Lower Bounds for Quantized Matrix Completion

Mary Wootters and Yaniv Plan
Department of Mathematics
University of Michigan
Ann Arbor, MI
Email: {wootters, yplan}@umich.edu

Mark A. Davenport
School of Elec. & Comp. Engineering
Georgia Institute of Technology
Atlanta, GA
Email: mdav@gatech.edu

Ewout van den Berg
Department of Statistics
Stanford University
Stanford, CA
Email: ewout@stanford.edu

Abstract—In this paper we consider the problem of 1-bit matrix completion, where instead of observing a subset of the real-valued entries of a matrix M, we obtain a small number of binary (1-bit) measurements generated according to a probability distribution determined by the real-valued entries of M. The central question we ask is whether or not it is possible to obtain an accurate estimate of M from this data. In general this would seem impossible, however, it has recently been shown in [1] that under certain assumptions it is possible to recover M by optimizing a simple convex program. In this paper we provide lower bounds showing that these estimates are near-optimal.

I. Introduction

The problem of recovering a matrix from an incomplete sampling of its entries—also known as *matrix completion*—arises in a wide variety of practical situations. In many of these settings, however, the observations are not only incomplete, but also highly *quantized*, often even to a single bit. In this paper we consider a statistical model for such data where instead of observing a real-valued entry as in the original matrix completion problem, we are now only able to see a positive or negative rating. This binary output is generated according to a probability distribution which is parameterized by the corresponding entry of the unknown low-rank matrix M. A natural question in this context is: "Given observations of this form, can we recover the underlying matrix?"

The first theoretical accuracy guarantees in this context were recently established by the current authors in [1]. Inspired by recent theoretical results on unquantized matrix completion, [1] establishes that $O(rd)$ binary observations are sufficient to accurately recover a $d \times d$, rank-r matrix by convex programming. In this paper we review the upper bounds of [1], and provide lower bounds showing that these upper bounds are nearly optimal. First, however, we provide a brief review of the existing results on matrix completion.

A. Matrix completion

Matrix completion arises in a wide variety of practical contexts, including collaborative filtering [2], system identification [3], sensor localization [4–6], rank aggregation [7], and many more. There is now a strong base of theoretical results concerning matrix completion [8–11]; a typical result is that a generic $d \times d$ matrix of rank r can be exactly recovered from $O(r\,d\,\mathrm{polylog}(d))$ randomly chosen entries.

Although these results are quite impressive, there is an important gap between the statement of the problem as considered in the matrix completion literature and many of the most common applications discussed therein. As an example, consider collaborative filtering and the now-famous "Netflix problem." In this setting, we assume that there is some unknown matrix whose entries each represent a rating for a particular user on a particular movie. Since any user will rate only a small subset of possible movies, we are only able to observe a small fraction of the total entries in the matrix, and our goal is to infer the unseen ratings from the observed ones. If the rating matrix is low-rank, then this would seem to be the exact problem studied in the matrix completion literature. However, there is a subtle difference: the theory developed in this literature generally assumes that observations consist of (possibly noisy) continuous-valued entries of the matrix, whereas in the Netflix problem the observations are "quantized" to the set of integers between 1 and 5. If we believe that it is possible for a user's true rating for a particular movie to be, for example, 4.5, then we must account for the impact of this "quantization noise" on our recovery. One could potentially treat quantization simply as a form of bounded noise, but this is unsatisfying because the ratings aren't just quantized — there are also hard limits placed on the minimum and maximum allowable ratings. (Why should we suppose that a movie given a rating of 5 could not have a true underlying rating of 6 or 7 or 10?) The inadequacy of standard matrix completion techniques in dealing with this effect is particularly pronounced when we consider recommender systems where each rating consists of a single bit representing a positive or negative rating (consider for example rating music on Pandora, the relevance of advertisements on Hulu, or posts on Reddit or MathOverflow). In such a case, the assumptions made in the existing theory of matrix completion do not apply, standard algorithms are ill-posed, and a new theory is required.

B. Outline and notation

The remainder of the paper is organized as follows. In Section II we provide an overview of the 1-bit matrix completion problem, introducing the observation model, describing the assumptions necessary for 1-bit matrix completion to be well-posed, and summarizing the upper bounds established in [1]. In Section III we describe a series of lower bounds that show that the upper bounds presented in Section II are nearly optimal.

Finally, Section IV concludes with a discussion of future work.

We now provide a brief summary of some of the key notation used in this paper. We use $[d]$ to denote the set of integers $\{1, \ldots, d\}$. We use capital boldface to denote a matrix (e.g., $\boldsymbol{M}$) and standard text to denote its entries (e.g., $M_{i,j}$). We let $\|\boldsymbol{M}\|$ denote the operator norm of $\boldsymbol{M}$, $\|\boldsymbol{M}\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ denote the Frobenius norm of $\boldsymbol{M}$, $\|\boldsymbol{M}\|_*$ denote the nuclear or Schatten-1 norm of $\boldsymbol{M}$ (the sum of the singular values), and $\|\boldsymbol{M}\|_\infty = \max_{i,j} |M_{i,j}|$ denote the entry-wise infinity-norm of $\boldsymbol{M}$. For two scalars $p, q \in [0, 1]$, the *Hellinger distance*,

$$d_H^2(p, q) := (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2$$

is a standard notion of distance between two binary probability distributions. We allow the Hellinger distance to act on matrices in a natural way: for matrices $\boldsymbol{P}, \boldsymbol{Q} \in [0, 1]^{d_1 \times d_2}$, we define

$$d_H^2(\boldsymbol{P}, \boldsymbol{Q}) := \frac{1}{d_1 d_2} \sum_{i,j} d_H^2(P_{i,j}, Q_{i,j}).$$

II. THE 1-BIT MATRIX COMPLETION PROBLEM

A. Observation model

We now introduce the 1-bit observation model considered in [1]. Given a matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$, a subset of indices $\Omega \subset [d_1] \times [d_2]$, and a differentiable function $f : \mathbb{R} \to [0, 1]$, we observe

$$Y_{i,j} = \begin{cases} +1 & \text{w. p. } f(M_{i,j}), \\ -1 & \text{w. p. } 1 - f(M_{i,j}) \end{cases} \quad \text{for } (i,j) \in \Omega. \quad (1)$$

When the function f behaves like a cumulative distribution function, the model (1) is equivalent to the following natural *latent variable* model. Suppose that $\boldsymbol{Z}$ is a random matrix with i.i.d. entries, and choose $f(x) := \mathbb{P}(Z_{1,1} \geq -x)$. Then we can rewrite (1) as

$$Y_{i,j} = \begin{cases} +1 & \text{if } M_{i,j} + Z_{i,j} \geq 0 \\ -1 & \text{if } M_{i,j} + Z_{i,j} < 0 \end{cases} \quad \text{for } (i,j) \in \Omega. \quad (2)$$

As examples, we consider two natural choices for f (or equivalently, for $\boldsymbol{Z}$).

Example 1 (Logistic regression/Logistic noise)**.** The logistic regression model, which is common in statistics, is captured by (1) with $f(x) = \frac{e^x}{1+e^x}$ and by (2) with $Z_{i,j}$ i.i.d. according to the standard logistic distribution.

Example 2 (Probit regression/Gaussian noise)**.** The probit regression model is captured by (1) by setting $f(x) = 1 - \Phi(-x/\sigma) = \Phi(x/\sigma)$ where Φ is the cumulative distribution function of a standard Gaussian and by (2) with $Z_{i,j}$ i.i.d. according to a mean-zero Gaussian distribution with variance σ^2.

As has been important in previous work on matrix completion, in order to show that 1-bit matrix completion is feasible we will assume that Ω is chosen at random with $\mathbb{E}|\Omega| = m$. Specifically, Ω follows a binomial model in which each entry

$(i, j) \in [d_1] \times [d_2]$ is included in Ω with probability $\frac{m}{d_1 d_2}$, independently. However, our lower bounds will not require this assumption, and hold for arbitrary sets Ω.

B. Assumptions

The majority of the literature on matrix completion assumes that the first r singular values of $\boldsymbol{M}$ are nonzero and the remainder are exactly zero. However, in many applications the singular values instead exhibit only a gradual decay towards zero. Thus, in this paper we allow a relaxation of the assumption that $\boldsymbol{M}$ has rank exactly r. Instead, we assume that $\|\boldsymbol{M}\|_* \leq \alpha\sqrt{r d_1 d_2}$, where α is a parameter left to be determined, but which will often be of constant order. In other words, the singular values of $\boldsymbol{M}$ belong to a scaled ℓ_1 ball. In compressed sensing, belonging to an ℓ_p ball with $p \in (0, 1]$ is a common relaxation of exact sparsity; in matrix completion, the nuclear-norm ball (or Schatten-1 ball) plays an analogous role.

The particular choice of scaling, $\alpha\sqrt{r d_1 d_2}$, arises from the following considerations. Suppose that each entry of $\boldsymbol{M}$ is bounded in magnitude by α and that $\operatorname{rank}(\boldsymbol{M}) \leq r$. Then

$$\|\boldsymbol{M}\|_* \leq \sqrt{r}\,\|\boldsymbol{M}\|_F \leq \sqrt{r d_1 d_2}\,\|\boldsymbol{M}\|_\infty \leq \alpha\sqrt{r d_1 d_2}.$$

Thus, the assumption that $\|\boldsymbol{M}\|_* \leq \alpha\sqrt{r d_1 d_2}$ is a relaxation of the conditions that $\operatorname{rank}(\boldsymbol{M}) \leq r$ and $\|\boldsymbol{M}\|_\infty \leq \alpha$. The condition that $\|\boldsymbol{M}\|_\infty \leq \alpha$ essentially means that the probability of seeing a $+1$ or -1 does not depend on the dimension. It is also a way of enforcing that $\boldsymbol{M}$ should not be too "spiky"; this is an important assumption in order to make the recovery of $\boldsymbol{M}$ well-posed (for example, see [12]).

Finally, some assumptions must be made on f for recovery of $\boldsymbol{M}$ to be feasible. We define two quantities L_α and β_α which control the "steepness" and "flatness" of f, respectively:

$$L_\alpha := \sup_{|x| \leq \alpha} \frac{|f'(x)|}{f(x)(1 - f(x))} \quad (3)$$

and

$$\beta_\alpha := \sup_{|x| \leq \alpha} \frac{f(x)(1 - f(x))}{(f'(x))^2}. \quad (4)$$

We will restrict our attention to f such that L_α and β_α are well-defined. In particular, we assume that f and f' are nonzero in $[-\alpha, \alpha]$. This assumption is fairly mild—for example, it includes the logistic and probit models (as we will see below in Remark 1). The quantity L_α appears only in the upper bounds, but it is generally well behaved. The quantity β_α appears both in the upper and lower bounds. Intuitively, it controls the "flatness" of f in the interval $[-\alpha, \alpha]$—the flatter f is, the larger β_α is. It is clear that some dependence on β_α is necessary. Indeed, if f is perfectly flat, then the magnitudes of the entries of $\boldsymbol{M}$ cannot be recovered. For example, if $\boldsymbol{M}$ is a rank-1 matrix, i.e., we can write $\boldsymbol{M} = \boldsymbol{u}\boldsymbol{v}^T$ for some pair of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, then without noise, $\boldsymbol{M}$ is indistinguishable from $\widetilde{\boldsymbol{M}} = \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^T$, where $\widetilde{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{v}}$ have the same sign patterns and $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Of course, when α is a fixed constant and f is a fixed function, both L_α and β_α are bounded by fixed constants independent of the dimension.

C. Upper bounds for 1-bit matrix completion

In this section, we review the upper bounds of [1], which have two goals. The first goal is to accurately recover M itself, and the second is to accurately recover the distribution of Y given by $f(M)$.[1] Both of these results are obtained by maximizing the log-likelihood function of the optimization variable X given our observations subject to a set of convex constraints. In many cases, including the logistic and probit models, the log-likelihood function is concave, and hence recovery amounts to solving a convex program.

Theorem 1. *Assume that $\|M\|_* \leq \alpha\sqrt{d_1 d_2 r}$ and $\|M\|_\infty \leq \alpha$. Suppose that Ω is chosen at random following the binomial model of Section II-A with $\mathbb{E}|\Omega| = m$. Suppose that Y is generated as in (1). Let L_α and β_α be as in (3) and (4). If $m \geq (d_1 + d_2)\log(d_1 d_2)$, then there is an algorithm which outputs $\widehat{M}$ such that with probability at least $1 - C_1/(d_1 + d_2)$,*

$$\frac{1}{d_1 d_2}\|\widehat{M} - M\|_F^2 \leq C_\alpha \sqrt{\frac{r(d_1 + d_2)}{m}} \tag{5}$$

with $C_\alpha := C_2 \alpha L_\alpha \beta_\alpha$. Above, C_1 and C_2 are absolute constants.

Remark 1 (Recovery in the logistic and probit models). The logistic model satisfies the hypotheses of Theorem 1 with $\beta_\alpha = \frac{(1+e^\alpha)^2}{e^\alpha} \approx e^\alpha$ and $L_\alpha = 1$. The probit model has

$$\beta_\alpha \leq c_1 \sigma^2 e^{\frac{\alpha^2}{2\sigma^2}} \quad \text{and} \quad L_\alpha \leq c_2 \frac{\frac{\alpha}{\sigma} + 1}{\sigma}$$

where we can take $c_1 = \pi$ and $c_2 = 8$. In particular, in the probit model the bound in (5) reduces to

$$\frac{1}{d_1 d_2}\|\widehat{M} - M\|_F^2 \leq C\alpha(\alpha + \sigma) e^{\frac{\alpha^2}{2\sigma^2}} \sqrt{\frac{r(d_1 + d_2)}{m}}. \tag{6}$$

Hence, when $\sigma < \alpha$, increasing the size of the noise leads to significantly improved error bounds—this is not an artifact of the proof. We will see in Section III that the exponential dependence on α in the logistic model (and on α/σ in the probit model) is intrinsic to the problem. Intuitively we should expect this since for such models, as $\|M\|_\infty$ grows large, we essentially revert to the noiseless setting where estimation of M is impossible. Furthermore, in Section III we will also see that when α (or α/σ) is bounded by a constant, the error bound (5) is optimal up to a constant factor. Fortunately, in many applications, one would expect α to be small, and in particular to have little, if any, dependence on the dimension.

In many situations, we might not be interested in the underlying matrix M, but rather in determining the distribution of the unknown entries of Y. For example, in recommender systems, a natural question would be to determine the likelihood that a user would enjoy a particular unrated item. Surprisingly, this distribution may be accurately recovered without any restriction on the infinity-norm of M. This may be unexpected

to those familiar with the matrix completion literature in which "non-spikiness" constraints seem to be unavoidable. In fact, we will show in Section III that the bound in Theorem 2 is near-optimal—even under the added constraint that $\|M\|_\infty \leq \alpha$, it would be impossible to estimate $f(M)$ significantly more accurately.

Theorem 2. *Assume that $\|M\|_* \leq \alpha\sqrt{d_1 d_2 r}$. Suppose that Ω is chosen at random following the binomial model of Section II-A with $\mathbb{E}|\Omega| = m$. Suppose that Y is generated as in (1), and let $L = \lim_{\alpha\to\infty} L_\alpha$. If $m \geq (d_1 + d_2)\log(d_1 d_2)$, then there is an algorithm which outputs $\widehat{M}$ such that with probability at least $1 - C_1/(d_1 + d_2)$,*

$$d_H^2(f(\widehat{M}), f(M)) \leq C_2 \alpha L \sqrt{\frac{r(d_1 + d_2)}{m}}. \tag{7}$$

Above, C_1 and C_2 are absolute constants.

While $L = 1$ for the logistic model, the astute reader will have noticed that for the probit model L is unbounded—that is, L_α tends to ∞ as $\alpha \to \infty$. L would also be unbounded for the case where $f(x)$ takes values of 1 or 0 outside of some range (as would be the case in (2) if the distribution of the noise had compact support). Fortunately, one can recover a result for these cases by enforcing an infinity-norm constraint [1].

III. LOWER BOUNDS

The upper bounds in Theorems 1 and 2 show that we may recover M or $f(M)$ from incomplete, binary measurements. In this section, we discuss the extent to which these theorems are optimal. We give three theorems, all proved using information theoretic methods, which show that these results are nearly tight, even when some of the assumptions are relaxed. Theorem 3 gives a lower bound to nearly match the upper bound on the error in recovering M derived in Theorem 1. Theorem 4 compares our upper bounds to those available without discretization and shows that very little is lost when discretizing to a single bit. Finally, Theorem 5 gives a lower bound matching, up to a constant factor, the upper bound on the error in recovering the distribution $f(M)$ given in Theorem 2. Theorem 5 also shows that Theorem 2 does not suffer by dropping the canonical "spikiness" constraint.

Our lower bounds require a few assumptions, so before we delve into the bounds themselves, we briefly argue that these assumptions are innocuous. For notational simplicity, we will assume without loss of generality that $d_1 \geq d_2$ throughout this section. Similarly, without loss of generality (since we can always adjust f to account for rescaling M), we assume that $\alpha \geq 1$. Next, we require that the parameters be sufficiently large so that

$$\alpha^2 r d_1 \geq C_0 \tag{8}$$

for an absolute constant C_0. Note that we could replace this with a simpler, but still mild, condition that $d_1 > C_0$. Finally, we also require that $r \geq c$ where c is a constant and that $r \leq O(d_2/\alpha^2)$, where $O(\cdot)$ hides parameters (which may differ in each theorem) that we make explicit below. This last

---

[1]Strictly speaking, $f(M) \in [0,1]^{d_1 \times d_2}$ is simply a matrix of scalars, but these scalars implicitly define the distribution of Y, so we will sometimes abuse notation slightly and refer to $f(M)$ as the distribution of Y.

assumption simply means that we are in the situation where r is significantly smaller than d_1 and d_2, i.e., the matrix is of approximately low rank.

In the following, let

$$K = \left\{ \boldsymbol{M} \ : \ \|\boldsymbol{M}\|_* \le \alpha\sqrt{rd_1 d_2}, \|\boldsymbol{M}\|_\infty \le \alpha \right\} \quad (9)$$

denote the set of matrices whose recovery is guaranteed by Theorem 1.

A. Recovery from 1-bit measurements

First, we show that the error in Theorem 1 is nearly optimal: there is no algorithm which can approximate $\boldsymbol{M}$ much better than the algorithms in [1].

Theorem 3. *Fix α, r, d_1, and d_2 to be such that $r \ge 4$ and (8) holds. Let β_α be defined as in (4), and suppose that $f'(x)$ is decreasing for $x > 0$. Let Ω be any subset of $[d_1] \times [d_2]$ with cardinality m, and let $\boldsymbol{Y}$ be as in (1). Consider any algorithm which, for any $\boldsymbol{M} \in K$, takes as input $Y_{i,j}$ for $(i,j) \in \Omega$ and returns $\widehat{\boldsymbol{M}}$. Then, there exists $\boldsymbol{M} \in K$ such that with probability at least $3/4$,*

$$\frac{1}{d_1 d_2}\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\|_F^2 \ge \min\left\{ C_1, C_2 \alpha \sqrt{\beta_{\frac{3}{4}\alpha}} \sqrt{\frac{rd_1}{m}} \right\} \quad (10)$$

as long as the right-hand side of (10) exceeds $r\alpha^2/d_2$. Above, C_1 and C_2 are absolute constants.[2]

The requirement that the right-hand side of (10) be larger than $r\alpha^2/d_2$ is satisfied as long as $r \le O(d_2/\alpha^2)$. In particular, it is satisfied whenever

$$r \le C_3 \frac{\min\{1, \beta_0\} \cdot d_2}{\alpha^2}$$

for a fixed constant C_3. Note also that in the latent variable model in (2), $f'(x)$ is simply the probability density of $Z_{i,j}$. Thus, the requirement that $f'(x)$ be decreasing is simply that the probability density have decreasing tails. One can easily check that this is satisfied for the logistic and probit models.

Note that if α is bounded by a constant and f is fixed (in which case β_α and $\beta_{\alpha'}$ are bounded by a constant), then the lower bound of Theorem 3 matches the upper bound given in (5) up to a constant. When α is not treated as a constant, the bounds differ by a factor of $\sqrt{\beta_\alpha}$. In the logistic model $\beta_\alpha \approx e^\alpha$ and so this amounts to the difference between $e^{\alpha/2}$ and e^α. The probit model has a similar change in the constant of the exponent.

B. Recovery from unquantized measurements

Next we show that, surprisingly, very little is lost by discretizing to a single bit. In Theorem 4, we consider an "unquantized" version of the latent variable model in (2) with Gaussian noise. That is, let $\boldsymbol{Z}$ be a matrix of i.i.d. Gaussian random variables, and suppose that the noisy entries

---

[2] Here and in the theorems below, the choice of $3/4$ in the probability bound is arbitrary, and can be adjusted at the cost of changing C_0 in (8) and C_1 and C_2. Similarly, $\beta_{\frac{3}{4}\alpha}$ can be replaced by $\beta_{(1-\epsilon)\alpha}$ for any $\epsilon > 0$.

$M_{i,j} + Z_{i,j}$ are observed directly, without discretization. In this setting, we give a lower bound that still nearly matches the upper bound given in Theorem 1, up to the β_α term.

Theorem 4. *Fix α, r, d_1, and d_2 to be such that $r \ge 1$ and (8) holds. Let Ω be any subset of $[d_1] \times [d_2]$ with cardinality m, and let $\boldsymbol{Z}$ be a $d_1 \times d_2$ matrix with i.i.d. Gaussian entries with variance σ^2. Consider any algorithm which, for any $\boldsymbol{M} \in K$, takes as input $Y_{i,j} = M_{i,j} + Z_{i,j}$ for $(i,j) \in \Omega$ and returns $\widehat{\boldsymbol{M}}$. Then, there exists $\boldsymbol{M} \in K$ such that with probability at least $3/4$,*

$$\frac{1}{d_1 d_2}\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\|_F^2 \ge \min\left\{ C_1, C_2 \alpha\sigma \sqrt{\frac{rd_1}{m}} \right\} \quad (11)$$

as long as the right-hand side of (11) exceeds $r\alpha^2/d_2$. Above, C_1 and C_2 are absolute constants.

The requirement that the right-hand side of (11) be larger than $r\alpha^2/d_2$ is satisfied whenever

$$r \le C_3 \frac{\min\{1, \sigma^2\}d_2}{\alpha^2}$$

for a fixed constant C_3.

Following Remark 1, the lower bound given in (11) matches the upper bound proven in Theorem 1 up to a constant, as long as α/σ is bounded by a constant. In other words:

> *When the signal-to-noise ratio is constant, almost nothing is lost by quantizing to a single bit.*

Perhaps it is not particularly surprising that 1-bit quantization induces little loss of information in the regime where the noise is comparable to the underlying quantity we wish to estimate—however, what *is* somewhat of a surprise is that a simple convex program can successfully recover all of the information contained in these 1-bit measurements.

Before proceeding, we also briefly note that our Theorem 4 is somewhat similar to Theorem 3 in [12]. The authors in [12] consider slightly different sets K: these sets are more restrictive in the sense that it is required that $\alpha \ge \sqrt{32\log n}$ and less restrictive because the nuclear-norm constraint may be replaced by a general Schatten-p norm constraint. It was important for us to allow $\alpha = O(1)$ in order to compare with our upper bounds due to the exponential dependence of β_α on α in Theorem 1 for the probit model. This led to some new challenges in the proof. Finally, it is also noteworthy that our statements hold for arbitrary sets Ω, while the argument in [12] is only valid for a random choice of Ω.

C. Recovery of the distribution from 1-bit measurements

Finally, we address the optimality of Theorem 2. We show that under mild conditions on f, any algorithm that recovers the distribution $f(\boldsymbol{M})$ must yield an estimate whose Hellinger distance deviates from the true distribution by an amount proportional to $\alpha\sqrt{rd_1 d_2/m}$, matching the upper bound of (7) up to a constant. Notice that the lower bound holds even if the algorithm is promised that $\|\boldsymbol{M}\|_\infty \le \alpha$, which the upper bound did not require.

Theorem 5. *Fix α, r, d_1, and d_2 to be such that $r \geq 4$ and (8) holds. Let L_1 be defined as in (3), and suppose that $f'(x) \geq c$ and $c' \leq f(x) \leq 1 - c'$ for $x \in [-1, 1]$, for some constants $c, c' > 0$. Let Ω be any subset of $[d_1] \times [d_2]$ with cardinality m, and let Y be as in (1). Consider any algorithm which, for any $M \in K$, takes as input $Y_{i,j}$ for $(i, j) \in \Omega$ and returns $\widehat{M}$. Then, there exists $M \in K$ such that with probability at least $3/4$,*

$$d_H^2(f(M), f(\widehat{M})) \geq \min\left\{ C_1, C_2 \frac{\alpha}{L_1} \sqrt{\frac{rd_1}{m}} \right\} \qquad (12)$$

as long as the right-hand side of (12) exceeds $r\alpha^2/d_2$. Above, C_1 and C_2 are constants that depend on c, c'.

The requirement that the right-hand side of (12) be larger than $r\alpha^2/d_2$ is satisfied whenever

$$r \leq C_3 \min\left\{ 1, \frac{1}{L_1^2} \right\} \frac{d_2}{\alpha^2}$$

for a constant C_3 that depends only on c, c'. Note also that the condition that f and f' be well-behaved in the interval $[-1, 1]$ is satisfied for the logistic model with $c = 1/4$ and $c' = \frac{1}{1+e} \leq 0.269$. Similarly, we may take $c = 0.242$ and $c' = 0.159$ in the probit model.

D. Proof sketches

All of our lower bounds follow the same general outline, and rest on Fano's inequality and the following lemma.

Lemma 1. *Let K be defined as in (9), let $\gamma \leq 1$ be such that $\frac{r}{\gamma^2}$ is an integer, and suppose that $\frac{r}{\gamma^2} \leq d_1$. There is a set $\mathcal{X} \subset K$ with*

$$|\mathcal{X}| \geq \exp\left(\frac{rd_2}{16\gamma^2} \right)$$

with the following properties:

1) *For all $X \in \mathcal{X}$, each entry has $|X_{i,j}| = \alpha\gamma$.*
2) *For all $X^{(i)}, X^{(j)} \in \mathcal{X}$, $i \neq j$,*

$$\|X^{(i)} - X^{(j)}\|_F^2 > \frac{\alpha^2\gamma^2 d_1 d_2}{2}.$$

The proof of Lemma 1 is probabilistic. Briefly, we consider drawing each the matrices $X^{(i)}$ independently, as follows: we draw independent sign flips (appropriately normalized) to fill an appropriately sized "block" of $X^{(i)}$. To fill out the rest of $X^{(i)}$, we simply repeat this block. It will turn out that the expected distance $\mathbb{E} \|X^{(i)} - X^{(j)}\|_F^2$ is large for distinct i, j, and so the proof rests on showing that these distances (which are random variables) are sufficiently concentrated that the probability of choosing a set with the required minimum distance is greater than zero.

With Lemma 1 in hand, the proofs of our lower bounds follow a similar framework, with different parameter settings in Lemma 1. Suppose that a recovery algorithm exists, which approximately recovers M or $f(M)$ with few observations $Y|_\Omega$. We will imagine running the recovery algorithm on

a randomly chosen element $X^{(i)}$ from the set $\mathcal{X}$. If the algorithm is successful, we will arrange the parameters so that it recovers the correct $X^{(i)}$ to within a Frobenius distance at most half the minimum distance of $\mathcal{X}$. Thus, we may use this algorithm to recover $X^{(i)}$ exactly, with good probability. On the other hand, we will show that the mutual information between the correct choice of $X^{(i)}$ and the output of the algorithm cannot be too large. Thus, Fano's inequality will imply that such recovery is impossible, and thus no such algorithm exists. The interested reader is referred to the preprint [1] for more details about the proofs.

IV. CONCLUSION

Many of the applications of matrix completion consider discrete data, sometimes consisting of binary measurements. This work addresses such situations. However, matrix completion from noiseless binary measurements is extremely ill-posed, even if one collects a binary measurement from all of the matrix entries. Fortunately, when there are some stochastic variations (noise) in the problem, matrix reconstruction becomes well-posed, and in [1] it was shown that the unknown matrix can be accurately and efficiently recovered from binary measurements. In this work, we focus on lower bounds, and show that the aforementioned upper bounds are nearly optimal, even when some restrictions are dropped.

REFERENCES

[1] M. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Arxiv preprint arxiv:1209.3672*, 2012.
[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Comm. ACM*, vol. 35, no. 12, pp. 61–70, 1992.
[3] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.
[4] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, "Semidefinite programming based algorithms for sensor network localization," *ACM Trans. Sen. Netw.*, vol. 2, no. 2, pp. 188–220, 2006.
[5] A. Singer, "A remark on global positioning from local distances," *Proc. Natl. Acad. Sci.*, vol. 105, no. 28, pp. 9507–9511, 2008.
[6] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM J. Matrix Analysis and Applications*, vol. 31, no. 4, pp. 1621–1641, 2010.
[7] D. Gleich and L.-H. Lim, "Rank aggregation via nuclear norm minimization," in *Proc. ACM SIGKDD Int. Conf. on Knowledge, Discovery, and Data Mining (KDD)*, San Diego, CA, Aug. 2011.
[8] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.
[9] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[10] E. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, 2010.
[11] B. Recht, "A simpler approach to matrix completion," *J. Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.
[12] S. Negahban and M. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Machine Learning Research*, vol. 13, pp. 1665–1697, 2012.