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 Jain35 and Li, Liang, and Risteski36 all relate the sampling pattern Ω to a graph whose adjacency matrix is given by Ω1Ω. 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 Shraibman33 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 ∥·∥ norm,” for any norm ∥·∥. 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 \( M \in \mathbb{R}^{d_1 \times d_2} \) 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 \( r = \text{rank}(M) \). 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 \( Z \in \mathbb{R}^{d_1 \times d_2} \) 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

\[
Y_{i,j} = M_{i,j} + \sigma Z_{i,j}, \quad (i, j) \in \Omega.
\]

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

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

Given a matrix \( X \in \mathbb{R}^{d_1 \times d_2} \), we define

\[
X_{\Omega} := P_{\Omega}(X).
\]

Then the observation model may be succinctly written as

\[
Y_{\Omega} = M_{\Omega} + Z_{\Omega}.
\]

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

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\(^{43,46}\). In those previous works (except for that of Lee and Shraibman\(^{43}\) which we discussed above), the matrix \( 1_{\Omega} \) 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 $1_{\Omega}$ be the all-ones vectors; in particular, the best rank-1 approximation of $1_{\Omega}$ 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 $1_{\Omega}$ is proportional to $1_{\Omega}$, and we can hope to recover all of the entries of $M$. In the second case, suppose that $1_{\Omega}$ contains only a single row, which is all 1, and the rest is zero; thus, $1_{\Omega}$ is itself rank 1, and is its own best rank-1 approximation. In this second case, it is clear that we can recover $M \circ 1_{\Omega}$, and nothing else. This motivated the following weighted approach: suppose that $W$ is the best rank-1 approximation to $1_{\Omega}$; 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 $1_{\Omega}$, although we encourage the reader to think of it that way). Let

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

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

We also show that the structure of $W$ is important. Indeed, when $W$ is not flat, the projection method of Keshevan, Montanari and Oh 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. Thus, let

$$\hat{M}_0 := \arg\min_{\text{rank}(X) \leq r} \|X - Y_{\Omega}\|_F.$$ 

In words, we find the closest rank-$r$ matrix to the zero-padded matrix $Y_{\Omega}$. This can be accomplished simply (and quickly) via a truncated singular value decomposition. Previous work 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 $\hat{M}_0$ is biased, and different parts of the matrix need to be rescaled by different weights. Thus set

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

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

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

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

$$\nu = \nu(\Omega) := \max(\|1_{\Omega}\|_{2,\infty}, \|1_{\Omega}^T\|_{2,\infty}).$$
**Theorem 2.1 (Low-rank projection).** Let \( M, Y_\Omega \) follow the model in Section 1.1 and let \( d = d_1 + d_2 \). Then, with probability at least \( 1 - \frac{1}{d} \),

\[
\|M_0 - W \circ M\|_F \leq 2\sqrt{2}r\lambda \gamma + 4\sqrt{2}\sqrt{r\nu} \sqrt{\log d}\sigma
\]

or, equivalently,

\[
\|W \circ (\hat{M}_{\text{debias}} - M)\|_F \leq 2\sqrt{2}r\lambda \gamma + 4\sqrt{2}\sqrt{r\nu} \sqrt{\log d}\sigma.
\]

**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 \( E1_\Omega = p1 \), a natural rank-1 estimate of \( 1_\Omega \). Further, it follows from Seginer’s theorem \(^{12}\) (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 \( \eta \) 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 \( E_m \), i.e., \( E_m = pd_1d_2 \).

Then we may consider normalized parameters

\[
W' := \frac{d_1d_2W}{m}, \quad \lambda' = \sqrt{\frac{d_1d_2}{md}} \lambda, \quad \eta' = \sqrt{\frac{d_1d_2}{md}} \eta
\]

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 \( \lambda', \eta' \) are \( O(1) \) with high probability. Then, one may write the error bound of Theorem 2.1 as

\[
\frac{\|W' \circ (\hat{M}_{\text{debias}} - M)\|_F}{\sqrt{d_1d_2}} \leq 2\sqrt{2}\lambda' \sqrt{\frac{dr^2}{m}} + 4\sqrt{2}\nu' \sqrt{\frac{dr \log(d)}{m}}.
\]

or, ignoring constants,

\[
\left\|\hat{M}_{\text{debias}} - M\right\|_F^2 \leq \left\|W' \circ (\hat{M}_{\text{debias}} - M)\right\|_F^2 \leq O \left( (\lambda')^2 \frac{dr^2}{m} + (\nu')^2 \sigma^2 \frac{dr \log(d)}{m} \right).
\]

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 \( \nu \). This is maximized if a single row or column of \( \Omega \) 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 \( M_0 \) is the rank-1 projection of the noise, and the factor of \( \nu \) 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

$$
\|X\|_{\text{max}} := \min_{X = UV^T} \|U\|_{2,\infty}\|V\|_{2,\infty}.
$$

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

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

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

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

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

**Theorem 2.2 (Grothendieck’s Inequality).**

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

where $K_G \leq 1.783$ is Grothendieck’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 $\|\cdot\|_{\text{max}^*}$, the dual norm to the max norm, i.e., for a matrix $X$

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

Grothendieck’s inequality implies that for a matrix $X$

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

(4)

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

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

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 $Y_\Omega$ onto the max-norm ball, where the norm used to define distance in the projection is the dual to the max norm.$^*$

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

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

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

$$
\hat{\lambda} := \left\|1 - W(-1) \circ 1_\Omega \right\|_{\text{max}^*}.
$$

(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 $\lambda$ and $\hat{\lambda}$).** We note that it is less clear how to choose $W$ to minimize $\hat{\lambda}$ in comparison to $\lambda$. However, $\hat{\lambda}$ can be bounded proportionally to $\lambda$. Indeed, Grothendieck’s inequality implies that

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

$^*$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 $K_G \sqrt{d_1 d_2} \|W^{(-1)}\|_\infty \lambda$. 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 $M, Y_\Omega$ follow the model in Section 1.1 and let $d = d_1 + d_2$. Assume that $\|W \circ M\|_{\max} \leq \alpha$. Then with probability at least $1 - d$,

$$\|\hat{M}_0 - W \circ M\|_F^2 \leq 7.2\tilde{\lambda} \alpha^2 + 10\alpha \sigma \sqrt{d_1 d_2}.$$

or, equivalently,

$$\|W \circ (\hat{M}_{\text{debias}} - M)\|_F^2 \leq 7.2\tilde{\lambda} \alpha^2 + 10\alpha \sigma \sqrt{d_1 d_2}.$$

### 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 $\Omega$ 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 $M$.

$$\hat{M} := \arg\min_{\|X\|_{\max} \leq \alpha} \|X_\Omega - Y_\Omega\|_{\max}^*.$$

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 $M, Y_\Omega$ follow the model in Section 1.1 and let $d = d_1 + d_2$. Assume that $\|W \circ M\|_{\max} \leq \alpha$. Then with probability at least $1 - d$,

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

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

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

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 $\hat{M}_{\text{debias}}$ as in (2) and (3), respectively, which are computed with a truncated singular value decomposition given the correct value of $r$.

---

*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.
give names vanilla method associated with $\hat{M}_0$ and debiased method associated with $\hat{M}_{\text{debias}}$. For an estimate $\hat{M}$, we consider both the unweighted relative error: $\frac{\|\hat{M} - M\|_F}{\|M\|_F}$ and the weighted relative error $\frac{\|W \circ (\hat{M} - M)\|_F}{\|W \circ M\|_F}$. 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 $\hat{M}_{\text{debias}}$ 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 $\frac{\|Z\|_F}{\|M\|_F}$. 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. 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 $1 - 1/d$,

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

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

Observe that $\text{rank}(H) \leq \text{rank}(W \circ M) + \text{rank}(\hat{M}) \leq 2r$. 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

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

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

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

where the last inequality follows from the triangle inequality.

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

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

$$\|1_{\Omega} \circ Z\| \leq 2\nu \sqrt{\log d} \sigma.$$  \hfill (8)

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

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

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

**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 $Z \circ B$).** Let $Z \in \mathbb{R}^{d_1 \times d_2}$ have independent standard normal entries. Let $B \in \mathbb{R}^{d_1 \times d_2}$ be a fixed matrix. Set

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

Then, for all $t \geq 0$,

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

The lemma is then proven by applying this corollary with $B = 1_{\Omega}$. \(\square\)
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
\[ \|H\|_F^2 \leq \|H\|_{\text{max}} \cdot \|H\|_{\text{max}^*}. \]

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

**Lemma 4.4.** With probability at least $1 - 1/d$
\[ \|H\|_{\text{max}^*} \leq 2K_G \tilde{\lambda} \alpha + 5\sqrt{d}m\sigma \leq 3.6\tilde{\lambda} \alpha + 5\sqrt{d}m\sigma. \]

**Proof.** [Proof of Lemma 4.4] We have
\[ \|H\|_{\text{max}^*} = \left\| \hat{M}_0 - Y_{\Omega} + Y_{\Omega} - W \circ M \right\|_{\text{max}^*} \leq \left\| \hat{M}_0 - Y_{\Omega} \right\|_{\text{max}^*} + \|W \circ M - Y_{\Omega}\|_{\text{max}^*} \leq 2\left\|W \circ M - Y_{\Omega}\right\|_{\text{max}^*}, \]

since $\hat{M}$ is the minimizer of the right-hand side. Further, by definition of $Y$,\[ \|W \circ M - Y_{\Omega}\|_{\text{max}^*} \leq \|W \circ M - M_{\Omega}\|_{\text{max}^*} + \|Z_{\Omega}\|_{\text{max}^*} = \left\| (1 - W^{(-1)} \circ 1_{\Omega}) \circ W \circ M \right\|_{\text{max}^*} + \|Z_{\Omega}\|_{\text{max}^*}. \]

The following lemma bounds the noise term.

**Lemma 4.5.** With probability at least $1 - 1/d$
\[ \|Z_{\Omega}\|_{\text{max}^*} \leq 2.5\sqrt{d}m\sigma. \]

It remains to bound $\left\| (1 - W^{(-1)} \circ 1_{\Omega}) \circ W \circ M \right\|_{\text{max}^*}$. By Grothendieck’s inequality 2.2 we have
\[ \left\| (1 - W^{(-1)} \circ 1_{\Omega}) \circ W \circ M \right\|_{\text{max}^*} \leq K_G \max_{B \in \mathcal{F}} \left\| (1 - W^{(-1)} \circ 1_{\Omega}) \circ W \circ M, B \right\|_{\text{max}^*} = K_G \max_{B \in \mathcal{F}} \left\| (1 - W^{(-1)} \circ 1_{\Omega}, W^T \circ M^T \circ B \right\|_{\text{max}^*} \leq K_G \left\| 1 - W^{(-1)} \circ 1_{\Omega} \right\|_{\text{max}^*} \cdot \max_{B \in \mathcal{F}} \|W \circ M \circ B\|_{\text{max}^*} = K_G \left\| 1 - W^{(-1)} \circ 1_{\Omega} \right\|_{\text{max}^*} \cdot \|W \circ M\|_{\text{max}} \leq K_G \tilde{\lambda} \alpha. \]

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
\[ \|Z_{\Omega}\|_{\text{max}^*} \leq K_G \cdot \max_{X \in \mathcal{F}} \|Z_{\Omega}, X\|. \]

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
\[ \Pr(N(0,1) > t) \leq \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad t > 0. \]
Thus, by the union bound, for any $t > 0$,
\[
\max_{X \in \mathcal{F}} \langle Z_\Omega, X \rangle \leq t \sqrt{m} \sigma, \quad \text{with probability at least } 1 - \frac{2^{d-1}}{t \sqrt{2\pi}} e^{-t^2}.
\]
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. \(\square\)

### 4.3 Proof of Theorem 2.4

Our proof is similar to the previous two. Set $H := \hat{M} - M$. By duality,
\[
\|W^{(1/2)} \circ H\|_F^2 = \langle H, W \circ H \rangle \leq \|H\|_{\max} \|W \circ H\|_{\max^*}.
\]
We have $\|H\|_{\max} \leq \|M\|_{\max} + \|\hat{M}\|_{\max} \leq 2\alpha$. The theorem is then proven by bounding $\|W \circ H\|_{\max^*}$ as given in the following lemma.

**Lemma 4.6.** With probability at least $1 - 1/d$
\[
\|W \circ H\|_{\max^*} \leq 3.6\alpha \sqrt{d_1 d_2} + 5\sqrt{dm} \sigma.
\]

**Proof.** [Proof of Lemma 4.6] By triangle inequality
\[
\|W \circ H\|_{\max^*} \leq \|(I - W) \circ H\|_{\max^*} + \|H\|_{\max^*} = I + II.
\]
Using the same steps that were used to derive Equation (10), we have
\[
I \leq 2K_G \alpha \|1 - W\|_{\max^*} \leq 2K_G \alpha \sqrt{d_1 d_2} \lambda \leq 3.6\alpha \sqrt{d_1 d_2} \lambda.
\]
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 $Y_\Omega$ inside the norm, and use triangle inequality to give
\[
II \leq \|\hat{M} - \hat{Z}\|_{\max^*} + \|M - Y\|_{\max^*} \leq 2 \|M - Y\|_{\max^*} = 2 \|Z_\Omega\|_{\max^*},
\]
where the second inequality follows since $\hat{M}$ is the minimizer of the misfit. Finally, from Lemma 4.5 we have
\[
\|Z_\Omega\|_{\max^*} \leq 2.5\sqrt{dm} \sigma, \quad \text{with probability at least } 1 - \frac{1}{d}.
\]
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. \(\square\)

### 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


