, vol. 12, pp. 3413–3430, Jan. 2011.

[54] A. Rohde and A. B. Tsybakov, “Estimation of high-dimensional low-rank matrices,” *Ann. Statist.*, vol. 39, no. 2, pp. 887–930, Apr. 2011.

[55] M. Rudelson and R. Vershynin, “Hanson-wright inequality and sub-Gaussian concentration,” *Electron. Commun. Probab.*, vol. 18, pp. 1–9, Dec. 2013.

[56] A. Shapiro, Y. Xie, and R. Zhang, “Matrix completion with deterministic pattern—A geometric perspective,” 2018, *arXiv:1802.00047*. [Online]. Available: http://arxiv.org/abs/1802.00047

[57] A. Singer, “A remark on global positioning from local distances,” *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 28, pp. 9507–9511, Jul. 2008.

[58] A. Singer and M. Cucuringu, “Uniqueness of low-rank matrix completion by rigidity theory,” *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 4, pp. 1621–1641, Jan. 2010.

[59] N. Srebro, N. Alon, and T. S. Jaakkola, “Generalization error bounds for collaborative prediction with low-rank matrices,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1321–1328.

[60] N. Srebro, J. D. Rennie, and T. Jaakkola, “Maximum-margin matrix factorization,” in *Proc. Adv. Neural Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2004, pp. 1–5.

[61] N. Srebro and R. R. Salakhutdinov, “Collaborative filtering in a non-uniform world: Learning with the weighted trace norm,” in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2056–2064, 2010.

[62] N. Srebro and A. Shraibman, “Rank, trace-norm and max-norm,” in *Learning Theory*, P. Auer and R. Meir, Eds. Berlin, Germany: Springer, 2005, pp. 545–560.

[63] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: A factorization method,” *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[64] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, Aug. 2012.

[65] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” 2010, *arXiv:1011.3027*. [Online]. Available: https://arxiv.org/abs/1011.3027

Simon Foucart received the M.Eng. degree from the Ecole Centrale Paris and the master’s degree in mathematics from the University of Cambridge in 2001, and the Ph.D. degree in mathematics from the University of Cambridge, in 2006, specializing in approximation theory. After two post-doctoral positions at Vanderbilt University and Université Paris 6, he joined Drexel University in 2010, before moving to the University of Georgia in 2013. Since 2015, he has been with Texas A&M University, where he became a Professor and a Presidential Impact Fellow in 2019. His research was recognized by the *Journal of Complexity*, from which he received the 2010 Best Paper Award. His recent work focuses on the field of compressive sensing, whose theory is exposed in the book *A Mathematical Introduction to Compressive Sensing* he coauthored with Holger Rauhut. His current research interests include the mathematical aspects of metagenomics, optimization, deep learning, and data science at large.

Deanna Needell (Member, IEEE) received the Ph.D. degree from UC Davis, before working as a Post-Doctoral Fellow at Stanford University. She is currently a Full Professor of mathematics with UCLA. She received several awards, including the IEEE Best Young Author Award, the Hottest paper in Applied and Computational Harmonic Analysis Award, the Alfred P. Sloan Fellowship, the NSF CAREER and NSF BIGDATA Award, and the Prestigious IMA Prize in Applied Mathematics. She has been a Research Professor Fellow with several top research institutes, including the Mathematical Sciences Research Institute and Simons Institute in Berkeley. She also serves as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, *Linear Algebra and its Applications*, the *SIAM Journal on Imaging Sciences*, and *Transactions of Mathematics and its Applications* as well as on the organizing committee for SIAM sessions and the Association for Women in Mathematics.

Reese Pathak received the B.S. degree in computer science and mathematics from Stanford University in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Sciences (EECS), UC Berkeley, advised by M. Wainwright and M. I. Jordan. His research interests include algorithm design for high-dimensional statistics and continuous optimization.

Yaniv Plan received the Ph.D. degree in applied and computational mathematics program with Caltech. He was a Hildebrand Assistant Professor, and also an NSF Post-doctoral Researcher with the mathematics Department, University of Michigan. He is currently an Assistant Professor of mathematics with The University of British Columbia. He was awarded a Tier 2 Canada Research Chair for this position. In 2016, he won the UBC Mathematics and the Pacific Institute for the Mathematical Sciences Faculty Award. He received the W. P. Carey and Co. Inc. Prize for an outstanding Ph.D. dissertation. He also spent two years as a Visiting Researcher at Stanford University during a Sabbatical from his Ph.D. His research interests include applied probability, high-dimensional inference, random matrix theory, compressive sensing, matrix completion, and learning theory.

Mary Wootters (Member, IEEE) received the B.A. degree in mathematics and computer science from the Swarthmore College in 2008, and the Ph.D. degree in mathematics from the University of Michigan in 2014. She was an NSF Post-Doctoral Fellow with Carnegie Mellon University from 2014 to 2016. She is currently an Assistant Professor of computer science and electrical engineering with Stanford University. She also works in theoretical computer science, applied mathematics, and information theory; her research interests include error correcting codes and randomized algorithms for dealing with high-dimensional data. She was a recipient of the NSF CAREER Award and was named a Sloan Research Fellow, in 2019.