De-biasing low-rank projection for matrix completion

Simon Foucart[*], Deanna Needell[†], Yaniv Plan[‡], Mary Wootters[§]

July 28, 2017

ABSTRACT

We study matrix completion with non-uniform, deterministic sampling patterns. We introduce a computable parameter, which is a function of the sampling pattern, and show that if this parameter is small, then we may recover missing entries of the matrix, with appropriate weights. We theoretically analyze a simple and well-known recovery method, which simply projects the (zero-padded) subsampled matrix onto the set of low-rank matrices. We show that under non-uniform deterministic sampling, this method yields a biased solution, and we propose an algorithm to de-bias it. Numerical simulations demonstrate that de-biasing significantly improves the estimate. However, when the observations are noisy, the error of this method can be sub-optimal when the sampling is highly non-uniform. To remedy this, we suggest an alternative which is based on projection onto the max-norm ball whose robustness to noise tolerates arbitrarily non-uniform sampling. Finally, we analyze convex optimization in this framework.

1. INTRODUCTION

In the matrix completion problem, one is given a subset of the entries of a low-rank matrix and the goal is to fill in the missing entries. There are broad applications including collaborative filtering,[1] system identification,[2] sensor localization,[3–5] rank aggregation,[6] scene recovery in imaging,[7,8] multi-class learning,[9–11] and many others. There are several proposed programs and algorithms which can take advantage of the low-rank structure to complete the matrix. These include low-rank projection,[12,13] and convex optimization.[14,15] Fortunately, there is a strong theoretical backing for these methods.[12,13,15–26] However, much of the theory of matrix completion requires uniformly random sampling; or at least random (possibly non-uniform) sampling.[27–32] Thus, in practical applications in which one is given a sampling pattern with no model of a random generating process, it is unclear the extent to which these theoretical results apply.

In this work, we examine the problem of matrix completion with deterministic sampling. To our knowledge, there are not many works which give theoretical guarantees for matrix completion with deterministic sampling pattern.[33–36] The work of Heiman, Schechtman, and Shraibman,[34] Bhojanapalli and Jain[35] and Li, Liang, and Risteski[36] all 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. The work of Heiman et al. also discusses non-uniform sampling patterns, and they show that if a matrix completion algorithm works well on a non-uniform random sampling pattern, then there exists a deterministic sampling pattern on which it also does well. However, that work does not characterize which deterministic, non-uniform sampling patterns are good. The work of Lee and Shraibman[33] addresses this by introducing a parameter which measures the complexity of the sampling pattern. Their parameter is given by solving a semidefinite program involving the sampling pattern, and that work shows that if this parameter is small, then the entries of the matrix can be recovered with appropriate weights. That work is quite general, and bounds the reconstruction error under general algorithms of the form "find a matrix which is correct on the observed entries, with minimal $\|\cdot\|$ norm," for any norm $\|\cdot\|$. In particular, the results do not work with noisy observations and do not apply to projection-based methods.

[*] Simon Foucart, Department of Mathematics, Texas A&M University, College Station, TX, USA, E-mail: foucart@tamu.edu.
[†] Deanna Needell, Department of Mathematics, Univ. of California, Los Angeles, CA, USA, E-mail: deanna@math.ucla.edu.
[‡] Yaniv Plan, Department of Mathematics, Univ. of British Columbia, Vancouver, BC, Canada, Email: yaniv@math.ubc.edu.
[§] Mary Wootters, Department of Computer Science and Electrical Engineering, Stanford University, Stanford, CA, USA, Email: marykw@stanford.edu.

In this work, we continue in the same vein as Lee and Shraibman, by identifying a parameter of the sampling pattern that guarantees good weighted recovery. Our parameter is in some sense a special case of Lee and Shraibman, although because our choices are more specific, we are able to say more. In particular, our parameter is easy to compute (without solving an SDP), and is a generalization of the spectral gap approach in the uniform case. We show that if our parameter is small, then the missing entries of the matrix can be estimated well, even in the presence of noise, and we analyze several efficient algorithms for doing this.

More precisely, we show that projection-based methods for matrix completion are biased under such non-uniform sampling, and we give a method to de-bias them, resulting in an algorithm to reconstruct a matrix given deterministic non-uniform sampling patterns. Additionally, we propose and analyze another method based on convex optimization. We conduct numerical simulations for a projection-based method using real-data sampling patterns. These show that debiasing is vital for accurate matrix recovery.

1.1 Model

We wish to estimate a low-rank matrix $M \in \boldsymbol{R}^{d_1 \times d_2}$ from noisy entries. As is standard in matrix completion, we make a few assumptions about M. First, we assume that M is low-rank, and set $r = \operatorname{rank}(M)$. Second, we need to assume that M is not too "spiky." We set $\gamma := \|M\|_\infty$ to be the largest entry of M in absolute value.

Let $\Omega \subset \{1, 2, \ldots, d_1\} \times \{1, 2, \ldots, d_2\}$ be the sampling pattern, i.e., the set of pairs of natural numbers indexing the observed matrix entries. Let $m := |\Omega|$. Let $\sigma > 0$ and let $Z \in \boldsymbol{R}^{d_1 \times d_2}$ be a noise matrix with independent $N(0, \sigma^2)$ entries. In this work, we suppose that we observe m noisy observations of the form

$$Y_{i,j} = M_{i,j} + \sigma Z_{i,j}, \qquad (i,j) \in \Omega. \tag{1}$$

Let $P_\Omega : \boldsymbol{R}^{d_1 \times d_2} \to \boldsymbol{R}^{d_1 \times d_2}$ be the projection onto the sampling pattern, i.e.,

$$P_\Omega(X)_{i,j} = \begin{cases} X_{i,j} & \text{if} \quad (i,j) \in \Omega, \\ 0 & \text{if} \quad (i,j) \notin \Omega. \end{cases}$$

Given a matrix $X \in \boldsymbol{R}^{d_1 \times d_2}$, we define

$$X_\Omega := P_\Omega(X).$$

Then the observation model may be succinctly written as

$$Y_\Omega = M_\Omega + Z_\Omega.$$

We use $\mathbf{1} \in \boldsymbol{R}^{d_1 \times d_2}$ to denote the matrix with all entries equal to 1 and $\circ$ to denote the Hadamard (entry-wise) product. Thus, $P_\Omega(X) = \mathbf{1}_\Omega \circ X$.

1.2 Computable parameter

In the literature on matrix completion, and the closely related fields of *low-rank matrix recovery*[37,38] and *compressive sensing*,[39–41] parameters which determine that the problem set up is *well conditioned* are generally NP-hard to compute. For example, the *restricted isometry constant* of compressive sensing is NP-hard to verify,[42] and similarly for restricted isometry constants defined in the context of low-rank matrix recovery, since the latter reduces to the former in the diagonal case.[37] Despite much research into computable parameters,[43–46] the guarantees that can be made using parameters that are known to be computable in polynomial time is generally far from optimal.[43]

Fortunately, the particular observation model of matrix completion has a special structure which allows a computable parameter, as discussed above.[33–36] In those previous works (except for that of Lee and Shraibman[33] which we discussed above), the matrix $\mathbf{1}_\Omega$ is thought of as the adjacency matrix of a d-regular graph. If the *spectral gap* of that graph is large—that is, if the second-largest eigenvalue is much smaller than the largest—then the matrix can be efficiently estimated with near-optimal guarantees. These results require that the top singular vectors of $\mathbf{1}_\Omega$ be the all-ones vectors; in particular, the best rank-1 approximation of $\mathbf{1}_\Omega$ must be equal to the all-ones matrix (after re-scaling). In this work, we extend this approach to settings where this may not be the case.

To motivate our approach, consider two extreme cases. The first is the case discussed above. In this case, the best rank-1 approximate to $\mathbf{1}_\Omega$ is proportional to $\mathbf{1}$, and we can hope to recover all of the entries of M. In the second case, suppose that $\mathbf{1}_\Omega$ contains only a single row, which is all 1, and the rest is zero; thus, $\mathbf{1}_\Omega$ is itself rank 1, and is its own best rank-1 approximation. In this second case, it is clear that we can recover $M \circ \mathbf{1}_\Omega$, and nothing else. This motivated the following weighted approach: suppose that W is the best rank-1 approximation to $\mathbf{1}_\Omega$; we will attempt to recover $W \circ M$.

To be precise, let W be any rank-1 matrix, so that every entry of W strictly greater than 0, but no other assumption. (Our theory works whether or not W is the *best* rank-1 approximation to $\mathbf{1}_\Omega$, although we encourage the reader to think of it that way). Let

$$\lambda := \|\mathbf{1}_\Omega - W\|.$$

If W does happen to be the best rank-1 approximation to $\mathbf{1}_\Omega$, and if $\mathbf{1}_\Omega$ is irreducible, then the Perron-Frobenius theorem implies that all entries of W are greater than 0.[47,48] In this case, λ is the second singular value of $\mathbf{1}_\Omega$. Below, we show that if λ is small, then a low-rank matrix can be well-approximated just from viewing entries on Ω.

We also show that the structure of W is important. Indeed, when W is not flat, the projection method of Keshevan, Montanari and Oh[12] gives a biased estimate; we show how to debias the estimate. Further, the error bounds are weighted proportionally to W. This reflects the fact that W is larger in rows/columns that are sampled more highly, and thus one expects to have more accurate estimation on these rows and columns.

2. MAIN RESULTS

2.1 Low-rank projection

In this section, we estimate M using low-rank projection as suggested by Keshavan et al.[12] Thus, let

$$\hat{M}_0 := \underset{\mathrm{rank}(X) \le r}{\arg\min} \|X - Y_\Omega\|_F. \tag{2}$$

In words, we find the closest rank-r matrix to the zero-padded matrix Y_Ω. This can be accomplished simply (and quickly) via a truncated singular value decomposition. Previous works[12,13] have considered matrix completion with uniform at random sampling—made even more uniform by trimming any rows or columns that are sampled significantly more than the expected value. In this setting, those works suggest to rescale the entire matrix $\hat{M}_0$ to give an estimate of M, and they show that the estimate is quite accurate. However, in the non-uniform, deterministic case, it is unclear whether such a result should persist. In fact, in the non-uniform case, we find that the estimate $\hat{M}$ is biased, and different parts of the matrix need to be rescaled by different weights. Thus set

$$\hat{M}_{debias} := W^{(-1)} \circ \hat{M}_0. \tag{3}$$

Above and below, for a matrix X and real number t, we define $X^{(t)}$ to be entry-wise exponentiation, i.e.,

$$X^{(t)}_{i,j} := X^t_{i,j}.$$

We will need one extra piece of notation. For a matrix, A, let $\|A\|_{2,\infty}$ by the largest Euclidean norm of a row. Then set

$$\nu = \nu(\Omega) := \max(\|\mathbf{1}_\Omega\|_{2,\infty}, \|\mathbf{1}_\Omega^T\|_{2,\infty}).$$

Theorem 2.1 (Low-rank projection). *Let M, Y_Ω follow the model in Section 1.1 and let $d = d_1 + d_2$. Then, with probability at least $1 - 1/d$,*

$$\left\| \hat{M}_0 - W \circ M \right\|_F \leq 2\sqrt{2}\, r\lambda\gamma + 4\sqrt{2}\sqrt{r}\nu\sqrt{\log d}\sigma$$

or, equivalently,

$$\left\| W \circ (\hat{M}_{debias} - M) \right\|_F \leq 2\sqrt{2}\, r\lambda\gamma + 4\sqrt{2}\sqrt{r}\nu\sqrt{\log d}\sigma.$$

Remark 1 (Interpreting the error bound). *To help understand the error bound, let us compare to the case of uniform random sampling, in which each entry of the matrix is observed with probability p (independent of other entries). Then the expectation of the sampling pattern is $\mathbb{E}\mathbf{1}_\Omega = p\mathbf{1}$, a natural* rank-1 *estimate of $\mathbf{1}_\Omega$. Further, it follows from Seginer's theorem*[49] *(see also [50, proof of Lemma 1]) that with high probability $\lambda \leq O(\sqrt{pd})$ if $p \geq \log(d)/d$. In this setting, the parameter η is simple to bound by, e.g., Hoeffding's inequality, giving $\eta \leq O(\sqrt{pd})$ with high probability. Also note the relationship between p and $\mathbb{E}m$, i.e., $\mathbb{E}m = pd_1d_2$.*

Then we may consider normalized parameters

$$W' := \frac{d_1 d_2 W}{m}, \qquad \lambda' = \sqrt{\frac{d_1 d_2}{md}}\lambda, \qquad \eta' = \sqrt{\frac{d_1 d_2}{md}}\eta$$

and note that in the case of uniform sampling, the operator which takes the Hadamard product with W' (i.e. $W'\circ$) acts as the identity, and λ', η' are $O(1)$ with high probability. Then, one may write the error bound of Theorem 2.1 as

$$\frac{\left\| W' \circ (\hat{M}_{debias} - M) \right\|_F}{\sqrt{d_1 d_2}} \leq 2\sqrt{2}\lambda'\gamma\sqrt{\frac{dr^2}{m}} + 4\sqrt{2}\nu'\sigma\sqrt{\frac{dr\log(d)}{m}},$$

or, ignoring constants,

$$\frac{\left\| \hat{M}_{debias} - M \right\|_F^2}{d_1 d_2} = \frac{\left\| W' \circ (\hat{M}_{debias} - M) \right\|_F^2}{d_1 d_2} \leq O\left((\lambda')^2\gamma^2 \frac{dr^2}{m} + (\nu')^2\sigma^2\frac{dr\log d}{m} \right).$$

In other words, when $m \gtrsim dr\max(r, \log(d))$, the right hand side becomes small and the estimate is more accurate.

When the sampling is not uniform, one may still consider the normalized parameters; in this case the Hadamard product with W' puts higher weight on the rows/columns that are more prevalently observed.

Remark 2 (Noise bound). *Note that the noise term in the error bound is proportional ν. This is maximized if a single row or column of Ω is entirely sampled, which is paradoxical, since more sampling should not make the matrix completion problem harder. In fact, similar observations have been made before,*[12] *in which the authors suggested to "trim" rows or columns that had too many entries. Further, the error bound is tight which can be seen by taking $r = 1$ and $M = 0$ (or $M \approx 0$), in which case $\hat{M}_0$ is the* rank-*1 projection of the noise, and the factor of ν is unavoidable. While we found good noise resilience in our simulations, the above arguments suggest that it is worth exploring other recovery methods. In the next section, we propose a different projection-based algorithm whose noise bound tolerates non-uniform sampling. In the following section, we give a convex-optimization method whose noise bound also tolerates non-uniform sampling.*

2.2 Max-norm ball projection

The max-norm ball has proven to be an effective constraint set to promote low-rankness for the matrix completion problem.[14] We begin this section by describing some of its interesting and useful properties of the max norm.

Given a matrix X, the max-norm is defined by

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

Let $B_{\max}$ be the max-norm ball, that is

$$B_{\max} := \{X \in \boldsymbol{R}^{d_1\times d_2} : \|X\|_{\max} \leq 1\}.$$

A result due to Grothendiek shows that $B_{\max}$ closely contained in a polytope whose vertices are flat rank-1 matrices. Let $\mathcal{F}$ be the set of flat, rank-1, matrices, i.e.,

$$\mathcal{F} := \{uv^T : u \in \{+1,-1\}^{d_1},\, v \in \{+1,-1\}^{d_2}\}.$$

Then Grothendieck's inequality (see [51, Chapter 10]) states that $B_{\max}$ is nearly the convex hull of $\mathcal{F}$.

Theorem 2.2 (Grothendieck's Inequality).

$$conv(\mathcal{F}) \subset B_{\max} \subset K_G \cdot conv(\mathcal{F})$$

where $K_G \leq 1.783$ is Grothendiek's constant.

We also need the dual to the max norm. To be precise, given two real matrices A and B of the same dimensions, we use the standard inner product $\langle A, B\rangle := \sum_{i,j} A_{i,j}B_{i,j}$. We denote $\|\cdot\|_{\max^*}$ the dual norm to the max norm, i.e., for a matrix X

$$\|X\|_{\max^*} := \max_{A\in B_{\max}} \langle A, X\rangle.$$

Grothendiek's inequality implies that for a matrix X

$$\|X\|_{\max^*} \leq K_G \max_{B\in\mathcal{F}} \langle B, X\rangle \tag{4}$$

Remark 3 (Max norm and rank). *Given a* rank-r *matrix M with $\|M\|_\infty \leq \gamma$, one has [52, Corollary 2.2]*

$$\|M\|_{\max} \leq \sqrt{r}\gamma.$$

Thus, one may think of the max norm (squared) as a proxy for the rank of a flat matrix. In contrast to the rank, the max norm is robust to small perturbations. In this section we do not assume that M has small rank, only that (a weighted version of) it has small max norm.

We take our (biased) estimate of M to be the projection of Y_Ω onto the max-norm ball, where the norm used to define distance in the projection is the dual to the max norm.[¶]

$$\hat{M}_0 := \operatorname*{arg\,min}_{\|X\|_{\max}\leq\alpha} \|X - Y_\Omega\|_{\max^*}.$$

As in the previous section, it is vital to debias the estimate: $\hat{M}_{debias} = W^{(-1)} \circ \hat{M}_0$.

The recovery guarantees of this section will use a variation on the computable parameter of the last section. Thus let

$$\tilde{\lambda} := \left\|\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega\right\|_{\max^*}. \tag{5}$$

We note that this parameter is very similar to the one considered in the work of Lee and Shraibman.[33]

Remark 4 (Comparison of λ and $\tilde{\lambda}$). *We note that it is less clear how to choose W to minimize $\tilde{\lambda}$ in comparison to λ. However, $\tilde{\lambda}$ can be bounded proportionally to λ. Indeed, Grothendieck's inequality implies that*

$$\tilde{\lambda} \leq K_G\sqrt{d_1 d_2}\left\|\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega\right\|$$

[¶] We thank R. Vershynin for sharing the idea of projecting using the dual norm; this approach was also used by Lee and Shraibman.[33]

which can be further bounded by $K_G\sqrt{d_1 d_2}\left\|W^{(-1)}\right\|_\infty \lambda$. *Thus, by taking* W *equal to the* rank-*1 projection of* $\mathbf{1}_\Omega$ *one can bound* $\tilde{\lambda}$ *proportionally to the second singular value of* $\mathbf{1}_\Omega$.

We are now in position to state our main theorem for max-norm ball projection.

Theorem 2.3 (Max-norm ball projection). *Let* M, Y_Ω *follow the model in Section 1.1 and let* $d = d_1 + d_2$. *Assume that* $\|W \circ M\|_{\max} \le \alpha$. *Then with probability at least* $1 - d$,

$$\left\|\hat{M}_0 - W \circ M\right\|_F^2 \le 7.2\tilde{\lambda}\alpha^2 + 10\alpha\sigma\sqrt{dm}.$$

or, equivalently,

$$\left\|W \circ (\hat{M}_{debias} - M)\right\|_F^2 \le 7.2\tilde{\lambda}\alpha^2 + 10\alpha\sigma\sqrt{dm}.$$

2.3 Convex optimization with max norm

Above we considered projection based methods for matrix completion which are appealing due to their simplicity (in particular the method of Section 2.1). Nevertheless, since only the entries on Ω are observed, it is natural to minimize misfit constrained to those entries. We consider such a method in this section. As in the previous section, we assume a bound on the max norm of M.

$$\hat{M} := \underset{\|X\|_{\max} \le \alpha}{\arg\min} \; \|X_\Omega - Y_\Omega\|_{\max^*}.$$

The solution does not need to be debiased, and the error bound is weighted slightly differently.

Theorem 2.4 (Max-norm constrained convex optimization). *Let* M, Y_Ω *follow the model in Section 1.1 and let* $d = d_1 + d_2$. *Assume that* $\|W \circ M\|_{\max} \le \alpha$. *Then with probability at least* $1 - d$,

$$\left\|W^{(1/2)} \circ (\hat{M} - M)\right\|_F^2 \le 7.2\alpha^2\lambda\sqrt{d_1 d_2} + 10\alpha\sigma\sqrt{dm}.$$

Remark 5 (Tightening λ for non-uniform sampling pattern). *In fact,* λ *could be replaced with a strictly smaller, but more exotic parameter. As seen from Equation* (12) *below, one could replace* λ *with*

$$\frac{\|\mathbf{1}_\Omega - W\|_{\max^*}}{\sqrt{d_1 d_2}},$$

which could potentially result in an improved error guarantee, depending on the sampling pattern.

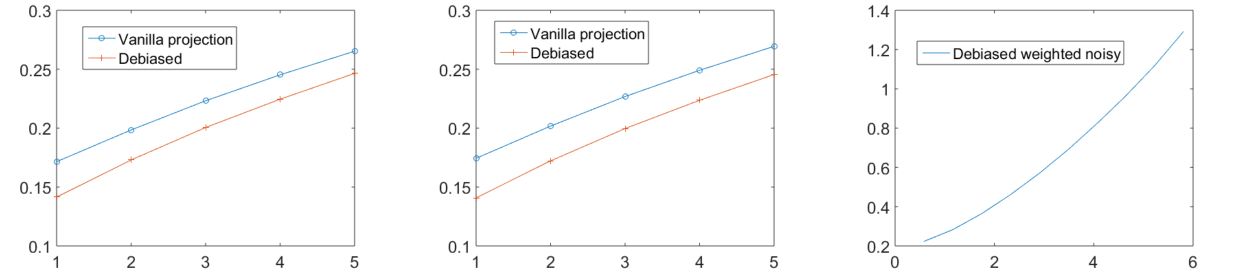
3. NUMERICAL RESULTS

In this section we illustrate the results of several numerical experiments for the method of Section 2.1. We consider both a uniform at random sampling pattern and also a non-uniform sampling pattern given by real data.

In our first experiments, for various values of r, we create a random 10,000 by 10,000 rank-r matrix.[‖] We (add noise and) subsample 2.1% of the entries. We subsample at this rate so that it will match a real data set described below. We consider estimates $\hat{M}_0$ and $\hat{M}_{debias}$ as in (2) and (3), respectively, which are computed with a truncated singular value decomposition given the correct value of r.[**] We

[‖] We construct a rank-r matrix by taking the product of two 10,000 by r factors with standard normal entries.

[**] In practice this would need to be estimated via cross validation.

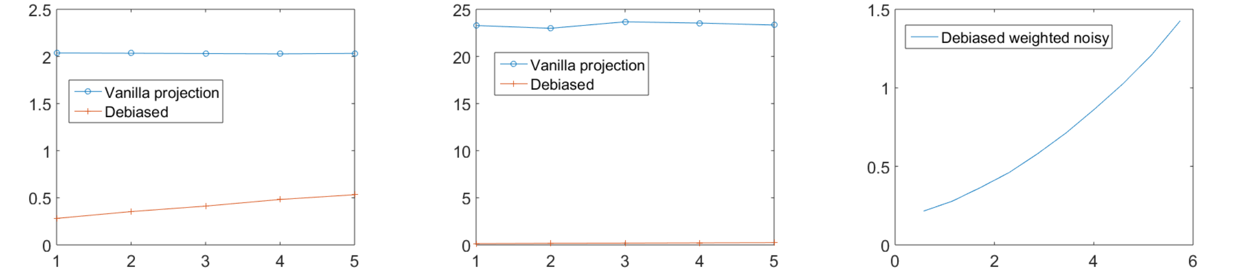


(a) Vanilla and debiased relative error, no noise, plotted against rank

(b) Vanilla and debiased weighted relative error, no noise, plotted against rank

(c) Debiased relative weighted error plotted against noise to signal ratio. Rank is 3.

Figure 1: 10,000 by 10,000 matrix, 2.1% of entries uniformly randomly subsampled.



(a) Vanilla and debiased relative error, no noise, plotted against rank

(b) Vanilla and debiased weighted relative error, no noise, plotted against rank

(c) Debiased relative weighted error plotted against noise to signal ratio. Rank is 3.

Figure 2: 10,000 by 10,000 matrix, 2.1% of entries sampled according to the taste music dataset.

give names *vanilla method* associated with $\hat{M}_0$ and *debiased method* associated with $\hat{M}_{debias}$. For an estimate $\hat{M}$, we consider both the unweighted relative error: $\left\|\hat{M} - M\right\|_F / \|M\|_F$ and the weighted relative error $\left\|W \circ (\hat{M} - M)\right\|_F / \|W \circ M\|_F$. We average each error over 20 experiments.

In Figure 1 we take the subsampling to be uniform at random, as is often considered in matrix completion, but is not the focus of this paper. One might not expect debiasing to have much effect under uniform sampling, but surprisingly, even in this setting we find that the debiased estimate $\hat{M}_{debias}$ performs significantly better than $\hat{M}_0$. However, we note that the weighted relative error and the unweighted relative error are nearly the same. We also plot the relative weighted error of the debiased method against the noise-to-signal ratio $\|Z_\Omega\|_F / \|M_\Omega\|_F$. Notably the relative error is still well below 1 even with 2.1% sampling and noise-to-signal ratio much higher than 1.

The setup of Figure 2, is the same as Figure 1 aside from one key difference: The sampling is taken from real data generated by music listening history (songs vs users) — the *taste music data set*.[53] We restrict to the 10,000 most prolific song listeners and 10,000 songs which are most listened to. The corresponding sampling pattern is sampled at 2.1%, albeit quite non-uniformly. In this case, one sees that debiasing is vital. The vanilla method has unweighted relative error much larger than 1, whereas the debiased method has unweighted relative error much less than 1. When we consider the weighted error, the vanilla method deteriorates further and the debiased method improves. As can be seen by comparison of the noisy plots, the debiased method, measured with weighted relative error, performs roughly the same with the non-uniform real data as it does with uniform at random data.

4. PROOFS

4.1 Proof of Theorem 2.1.

Set $H := \hat{M}_0 - W \circ M$. The following lemma controls the operator norm of H, denoted $\|H\|$.

LEMMA 4.1. *With probability at least* $1 - 1/d$,

$$\|H\| \leq 2\lambda \|M\|_{\max} + 4\nu\sqrt{\log d}\sigma. \tag{6}$$

The theorem then follows quickly from this lemma. First, [52, Corollary 2.2] implies that $\|M\|_{\max} \leq \sqrt{r}\gamma$. Further,

$$\|H\|_F \leq \sqrt{\operatorname{rank}(H)}\,\|H\|\,.$$

Observe that $\operatorname{rank}(H) \leq \operatorname{rank}(W \circ M) + \operatorname{rank}(\hat{M}) \leq 2r$. Combine these inequalities with the operator norm bound of Lemma 4.1 to prove the theorem.

Proof. [Proof of Lemma 4.1] By the triangle inequality

$$\|H\| = \left\|\hat{M}_0 - Y + Y - W \circ M\right\| \leq \left\|\hat{M}_0 - Y\right\| + \|Y - W \circ M\|\,.$$

By definition, $\hat{M}_0$ is a closest rank $\leq r$ matrix to Y in Frobenius norm, and equivalently, in operator norm. Since $\operatorname{rank} W \circ M \leq r$, we have $\left\|\hat{M}_0 - Y\right\| \leq \|Y - W \circ M\|$ and thus

$$\|H\| \leq 2\|Y - W \circ M\| = 2\|\mathbf{1}_\Omega \circ Z + (\mathbf{1}_\Omega - W) \circ M\| \leq 2\|\mathbf{1}_\Omega \circ Z\| + 2\|(\mathbf{1}_\Omega - W) \circ M\| \tag{7}$$

where the last inequality follows from the triangle inequality.

The following lemma controls $\|\mathbf{1}_\Omega \circ Z\|$.

LEMMA 4.2. *With probability at least* $1 - 1/d$,

$$\|\mathbf{1}_\Omega \circ \boldsymbol{Z}\| \leq 2\nu\sqrt{\log d}\sigma. \tag{8}$$

It remains to bound $\|(\mathbf{1}_\Omega - W) \circ M\|$. By [54, Corollary], we have

$$\|(\mathbf{1}_\Omega - W) \circ M\| \leq \|\mathbf{1}_\Omega - W\| \cdot \|M\|_{\max}\,.$$

Plug this and the result of Equation (8) into Equation (7) to complete the proof. □

Proof. [Proof of Lemma 4.2]

We will use[55][Theorem 4.1.1] to control the norm of this random matrix. We first specialize the above result to our setting.

COROLLARY 4.3 (OPERATOR NORM OF $\boldsymbol{Z} \circ \boldsymbol{B}$). *Let* $\boldsymbol{Z} \in \boldsymbol{R}^{d_1 \times d_2}$ *have independent standard normal entries. Let* $\boldsymbol{B} \in \boldsymbol{R}^{d_1 \times d_2}$ *be a fixed matrix. Set*

$$\nu(\boldsymbol{B}) := \max(\|B\|_{2,\infty}, \|B^T\|_{2,\infty}).$$

Then, for all $t \geq 0$,

$$\Pr(\|\boldsymbol{B} \circ \boldsymbol{Z}\| \geq t) \leq (d_1 + d_2)\exp\left(\frac{-t^2}{2\nu^2(\boldsymbol{B})}\right).$$

The lemma is then proven by applying this corollary with $\boldsymbol{B} = \mathbf{1}_\Omega$. □

4.2 Proof of Theorem 2.3

The proof is similar to the proof of Theorem 2.1. The main difference is in the (much better) control of the noise term. As above, set $H := \hat{M}_0 - W \circ M$. By duality, we have

$$\|H\|_F^2 \leq \|H\|_{\max} \cdot \|H\|_{\max^*}.$$

Observe that $\|H\|_{\max} \leq \left\|\hat{M}_0\right\|_{\max} + \|W \circ M\|_{\max} \leq 2\alpha$ be assumption on $W \circ M$ and definition of $\hat{M}_0$. The theorem is then proven by bounding $\|H\|_{\max^*}$ as given in the following lemma.

Lemma 4.4. *With probability at least* $1 - 1/d$

$$\|H\|_{\max^*} \leq 2K_G\tilde{\lambda}\alpha + 5\sqrt{dm}\sigma \leq 3.6\tilde{\lambda}\alpha + 5\sqrt{dm}\sigma.$$

Proof. [Proof of Lemma 4.4] We have

$$\|H\|_{\max^*} = \left\|\hat{M}_0 - Y_\Omega + Y_\Omega - W \circ M\right\|_{\max^*} \leq \left\|\hat{M}_0 - Y_\Omega\right\|_{\max^*} + \|W \circ M - Y_\Omega\|_{\max^*} \leq 2\,\|\boldsymbol{W} \circ M - Y_\Omega\|_{\max^*}$$

since $\hat{M}$ is the minimizer of the right-hand side. Further, by definition of Y,

$$\|\boldsymbol{W} \circ M - Y_\Omega\|_{\max^*} \leq \|W \circ M - M_\Omega\|_{\max^*} + \|Z_\Omega\|_{\max^*} = \left\|(\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega) \circ W \circ M\right\|_{\max^*} + \|Z_\Omega\|_{\max^*}. \tag{9}$$

The following lemma bounds the noise term.

Lemma 4.5. *With probability at least* $1 - 1/d$

$$\|Z_\Omega\|_{\max^*} \leq 2.5\sqrt{dm}\sigma.$$

It remains to bound $\left\|(\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega) \circ W \circ M\right\|_{\max^*}$. By Grothendiek's inequality 2.2 we have

$$\begin{aligned}
\left\|(\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega) \circ W \circ M\right\|_{\max^*} &\leq K_G \max_{B \in \mathcal{F}} \langle (\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega) \circ W \circ M, B\rangle \\
&= K_G \max_{B \in \mathcal{F}} \langle \mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega, W^T \circ M^T \circ B\rangle \\
&\leq K_G \left\|\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega\right\|_{\max^*} \cdot \max_{B \in \mathcal{F}} \|W \circ M \circ B\|_{\max} \\
&= K_G \left\|\mathbf{1} - W^{(-1)} \circ \mathbf{1}_\Omega\right\|_{\max^*} \cdot \|W \circ M\|_{\max} \\
&\leq K_G\tilde{\lambda}\alpha.
\end{aligned} \tag{10}$$

The second to last line follows since the max norm is invariant with respect to Hadamard product with any matrix in $\mathcal{F}$ and the last line follows by definition. Insert the last inequality, and also the noise bound of Lemma 4.5 into Equation (9), then insert the result into the equation above, to complete the proof. □

Proof. [Proof of Lemma 4.5] Grothendieck's inequality 2.2 implies that

$$\|Z_\Omega\|_{\max^*} \leq K_G \cdot \max_{X \in \mathcal{F}} \langle Z_\Omega, X\rangle.$$

Note that $|\mathcal{F}| = 2^{d-1}$ and thus the right hand side is the maximum of 2^{d-1} $N(0, m\sigma^2)$ random variables. We complete the proof with a standard tail bound. Indeed, it is well known that

$$\Pr(N(0,1) > t) \leq \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \qquad t > 0.$$

Thus, by the union bound, for any $t > 0$,

$$\max_{X \in \mathcal{F}} \langle Z_\Omega, X \rangle \leq t\sqrt{m}\sigma, \qquad \text{with probability at least } 1 - \frac{2^{d-1}}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Pick $t = 1.4\sqrt{d}$, and note that 1) for this value of t the probability estimate is bounded by $1 - 1/d$; 2) $K_G \cdot 1.4 \leq 2.5$ to complete the proof. □

4.3 Proof of Theorem 2.4

Our proof is similar to the previous two. Set $H := \hat{M} - M$. By duality,

$$\left\| W^{(1/2)} \circ H \right\|_F^2 = \langle H, W \circ H \rangle \leq \|H\|_{\max} \|W \circ H\|_{\max^*}.$$

We have $\|H\|_{\max} \leq \|M\|_{\max} + \left\|\hat{M}\right\|_{\max} \leq 2\alpha$. The theorem is then proven by bounding $\|W \circ H\|_{\max^*}$ as given in the following lemma.

Lemma 4.6. *With probability at least* $1 - 1/d$

$$\|W \circ H\|_{\max^*} \leq 3.6\alpha\sqrt{d_1 d_2}\lambda + 5\sqrt{dm}\sigma.$$

Proof. [Proof of Lemma 4.6] By triangle inequality

$$\|W \circ H\|_{\max^*} \leq \|(\mathbf{1}_\Omega - W) \circ H\|_{\max^*} + \|H_\Omega\|_{\max^*} = I + II. \tag{11}$$

Using the same steps that were used to derive Equation (10), we have

$$I \leq 2K_G\alpha \|\mathbf{1}_\Omega - W\|_{\max^*} \leq 2K_G\alpha\sqrt{d_1 d_2}\lambda \leq 3.6\alpha\sqrt{d_1 d_2}\lambda. \tag{12}$$

The second inequality follows by definition of $\|\cdot\|_{\max^*}$ and since every element of $\mathcal{F}$ has rank 1 and Frobenius norm equal to $\sqrt{d_1 d_2}$.

We now control II. Add and subtract Y_Ω inside the norm, and use triangle inequality to give

$$II \leq \left\|\hat{M}_\Omega - Y\right\|_{\max^*} + \|M_\Omega - Y\|_{\max^*} \leq 2\|M_\Omega - Y\|_{\max^*} = 2\|Z_\Omega\|_{\max^*},$$

where the second inequality follows since $\hat{M}$ is the minimizer of the misfit. Finally, from Lemma 4.5, we have

$$\|Z_\Omega\|_{\max^*} \leq 2.5\sqrt{dm}\sigma, \qquad \text{with probability at least} 1 - \frac{1}{d}.$$

Combine the previous two equations to bound II and insert this bound, together with the bound on I from Equation (12), into Equation (11) to complete the proof. □

Acknowledgements

The authors would like to thank Anna Ma for her help retrieving and processing the data used to generate the sampling patterns in our experiments. We would like to thank Roman Vershynin for sharing his idea of projecting using the dual norm. They would also like to thank the AIM SQuaRE program that funded and hosted our initial collaboration. In addition, Needell is partially supported by an Alfred P. Sloan Fellowship and NSF CAREER grant DMS-1348721; Plan by NSERC grant 22R23068; Foucart by the NSF grant DMS-1622134.

REFERENCES

[1] Goldberg, D., Nichols, D., Oki, B., and Terry, D., "Using collaborative filtering to weave an information tapestry," *Comm. ACM* **35**(12), 61–70 (1992).

[2] Liu, Z. and Vandenberghe, L., "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Analysis and Applications* **31**(3), 1235–1256 (2009).

[3] Biswas, P., Lian, T. C., Wang, T. C., and Ye, Y., "Semidefinite programming based algorithms for sensor network localization," *ACM Trans. Sen. Netw.* **2**(2), 188–220 (2006).

[4] Singer, A., "A remark on global positioning from local distances," *Proc. Natl. Acad. Sci.* **105**(28), 9507–9511 (2008).

[5] Singer, A. and Cucuringu, M., "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM J. Matrix Analysis and Applications* **31**(4), 1621–1641 (2010).

[6] Gleich, D. and Lim, L.-H., "Rank aggregation via nuclear norm minimization," in [*Proc. ACM SIGKDD Int. Conf. on Knowledge, Discovery, and Data Mining (KDD)*], (Aug. 2011).

[7] Chen, P. and Suter, D., "Recovering the missing components in a large noisy low-rank matrix: Application to sfm," *IEEE transactions on pattern analysis and machine intelligence* **26**(8), 1051–1063 (2004).

[8] Tomasi, C. and Kanade, T., "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision* **9**(2), 137–154 (1992).

[9] Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P., "Low-rank matrix factorization with attributes," *arXiv preprint cs/0611124* (2006).

[10] Amit, Y., Fink, M., Srebro, N., and Ullman, S., "Uncovering shared structures in multiclass classification," in [*Proceedings of the 24th international conference on Machine learning*], 17–24, ACM (2007).

[11] Argyriou, A., Evgeniou, T., and Pontil, M., "Multi-task feature learning," in [*Advances in neural information processing systems*], 41–48 (2007).

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

[13] Keshavan, R., Montanari, A., and Oh, S., "Matrix completion from noisy entries," *J. Machine Learning Research* **11**, 2057–2078 (2010).

[14] Srebro, N., Rennie, J. D., and Jaakkola, T., "Maximum-margin matrix factorization.," in [*Proc. Adv. in Neural Processing Systems (NIPS)*], (Dec. 2004).

[15] Candès, E. and Plan, Y., "Matrix completion with noise," *Proc. IEEE* **98**(6), 925–936 (2010).

[16] Gross, D., "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory* **57**(3), 1548–1566 (2011).

[17] Candès, E. and Recht, B., "Exact matrix completion via convex optimization," *Found. Comput. Math.* **9**(6), 717–772 (2009).

[18] Candès, E. and Tao, T., "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory* **56**(5), 2053–2080 (2010).

[19] Negahban, S. and Wainwright, M., "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Machine Learning Research* **13**, 1665–1697 (2012).

[20] Koltchinskii, V., Lounici, K., and Tsybakov, A., "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Stat.* **39**(5), 2302–2329 (2011).

[21] Recht, B., "A simpler approach to matrix completion," *J. Machine Learning Research* **12**, 3413–3430 (2011).

[22] Rohde, A. and Tsybakov, A., "Estimation of high-dimensional low-rank matrices," *Ann. Stat.* **39**(2), 887–930 (2011).

[23] Klopp, O., "Rank penalized estimators for high-dimensional matrices," *Elec. J. Stat.* **5**, 1161–1183 (2011).

[24] Gaïffas, S. and Lecué, G., "Sharp oracle inequalities for the prediction of a high-dimensional matrix," (2010). Arxiv preprint 1008.4886.

[25] Klopp, O., "High dimensional matrix estimation with unknown variance of the noise," (2011). Arxiv preprint 1112.3055.

[26] Koltchinskii, V., "Von Neumann entropy penalization and low-rank matrix estimation," *Ann. Stat.* **39**(6), 2936–2973 (2012).

[27] Meka, R., Jain, P., and Dhillon, I. S., "Matrix completion from power-law distributed samples," in [*Advances in neural information processing systems*], 1258–1266 (2009).

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

[29] Klopp, O. et al., "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli* **20**(1), 282–303 (2014).

[30] Bhojanapalli, S., Jain, P., and Sanghavi, S., "Tighter low-rank approximation via sampling the leveraged element," in [*Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*], 902–920, SIAM (2014).

[31] Eftekhari, A., Wakin, M. B., and Ward, R. A., "Mcˆ 2: A two-phase algorithm for leveraged matrix completion," *arXiv preprint arXiv:1609.01795* (2016).

[32] Eftekhari, A., Yang, D., and Wakin, M. B., "Weighted matrix completion and recovery with prior subspace information," *arXiv preprint arXiv:1612.01720* (2016).

[33] Lee, T. and Shraibman, A., "Matrix completion from any given set of observations," in [*Advances in Neural Information Processing Systems*], 1781–1787 (2013).

[34] Heiman, E., Schechtman, G., and Shraibman, A., "Deterministic algorithms for matrix completion," *Random Structures & Algorithms* **45**(2), 306–317 (2014).

[35] Bhojanapalli, S. and Jain, P., "Universal matrix completion," in [*International Conference on Machine Learning*], 1881–1889 (2014).

[36] Li, Y., Liang, Y., and Risteski, A., "Recovery guarantee of weighted low-rank approximation via alternating minimization," in [*International Conference on Machine Learning*], 2358–2367 (2016).

[37] Recht, B., Fazel, M., and Parrilo, P. A., "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review* **52**(3), 471–501 (2010).

[38] Candès, E. and Plan, Y., "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inform. Theory* **57**(4), 2342–2359 (2011).

[39] "Compressive sensing webpage," (2016). http://dsp.rice.edu/cs.

[40] Eldar, Y. C. and Kutyniok, G., [*Compressed sensing: theory and applications*], Cambridge University Press (2012).

[41] Foucart, S. and Rauhut, H., [*A mathematical introduction to compressive sensing*], Birkhäuser (2013).

[42] Bandeira, A., Dobriban, E., Mixon, D., and Sawin, W., "Certifying the restricted isometry property is hard," (2012). Available at `http://arxiv.org/abs/1204.1580`.

[43] d'Aspremont, A. and El Ghaoui, L., "Testing the nullspace property using semidefinite programming," *Mathematical programming* **127**(1), 123–144 (2011).

[44] Lee, K. and Bresler, Y., "Computing performance guarantees for compressed sensing," in [*Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*], 5129–5132, IEEE (2008).

[45] Juditsky, A. and Nemirovski, A., "On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization," *Mathematical programming* **127**(1), 57–88 (2011).

[46] d'Aspremont, A., Bach, F., and Ghaoui, L., "Optimal solutions for sparse principal component analysis," *The Journal of Machine Learning Research* **9**, 1269–1294 (2008).

[47] Perron, O., "Zur theorie der matrices," *Mathematische Annalen* **64**(2), 248–263 (1907).

[48] Frobenius, F. G., "Ueber matrizen aus nicht negativen elementen," 456–477 (1912).

[49] Seginer, Y., "The expected norm of random matrices," *Combinatorics, Probability, and Computing* **9**(2), 149–166 (2000).

[50] Davenport, M., Plan, Y., van den Berg, E., and Wootters, M., "One-bit matrix completion," Submitted. Available at `http://arxiv.org/abs/1209.3672`.

[51] Jameson, G. J. O., [*Summing and nuclear norms in Banach space theory*], vol. 8, Cambridge University Press (1987).

[52] Rashtchian, C., "Bounded matrix rigidity and john's theorem.," in [*Electronic Colloquium on Computational Complexity (ECCC)*], **23**, 93 (2016).

[53] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P., "The million song dataset," in [*Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*], (2011).

[54] Ando, T., Horn, R. A., and Johnson, C. R., "The singular values of a hadamard product: a basic inequality," *Linear and Multilinear Algebra* **21**(4), 345–365 (1987).

[55] Tropp, J. A., "An introduction to matrix concentration inequalities," *arXiv preprint arXiv:1501.01571* (2015).