

A Max-Norm Constrained Minimization Approach to 1-Bit Matrix Completion

Tony Cai

*Statistics Department
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA*

TCAI@WHARTON.UPENN.EDU

Wen-Xin Zhou

*Department of Mathematics and Statistics
University of Melbourne
Parkville, VIC 3010, Australia*

WENXIN.ZHOU@UNIMELB.EDU.AU

Editor: Nathan Srebro

Abstract

We consider in this paper the problem of noisy 1-bit matrix completion under a general non-uniform sampling distribution using the max-norm as a convex relaxation for the rank. A max-norm constrained maximum likelihood estimate is introduced and studied. The rate of convergence for the estimate is obtained. Information-theoretical methods are used to establish a minimax lower bound under the general sampling model. The minimax upper and lower bounds together yield the optimal rate of convergence for the Frobenius norm loss. Computational algorithms and numerical performance are also discussed.

Keywords: 1-bit matrix completion, low-rank matrix, max-norm, trace-norm, constrained optimization, maximum likelihood estimate, optimal rate of convergence

1. Introduction

Matrix completion, which aims to recover a low-rank matrix from a subset of its entries, has been an active area of research in the last few years. It has a range of successful applications. In some real-life situations, however, the observations are highly *quantized*, sometimes even to a single bit and thus the standard matrix completion techniques do not apply. Take the Netflix problem as an example, the observations are the ratings of movies, which are quantized to the set of integers from 1 to 5. In the more extreme case such as recommender systems, only a single bit of rating standing for a “thumbs up” or “thumbs down” is recorded at each occurrence. Another example of applications is targeted advertising, such as the relevance of advertisements on Hulu. Each user who is watching TV shows on Hulu is required to answer yes/no to the question “*Is this ad relevant to you?*”. Noise effect should be considered since there are users who just click no to all the advertisements. In general, people would prefer to have advertisement catered to them, rather than to endure random advertisement. Targeted marketing that uses customer needs tends to serve better than random, scattershot advertisements. Similar idea has already been employed in mail system (Goldberg et al., 1992). Other examples from recommender systems include rating music on Pandora and posts on Reddit or MathOverflow, in which each observation consists of a single bit representing a positive

or negative rating. Similar problem also arises in analyzing incomplete survey designs containing simple agree/disagree questions in the analysis of survey data, and distance matrix recovery in multidimensional scaling that incorporates binary responses with incomplete data (Green and Wind, 1973; Spence and Demoney, 1974). See Davenport et al. (2012) for more discussions on potential applications.

Motivated by these applications, Davenport et al. (2012) considered the *1-bit matrix completion problem* of recovering an approximately low-rank matrix $M^* \in \mathbb{R}^{d_1 \times d_2}$ from a set of n noise corrupted sign (1-bit) measurements. In particular, they proposed a trace-norm constrained maximum likelihood estimator to estimate M^* , based on a small number of binary samples observed according to a probability distribution determined by the entries of M^* . It was also shown that the trace-norm constrained optimization method is minimax rate-optimal under the uniform sampling model. This problem is closely connected to and in some respects more challenging than the *1-bit compressed sensing*, which was introduced and first studied in Boufounos and Baraniuk (2008). The 1-bit measurements are meant to model quantization in the extreme case, and a surprising fact is that when the signal-to-noise ratio is low, empirical evidence demonstrates that such extreme quantization can be optimal when constrained to a fixed bit budget (Laska and Baraniuk, 2012). See Plan and Vershynin (2013a) for the recent results and references on 1-bit compressed sensing.

To be more specific, consider an arbitrary unknown $d_1 \times d_2$ target matrix M^* with rank at most r . Suppose a subset $S = \{(i_1, j_1), \dots, (i_n, j_n)\}$ of entries of a binary matrix Y is observed, where the entries of Y depend on M^* in the following way:

$$Y_{i,j} = \begin{cases} +1, & \text{if } M_{i,j}^* + Z_{i,j} \geq 0, \\ -1, & \text{if } M_{i,j}^* + Z_{i,j} < 0. \end{cases}$$

Here $Z = (Z_{ij}) \in \mathbb{R}^{d_1 \times d_2}$ is a general noise matrix. This latent variable matrix model can be seen as a direct analogue to the usual 1-bit compressed sensing model, in which only the signs of measurements are observed. It is known that an s -sparse signal can still be approximately recovered from $O(s \log(d/s))$ random linear measurements. See, for example, Jacques et al. (2011), Plan and Vershynin (2013a), Plan and Vershynin (2013b) and Ai et al. (2013).

Contrary to the standard matrix completion model and many other statistical problems, random noise turns out to be helpful and has a positive effect in the 1-bit case, since the problem is ill-posed in the absence of noise as described in Davenport et al. (2012). In particular, when $Z = 0$ and $M^* = uv^T$ for some vectors $u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}$ having no zero coordinates, then the radically disparate matrix $\tilde{M} = \text{sign}(u)\text{sign}^T(v)$ will lead to the same observations Y . Thus M and \tilde{M} are indistinguishable. However, it has been surprisingly noticed that the problem may become well-posed when there are some additional stochastic variations, that is, $Z \neq 0$ is an appropriate random noise matrix. This phenomenon can be regarded as a ‘‘dithering’’ effect brought by random noise.

Although the trace-norm constrained optimization method has been shown to be minimax rate-optimal under the uniform sampling model, it remains unclear that the trace-norm is the best convex surrogate to the rank. A different convex relaxation for the rank, the matrix *max-norm*, has been duly noted in machine learning literature since Srebro et al. (2005), and it was shown to be empirically superior to the trace-norm for collaborative filtering problems. Regarding a real $d_1 \times d_2$ matrix as an operator that maps from \mathbb{R}^{d_2} to \mathbb{R}^{d_1} , its rank can be alternatively expressed as the smallest integer k , such that it is possible to express $M = UV^T$, where $U \in \mathbb{R}^{d_1 \times k}$ and $V \in \mathbb{R}^{d_2 \times k}$. In terms of the matrix factorization $M = UV^T$, we would like U and V to have a small number of columns. The number of columns of U and V can be relaxed in a different way from the usual trace-norm by the

so-called *max-norm* (Linial et al., 2007), which is defined by

$$\|M\|_{\max} = \min_{M=UV^T} \{\|U\|_{2,\infty} \|V\|_{2,\infty}\}, \quad (1)$$

where the infimum is over all factorizations $M = UV^T$ with $\|U\|_{2,\infty}$ being the operator norm of $U : \ell_2^k \rightarrow \ell_\infty^{d_1}$ and $\|V\|_{2,\infty}$ the operator norm of $V : \ell_2^k \rightarrow \ell_\infty^{d_2}$ (or, equivalently, $V^T : \ell_1^{d_2} \rightarrow \ell_2^k$) and $k = 1, \dots, \min(d_1, d_2)$. It is not hard to check that $\|U\|_{2,\infty}$ is equal to the largest ℓ_2 norm of the rows in U . Since ℓ_2 is a Hilbert space, $\|\cdot\|_{\max}$ indeed defines a norm on the space of operators between $\ell_1^{d_2}$ and $\ell_\infty^{d_1}$. Comparably, the trace-norm has a formulation similar to (1), as given below in Section 2.1.

Foygel and Srebro (2011) first used the max-norm for matrix completion under the uniform sampling distribution. Their results are direct consequences of a recent bound on the excess risk for a smooth loss function, such as the quadratic loss, with a bounded second derivative (Srebro et al., 2010). Matrix completion under a non-degenerate random sampling model was studied in Cai and Zhou (2013), where it was shown that the max-norm constrained minimization method is rate-optimal and it yields a more stable approximate recovery guarantee, with respect to the sampling distributions, than trace-norm based approaches.

Davenport et al. (2012) analyzed 1-bit matrix completion under the *uniform sampling model*, where observed entries are assumed to be sampled randomly and uniformly. In such a setting, the trace-norm constrained approach has been shown to achieve minimax rate of convergence. However, in certain application such as collaborative filtering, the uniform sampling model is over idealized. In the Netflix problem, for instance, the uniform sampling model is equivalent to assuming all users are equally likely to rate every movie and all movies are equally likely to be rated by any user. In practice, inevitably some users are more active than others and some movies are more popular and thus rated more frequently. Therefore, the sampling distribution is in fact non-uniform. In this scenario, Salakhutdinov and Srebro (2010) showed that the standard trace-norm relaxation can behave very poorly, and suggested to use a weighted variant of the trace-norm, which takes the sampling distribution into account. Since the true sampling distribution is most likely unknown and can only be estimated based on the locations of those entries that are revealed in the sample, what commonly used in practice is the empirically-weighted trace norm. Foygel et al. (2011) provided rigorous recovery guarantees for learning with the standard weighted, smoothed weighted and smoothed empirically-weighted trace-norms. In particular, they gave upper bounds on excess error, which show that there is no theoretical disadvantage of learning with smoothed empirical marginals as compared to learning with smoothed true marginals.

In this paper, we study matrix completion based on noisy 1-bit observations under a general (non-degenerate) sampling model using the max-norm as a convex relaxation for the rank. The rate of convergence for the max-norm constrained maximum likelihood estimate is obtained. A matching minimax lower bound is established under the general non-uniform sampling model using information-theoretical methods. The minimax upper and lower bounds together yield the optimal rate of convergence for the Frobenius norm loss. As a comparison with the max-norm constrained optimization approach, we also analyze the recovery guarantee of the weighted trace-norm constrained method in the setting of non-uniform sampling distributions. Our result includes an additional logarithmic factor, which might be an artifact of the proof technique. The numerical results in Section 5 show that, even when the sampling distribution is uniform, the max-norm based regularization might slightly outperform the corresponding trace-norm method. To sum up, the max-norm regularized approach indeed provides a unified and stable approximate recovery guarantee with re-

spect to the sampling distributions, while previously used approaches are based on different variants of the trace-norm which may sometimes seem artificial to practitioners.

When the noise distribution is Gaussian or more generally log-concave, the negative log-likelihood function for M , given the measurements, is convex, hence computing the max-norm constrained maximum likelihood estimate is a convex optimization problem. The computational effectiveness of this method is also studied, based on a first-order algorithm developed in Lee et al. (2010) for solving convex programs involving a max-norm constraint, which outperforms the semi-definite programming method used in Srebro et al. (2005). It will be shown in Section 4 that the convex optimization problem can be implemented in polynomial time as a function of the sample size and the matrix dimensions.

The rest of the paper is organized as follows. Section 2 begins with the basic notation and definitions, and then states a collection of useful results on the matrix norms, Rademacher complexity and distances between matrices that will be needed throughout the paper. Section 3 introduces the 1-bit matrix completion model and the estimation procedure and investigates the theoretical properties of the estimator. Both minimax upper and lower bounds are established. The results show that the max-norm constraint maximum likelihood estimator is rate-optimal over the parameter space. Section 3 also gives a comparison of our results with previous work. Computational algorithms are discussed in Section 4, and numerical performance of the proposed algorithm is presented in Section 5. The paper is concluded with a brief discussion in Section 6, and the proofs of the main results are given in Section 7.

2. Notations and Preliminaries

In this section, we introduce basic notation and definitions that will be used throughout the paper, and state some known results on the max-norm, trace-norm and Rademacher complexity that will be used repeatedly later.

2.1 Notation

For any positive integer d , we use $[d]$ to denote the set of integers $\{1, 2, \dots, d\}$. For any pair of real numbers a and b , set $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. For a vector $u \in \mathbb{R}^d$ and $0 < p < \infty$, denote its ℓ_p -norm by $\|u\|_p = (\sum_{i=1}^d |u_i|^p)^{1/p}$. In particular, $\|u\|_\infty = \max_{i=1, \dots, d} |u_i|$ is the ℓ_∞ -norm. For a matrix $M = (M_{k,l}) \in \mathbb{R}^{d_1 \times d_2}$, let $\|M\|_F = \sqrt{\sum_{k=1}^{d_1} \sum_{l=1}^{d_2} M_{k,l}^2}$ be the Frobenius norm and let $\|M\|_\infty = \max_{k,l} |M_{k,l}|$ denote the elementwise ℓ_∞ -norm. Given two norms ℓ_p and ℓ_q on \mathbb{R}^{d_1} and \mathbb{R}^{d_2} respectively, the corresponding operator norm $\|\cdot\|_{p,q}$ of a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ is defined by $\|M\|_{p,q} = \sup_{\|x\|_p=1} \|Mx\|_q$. It is easy to verify that $\|M\|_{p,q} = \|M^T\|_{q^*,p^*}$, where (p, p^*) and (q, q^*) are conjugate pairs, that is, $\frac{1}{p} + \frac{1}{p^*} = 1$ and $\frac{1}{q} + \frac{1}{q^*} = 1$. In particular, $\|M\| = \|M\|_{2,2}$ is the spectral norm and $\|M\|_{2,\infty} = \max_{k=1, \dots, d_1} \sqrt{\sum_{l=1}^{d_2} M_{k,l}^2}$ is the maximum row norm of M .

2.2 Max-Norm and Trace-Norm

For any matrix $M \in \mathbb{R}^{d_1 \times d_2}$, its *trace-norm* is defined to be the sum of the singular values of M (that is, the roots of the eigenvalues of MM^T), and can also equivalently written as

$$\|M\|_* = \inf \left\{ \sum_j |\sigma_j| : M = \sum_j \sigma_j u_j v_j^T, u_j \in \mathbb{R}^{d_1}, v_j \in \mathbb{R}^{d_2} \text{ satisfying } \|u_j\|_2 = \|v_j\|_2 = 1 \right\}.$$

Recall the definition (1) of the max-norm, the trace-norm can be analogously defined in terms of matrix factorization as

$$\|M\|_* = \min_{M=UV^T} \{ \|U\|_F \|V\|_F \} = \frac{1}{2} \min_{U,V:M=UV^T} (\|U\|_F^2 + \|V\|_F^2).$$

Since the ℓ_1 -norm of a vector is bounded by the product of its ℓ_2 -norm and the number of non-zero coordinates, we have the following relationship between the trace-norm and Frobenius norm

$$\|M\|_F \leq \|M\|_* \leq \sqrt{\text{rank}(M)} \cdot \|M\|_F.$$

By the elementary inequality $\|M_{m \times n}\|_F \leq \sqrt{m} \|M_{m \times n}\|_{2,\infty}$, we see that

$$\frac{\|M\|_*}{\sqrt{d_1 d_2}} \leq \|M\|_{\max}. \tag{2}$$

Furthermore, as was noticed in Lee et al. (2010), the max-norm, which is defined in (1), is comparable with a trace-norm more precisely in the following sense (Jameson, 1987):

$$\begin{aligned} & \|M\|_{\max} \tag{3} \\ & \approx \inf \left\{ \sum_j |\sigma_j| : M = \sum_j \sigma_j u_j v_j^T, u_j \in \mathbb{R}^{d_1}, v_j \in \mathbb{R}^{d_2} \text{ satisfying } \|u_j\|_\infty = \|v_j\|_\infty = 1 \right\}, \end{aligned}$$

where the factor of equivalence is $K_G \in (1.67, 1.79)$, denoting the Grothendieck's constant. What may be more surprising is the following bounds for the max-norm, in connection with element-wise ℓ_∞ -norm (Linial et al., 2007):

$$\|M\|_\infty \leq \|M\|_{\max} \leq \sqrt{\text{rank}(M)} \cdot \|M\|_{1,\infty} \leq \sqrt{\text{rank}(M)} \cdot \|M\|_\infty. \tag{4}$$

2.3 Rademacher Complexity

Considering matrices as functions from index pairs to entry values, a technical tool used in our proof involves data-dependent estimates of the *Rademacher complexity* of the classes that consist of low trace-norm and low max-norm matrices. We refer to Bartlett and Mendelson (2002) for a detailed introduction of this concept.

Definition 1 *Let \mathcal{P} be a probability distribution on a set \mathcal{X} . Suppose that X_1, \dots, X_n are independent samples drawn from \mathcal{X} according to \mathcal{P} , and set $S = \{X_1, \dots, X_n\}$. For a class \mathcal{F} of functions mapping from \mathcal{X} to \mathbb{R} , its empirical Rademacher complexity over the sample S is defined by*

$$\widehat{R}_S(\mathcal{F}) = \frac{2}{|S|} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \epsilon_i f(X_i) \right| \right],$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is a Rademacher sequence. The Rademacher complexity with respect to the distribution \mathcal{P} is the expectation, over a sample S of $|S|$ points drawn i.i.d. according to \mathcal{P} , denoted by

$$R_{|S|}(\mathcal{F}) = \mathbb{E}_{S \sim \mathcal{P}}[\widehat{R}_S(\mathcal{F})].$$

The following properties regarding $\widehat{R}_S(\mathcal{F})$ are useful.

Proposition 2 *We have*

1. If $\mathcal{F} \subseteq \mathcal{G}$, $\widehat{R}_S(\mathcal{F}) \leq \widehat{R}_S(\mathcal{G})$.
2. $\widehat{R}_S(\mathcal{F}) = \widehat{R}_S(\text{conv}(\mathcal{F})) = \widehat{R}_S(\text{absconv}(\mathcal{F}))$, where $\text{conv}(\mathcal{F})$ is the class of convex combinations of functions from \mathcal{F} , and $\text{absconv}(\mathcal{F})$ denotes the absolutely convex hull of \mathcal{F} , that is, the class of convex combinations of functions from \mathcal{F} and $-\mathcal{F}$.
3. For every $c \in \mathbb{R}$, $\widehat{R}_S(c\mathcal{F}) = |c|\widehat{R}_S(\mathcal{F})$, where $c\mathcal{F} \equiv \{cf : f \in \mathcal{F}\}$.

In particular, we are interested in calculating the Rademacher complexities of the trace-norm and max-norm balls. To this end, define for any radius $R > 0$ that

$$\begin{aligned} \mathbb{B}_*(R) &:= \{M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_* \leq R\} \quad \text{and} \\ \mathbb{B}_{\max}(R) &:= \{M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_{\max} \leq R\}. \end{aligned}$$

First, recall that any matrix with unit trace-norm is a convex combination of unit-norm rank-one matrices, and thus

$$\mathbb{B}_*(1) = \text{conv}(\mathcal{M}_1), \quad \text{where } \mathcal{M}_1 := \{uv^T : u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2}, \|u\|_2 = \|v\|_2 = 1\}. \quad (5)$$

Then $\widehat{R}_S(\mathbb{B}_*(1)) = \widehat{R}_S(\mathcal{M}_1)$. A sharp bound on the worst-case Rademacher complexity, defined as the supremum of $\widehat{R}_S(\cdot)$ over all sample sets S with size $|S| = n$, is $\frac{2}{\sqrt{n}}$ (See, expression (4) on page 551, Srebro and Shraibman, 2005). This bound, unfortunately, is barely useful in developing generalization error bounds. However, when the index pairs of a sample S are drawn uniformly at random from $[d_1] \times [d_2]$ (with replacement), Srebro and Shraibman (2005) showed that the *expected* Rademacher complexity is low, and Foygel and Srebro (2011) have improved this result by reducing the logarithmic factor. In particular, they proved that for a sample size $n \geq d = d_1 + d_2$,

$$\mathbb{E}_{S \sim \text{unif}, |S|=n}[\widehat{R}_S(\mathbb{B}_*(1))] \leq \frac{K}{\sqrt{d_1 d_2}} \sqrt{\frac{d \log(d)}{n}},$$

where $K > 0$ denotes a universal constant.

The unit max-norm ball, on the other hand, can be approximately characterized as a convex hull. Due to the Grothendieck's inequality, it was shown in Srebro and Shraibman (2005) that

$$\text{conv}(\mathcal{M}_{\pm}) \subset \mathbb{B}_{\max}(1) \subset K_G \cdot \text{conv}(\mathcal{M}_{\pm}),$$

where $\mathcal{M}_{\pm} := \{M \in \{\pm 1\}^{d_1 \times d_2} : \text{rank}(M) = 1\}$ is the class of rank-one sign matrices, and $K_G \in (1.67, 1.79)$ is the Grothendieck's constant. It is easy to see that \mathcal{M}_{\pm} is a finite class with cardinality

$|\mathcal{M}_\pm| = 2^{d-1}$, $d = d_1 + d_2$. For any $d_1, d_2 > 2$ and any sample of size $2 < |S| \leq d_1 d_2$, the empirical Rademacher complexity of the unit max-norm ball is bounded by

$$\widehat{R}_S(\mathbb{B}_{\max}(1)) \leq 12\sqrt{\frac{d}{|S|}}. \quad (6)$$

In other words, $\sup_{S:|S|=n} \widehat{R}_S(\mathbb{B}_{\max}(1)) \leq 12\sqrt{\frac{d}{n}}$.

2.4 Discrepancy

In order to get both upper and lower prediction error bounds on the weighted squared Frobenius norm between the proposed estimator, given by (13) below, and the target matrix described via model (9), we will need the following two concepts of discrepancies between matrices as well as their connections. In particular, we will focus on element-wise notion of discrepancy between two $d_1 \times d_2$ matrices P and Q .

First, for two matrices $P, Q: [d_1] \times [d_2] \rightarrow [0, 1]^{d_1 \times d_2}$, their Hellinger distance is given by

$$d_H^2(P; Q) = \frac{1}{d_1 d_2} \sum_{(k,l)} d_H^2(P_{k,l}; Q_{k,l}),$$

where $d_H^2(p; q) = (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2$ for $p, q \in [0, 1]$. Next, the Kullback-Leibler divergence between two matrices $P, Q: [d_1] \times [d_2] \rightarrow [0, 1]^{d_1 \times d_2}$ is defined by

$$\mathbb{K}(P\|Q) = \frac{1}{d_1 d_2} \sum_{(k,l)} K(P_{k,l}\|Q_{k,l}),$$

where $K(p\|q) = p \log(\frac{p}{q}) + (1-p) \log(\frac{1-p}{1-q})$, for $p, q \in [0, 1]$. Note that $\mathbb{K}(P\|Q)$ is not a distance; it is sufficient to observe that it is not symmetric.

The relationship between the two “distances” is as follows. For any two scalars $p, q \in [0, 1]$, we have

$$d_H^2(p; q) \leq K(p\|q), \quad (7)$$

which in turn implies that, for any two matrices $P, Q: [d_1] \times [d_2] \rightarrow [0, 1]^{d_1 \times d_2}$,

$$d_H^2(P; Q) \leq \mathbb{K}(P\|Q). \quad (8)$$

The proof of (7) is based on the Jensen’s inequality and an elementary inequality that $1 - x \leq -\log x$ for any $x > 0$.

3. Max-Norm Constrained Maximum Likelihood Estimate

In this section, we introduce the max-norm constrained maximum likelihood estimation procedure for 1-bit matrix completion and investigates the theoretical properties of the estimator. The results are also compared with other results in the literature.

3.1 Observation Model

We consider 1-bit matrix completion under a general random sampling model. The unknown low-rank matrix $M^* \in \mathbb{R}^{d_1 \times d_2}$ is the object of interest. Instead of observing noisy entries $M_{i,j}^* + Z_{i,j}$ directly in *unquantized* matrix completion, now we only observe with error the sign of a random subset of the entries of M^* . More specifically, assume that a random sample

$$S = \{(i_1, j_1), (i_2, j_2), \dots, (i_n, j_n)\} \subseteq ([d_1] \times [d_2])^n$$

of the index set is drawn i.i.d. with replacement according to a general sampling distribution $\Pi = \{\pi_{kl}\}$ on $[d_1] \times [d_2]$. That is, $\mathbb{P}\{(i_t, j_t) = (k, l)\} = \pi_{kl}$, for all t and (k, l) . Suppose that a (random) subset S of size $|S| = n$ of entries of a sign matrix Y is observed. The dependence of Y on the underlying matrix M^* is as follows:

$$Y_{i,j} = \begin{cases} +1, & \text{if } M_{i,j}^* + Z_{i,j} \geq 0, \\ -1, & \text{if } M_{i,j}^* + Z_{i,j} < 0, \end{cases} \quad (9)$$

where $Z = (Z_{i,j}) \in \mathbb{R}^{d_1 \times d_2}$ is a matrix consisting of i.i.d. noise variables. Let $F(\cdot)$ be the cumulative distribution function of $-Z_{1,1}$, then the above model can be recast as

$$Y_{i,j} = \begin{cases} +1, & \text{with probability } F(M_{i,j}^*), \\ -1, & \text{with probability } 1 - F(M_{i,j}^*), \end{cases} \quad (10)$$

and we observe noisy entries $\{Y_{i,j_t}\}_{t=1}^n$ indexed by S . More generally, we consider the model (10) with an arbitrary differentiable function $F : \mathbb{R} \rightarrow [0, 1]$. Particular assumptions on F will be discussed below.

Instead of assuming the uniform sampling distribution as in Davenport et al. (2012), here we allow a general sampling distribution $\Pi = \{\pi_{kl}\}$, satisfying $\sum_{(k,l) \in [d_1] \times [d_2]} \pi_{kl} = 1$, according to which we make n independent random choices of entries. The drawback of the setting is that, with fairly high probability, some entries will be sampled multiple times. Intuitively it would be more practical to assume that entries are sampled without replacement, or equivalently, to sample n of the $d_1 d_2$ binary entries observed with noise without replacing. Due to the requirement that the drawn entries be distinct, the n samples are not independent. This dependence structure turns out to impede the technical analysis of the learning guarantees. To avoid this complication, we will use the i.i.d. approach as a proxy for sampling without replacement throughout this paper. As has been noted in Gross and Nemes (2010) and Foygel and Srebro (2011), between sampling with and without replacement both in a uniform sense, that is, making n independent uniform choices of entries versus choosing a set S of entries uniformly at random over all subsets that consist of exactly n entries, the latter can be theoretically as good as the former. See Section 7.4 below for more details.

Next we list three natural choices for F , or equivalently, for the distribution of $\{Z_{i,j}\}$.

3.1.1 EXAMPLES

1. (Logistic regression/Logistic noise): The logistic regression model is described by (10) with

$$F(x) = \frac{e^x}{1 + e^x},$$

and equivalently by (9) with $Z_{i,j}$ i.i.d. following the standard logistic distribution.

2. (Probit regression/Gaussian noise): The probit regression model is described by (10) with

$$F(x) = \Phi\left(\frac{x}{\sigma}\right),$$

where Φ denotes the cumulative distribution function of $N(0, 1)$, and equivalently by (9) with $Z_{i,j}$ i.i.d. following $N(0, \sigma^2)$.

3. (Laplacian noise): Another interesting case is that the noise $Z_{i,j}$ are i.i.d. drawn from a Laplacian distribution $\text{Laplace}(0, b)$, with

$$F(x) = \begin{cases} \frac{1}{2} \exp(x/b), & \text{if } x < 0, \\ 1 - \frac{1}{2} \exp(-x/b), & \text{if } x \geq 0, \end{cases}$$

where $b > 0$ is the scale parameter.

Davenport et al. (2012) have focused on approximately low-rank matrices recovery by considering the following class of matrices

$$K_*(\alpha, r) = \left\{ M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_\infty \leq \alpha, \frac{\|M\|_*}{\sqrt{d_1 d_2}} \leq \alpha \sqrt{r} \right\}, \quad (11)$$

where $1 \leq r \leq \min(d_1, d_2)$ and $\alpha > 0$ is a free parameter to be determined. Clearly, any matrix M with rank at most r satisfying $\|M\|_\infty \leq \alpha$ belongs to $K_*(\alpha, r)$. Alternatively, using max-norm as a convex relaxation for the rank, we consider recovery of matrices with ℓ_∞ -norm and max-norm constraints defined by

$$K_{\max}(\alpha, R) := \left\{ M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_\infty \leq \alpha, \|M\|_{\max} \leq R \right\}. \quad (12)$$

Here both $\alpha > 0$ and $R > 0$ are free parameters to be determined. If M^* is of rank at most r and $\|M^*\|_\infty \leq \alpha$, then by (2) and (4) we have $M^* \in \mathbb{B}_{\max}(\alpha \sqrt{r})$ and hence

$$M^* \in K_{\max}(\alpha, \alpha \sqrt{r}) \subset K_*(\alpha, r).$$

3.2 Max-norm Constrained Maximum Likelihood Estimate

Now, given a collection of observations $Y_S = \{Y_{i,j_t}\}_{t=1}^n$ from the observation model (10), the negative log-likelihood function can be written as

$$\ell_S(M; Y) = \sum_{t=1}^n \left[\mathbf{1}_{\{Y_{i,j_t}=1\}} \log \left(\frac{1}{F(M_{i,j_t})} \right) + \mathbf{1}_{\{Y_{i,j_t}=-1\}} \log \left(\frac{1}{1 - F(M_{i,j_t})} \right) \right].$$

Then we consider estimating the unknown $M^* \in K_{\max}(\alpha, R)$ by maximizing the empirical likelihood function subject to a max-norm constraint:

$$\hat{M}_{\max} = \arg \min_{M \in K_{\max}(\alpha, R)} \ell_S(M; Y). \quad (13)$$

The optimization procedure requires that all the entries of M_0 are bounded in absolute value by a pre-defined constant α . This condition is reasonable while also critical in approximate low-rank

matrix recovery problems by controlling the *spikiness* of the solution. Indeed, the measure of the “spikiness” of matrices is much less restrictive than the incoherence conditions imposed in exact low-rank matrix recovery. See, for example, Koltchinskii et al. (2011), Negahban and Wainwright (2012), Klopp (2012) and Cai and Zhou (2013).

As has been noted in Srebro et al. (2005), a large gap between the max-complexity (related to max-norm) and the dimensional-complexity (related to rank) is possible only when the underlying low-rank matrix has entries of vastly varying magnitudes. Also, in view of (3), the max-norm promotes low-rank decomposition with factors in ℓ_∞ (ℓ_2 for the trace-norm). Motivated by these features, max-norm regularization is expected to be reasonably effective for uniformly bounded data.

When the noise distribution is log-concave so that the log-likelihood is a concave function, the max-norm constrained minimization problem (13) is a convex program and we recommend a fast and efficient algorithm developed in Lee et al. (2010) for solving large-scale optimization problems that incorporate the max-norm. We will show in Section 4 that the convex optimization problem (13) can indeed be implemented in polynomial time as a function of the sample size n and the matrix dimensions d_1 and d_2 .

3.3 Upper Bounds

To establish an upper bound on the prediction error of estimator \widehat{M}_{\max} given by (13), we need the following assumption on the unknown matrix M^* as well as the regularity conditions on the function F in (10).

3.3.1 CONDITION U

Assume that there exist positive constants R and α such that

(U1) $M^* \in K_{\max}(\alpha, R)$;

(U2) F and F' are non-zero in $[-\alpha, \alpha]$, and

(U3) both

$$L_\alpha := \sup_{|x| \leq \alpha} \frac{|F'(x)|}{F(x)(1-F(x))}, \quad \text{and} \quad \beta_\alpha := \sup_{|x| \leq \alpha} \frac{F(x)(1-F(x))}{(F'(x))^2} \tag{14}$$

are finite.

In particular, under condition (U2), the quantity

$$U_\alpha := \sup_{|x| \leq \alpha} \log \left(\frac{1}{F(x)(1-F(x))} \right),$$

is well-defined. As prototypical examples, we specify below the quantities L_α , β_α and U_α in the cases of Logistic, Gaussian and Laplacian noise:

- (Logistic regression/Logistic noise): For $F(x) = e^x/(1 + e^x)$, we have

$$L_\alpha \equiv 1, \quad \beta_\alpha = \frac{(1 + e^\alpha)^2}{e^\alpha} \quad \text{and} \quad U_\alpha = 2 \log(e^{\alpha/2} + e^{-\alpha/2}).$$

2. (Probit regression/Gaussian noise): For $F(x) = \Phi(x/\sigma)$, straightforward calculations show that

$$L_\alpha \leq \frac{4}{\sigma} \left(\frac{\alpha}{\sigma} + 1 \right), \quad \beta_\alpha \leq \pi\sigma^2 \exp\{\alpha^2/(2\sigma^2)\} \quad \text{and} \quad U_\alpha \leq \left(\frac{\alpha}{\sigma} + 1 \right)^2. \quad (15)$$

3. (Laplacian noise): For a Laplace(0, b) distribution function, we have

$$L_\alpha = \frac{2}{b}, \quad \beta_\alpha = b(2 \exp(\alpha/b) - 1) \quad \text{and} \quad U_\alpha \leq 2 \left(\frac{\alpha}{b} + \log 2 \right).$$

Now we are ready to state our main results concerning the recovery of an approximately low-rank matrix M^* using the max-norm constrained maximum likelihood estimate. We write hereafter $d = d_1 + d_2$ for brevity.

Theorem 3 *Suppose that Condition U holds and assume that the training set S follows a general weighted sampling model according to the distribution Π . Then there exists an absolute constant C such that, for a sample size $2 < n \leq d_1 d_2$ and for any $\delta > 0$, the minimizer \widehat{M}_{\max} of the optimization program (13) satisfies*

$$\|\widehat{M}_{\max} - M^*\|_\Pi^2 = \sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \pi_{kl} \{\widehat{M}_{\max} - M^*\}_{k,l}^2 \leq C\beta_\alpha \left\{ L_\alpha R \sqrt{\frac{d}{n}} + U_\alpha \sqrt{\frac{\log(4/\delta)}{n}} \right\}, \quad (16)$$

with probability at least $1 - \delta$. Here and below $\|\cdot\|_\Pi$ denotes the weighted Frobenius norm with respect to Π , that is,

$$\|M\|_\Pi = \sqrt{\sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \pi_{kl} M_{k,l}^2} \quad \text{for all } M \in \mathbb{R}^{d_1 \times d_2}.$$

Remark 4 (i) While using the trace-norm to study this general weighted sampling model, it is common to assume that each row and column is sampled with positive probability (Klopp, 2012; Negahban and Wainwright, 2012), though in some applications this assumption does not seem realistic. More precisely, assume that there exists a positive constant $\mu \geq 1$ such that

$$\pi_{kl} \geq \frac{1}{\mu d_1 d_2}, \quad \text{for all } (k, l) \in [d_1] \times [d_2]. \quad (17)$$

Then, under condition (17) and the conditions of Theorem 3,

$$\frac{1}{d_1 d_2} \|\widehat{M}_{\max} - M^*\|_F^2 \leq C\mu\beta_\alpha \left\{ L_\alpha R \sqrt{\frac{d}{n}} + U_\alpha \sqrt{\frac{\log(d)}{n}} \right\} \quad (18)$$

holds with probability at least $1 - 4/d$, where $C > 0$ denotes an absolute constant.

- (ii) Klopp (2012) studied the problem of standard matrix completion with noise, also in the case of general sampling distribution, using the trace-norm penalized approach. However, the Assumption 1 therein requires that the distribution π_{kl} over entries is bounded from above, which is quite restrictive especially in the Netflix problem. It is worth noticing that this upper bound condition on sampling distribution is not required in both results (16) and (18).

It is noteworthy that above results are directly comparable to those obtained in the case of approximately low-rank recovery from unquantized measurements, also using max-norm regularized approach (Cai and Zhou, 2013). Let $Z = (Z_{i,j})$ be a noise matrix consisting of i.i.d. $N(0, \sigma^2)$ entries for some $\sigma > 0$, and assume we have observations on a (random) subset $S = \{(i_1, j_1), \dots, (i_n, j_n)\}$ of entries of $\tilde{Y} = M^* + Z$. Cai and Zhou (2013) studied the unquantized problem under a general sampling model using max-norm as a convex relaxation for the rank. In particular, for the max-norm constrained least squares estimator

$$\tilde{M}_{\max} = \arg \min_{M \in K_{\max}(\alpha, R)} \frac{1}{n} \sum_{t=1}^n (\tilde{Y}_{i_t, j_t} - M_{i_t, j_t}^*)^2,$$

for $K_{\max}(\alpha, R)$ as in (12), it was shown that for any $\delta \in (0, 1)$ and a sample size $2 < n \leq d_1 d_2$,

$$\|\tilde{M}_{\max} - M^*\|_{\Pi}^2 \leq C' \left\{ (\alpha \vee \sigma) R \sqrt{\frac{d}{n}} + \frac{\alpha^2 \log(2/\delta)}{n} \right\} \quad (19)$$

holds with probability greater than $1 - \exp(-d) - \delta$, where $C' > 0$ is a universal constant.

In 1-bit observations case when $Z_{i,j} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, it is equivalent that the function F in model (10) is given by $F(\cdot) = \Phi(\cdot/\sigma)$. According to (15), we have

$$\|\hat{M}_{\max} - M^*\|_{\Pi}^2 \leq C \exp\left(\frac{\alpha^2}{2\sigma^2}\right) \left\{ (\alpha + \sigma) R \sqrt{\frac{d}{n}} + (\alpha + \sigma)^2 \sqrt{\frac{\log(4/\delta)}{n}} \right\} \quad (20)$$

holds with probability at least $1 - \delta$.

Comparing the upper bounds in (19) and (20) and note that $\alpha \vee \sigma \leq \alpha + \sigma \leq 2(\alpha \vee \sigma)$, we see that there is no essential loss of recovery accuracy by discretizing to binary measurements as long as $\frac{\alpha}{\sigma}$ is bounded by a constant (Davenport et al., 2012). On the other hand, as the signal-to-noise ratio $\frac{\alpha}{\sigma} \geq 1$ increases, the error bounds deteriorate significantly. In fact, the case $\alpha \gg \sigma$ essentially amounts to the noiseless setting, in which it is impossible to recover M^* based on any subset of the signs of its entries.

3.4 Information-Theoretic Lower Bounds

We now establish minimax lower bounds by using information-theoretic techniques. The lower bounds given in Theorem 5 below show that the rate attained by the max-norm constrained maximum likelihood estimator is optimal up to constant factors.

Theorem 5 *Assume that $F'(x)$ is decreasing and $\frac{F(x)(1-F(x))}{(F'(x))^2}$ is increasing for $x > 0$, and let S be any subset of $[d_1] \times [d_2]$ with cardinality n . Then, as long as the parameters (R, α) satisfy*

$$\max\left(2, \frac{4}{(d_1 \vee d_2)^{1/2}}\right) \leq \frac{R}{\alpha} \leq \frac{(d_1 \wedge d_2)^{1/2}}{2},$$

the minimax risk for estimating M over the parameter space $K_{\max}(\alpha, R)$ satisfies

$$\inf_{\hat{M}} \max_{M \in K_{\max}(\alpha, R)} \left\{ \frac{1}{d_1 d_2} \mathbb{E} \|\hat{M} - M\|_F^2 \right\} \geq \frac{1}{512} \min \left\{ \alpha^2, \frac{\sqrt{\beta_{\alpha/2}}}{2} R \sqrt{\frac{d}{n}} \right\}. \quad (21)$$

Remark 6 In fact, the lower bound (21) is a special case of the following general result, which will be proved in Sect. 7.2. Let $\gamma^* > 0$ be the solution of the following equation

$$\gamma^* = \min \left\{ \frac{1}{2}, \frac{R^{1/2}}{\alpha} \left(\frac{\beta_{(1-\gamma^*)\alpha}}{32} \cdot \frac{d_1 \vee d_2}{n} \right)^{1/4} \right\} \quad (22)$$

and assume that

$$\max \left(2, \frac{4}{(d_1 \vee d_2)^{1/2}} \right) \leq \frac{R}{\alpha} \leq (d_1 \wedge d_2)^{1/2} \gamma^*. \quad (23)$$

Then the minimax risk for estimating M over the parameter space $K_{\max}(\alpha, R)$ satisfies

$$\inf_{\hat{M}} \max_{M \in K_{\max}(\alpha, R)} \left\{ \frac{1}{d_1 d_2} \mathbb{E} \|\hat{M} - M\|_F^2 \right\} \geq \frac{1}{512} \min \left\{ \alpha^2, \frac{\sqrt{\beta_{(1-\gamma^*)\alpha}}}{2} R \sqrt{\frac{d}{n}} \right\}. \quad (24)$$

To see the existence of γ^* defined above, setting

$$h(\gamma) = \gamma \quad \text{and} \quad g(\gamma) = \min \left\{ \frac{1}{2}, \frac{R^{1/2}}{\alpha} \left(\frac{\beta_{(1-\gamma)\alpha}}{32} \cdot \frac{d_1 \vee d_2}{n} \right)^{1/4} \right\},$$

then it is easy to see that $h(\gamma)$ is strictly increasing and $g(\gamma)$ is decreasing for $\gamma \in (0, 1)$ with $h(0) = 0$ and $g(0) > 0$. Therefore, equation (22) has a unique solution $\gamma^* \in (0, \frac{1}{2}]$ so that $h(\gamma^*) = g(\gamma^*)$.

Assume that μ and α are bounded above by universal constants and let the function F be fixed, so that both L_α and β_α in (14) are bounded. Also notice that $\beta_{(1-\gamma^*)\alpha} \geq \beta_{\alpha/2}$ since $\gamma^* \leq 1/2$. Then comparing the lower bound (24) with the upper bound (18) shows that if the sample size $n \geq \frac{R^2 \beta_{\alpha/2}}{4\alpha^4} (d_1 + d_2)$, the optimal rate of convergence is $R \sqrt{\frac{d_1 + d_2}{n}}$:

$$\inf_{\hat{M}} \sup_{M \in K_{\max}(\alpha, R)} \frac{1}{d_1 d_2} \mathbb{E} \|\hat{M} - M\|_F^2 \asymp R \sqrt{\frac{d_1 + d_2}{n}},$$

and the max-norm constrained maximum likelihood estimate (13) is rate-optimal. If the target matrix M^* is known to have rank at most r , we can take $R = \alpha \sqrt{r}$, such that the requirement here on the sample size $n \geq \frac{\beta_{\alpha/2}}{4\alpha^2} r (d_1 + d_2)$ is weak and the optimal rate of convergence becomes $\alpha \sqrt{\frac{r(d_1 + d_2)}{n}}$.

3.5 Comparison to Prior Work

In this paper, we study a matrix completion model proposed in Davenport et al. (2012), in which it is assumed that a binary matrix is observed at random from a distribution parameterized by an unknown matrix which is (approximately) low-rank. It is noteworthy that some earlier papers on collaborative filtering or matrix completion, including Srebro et al. (2005) and references therein, also dealt with binary observations that are assumed to be noisy versions of the underlying matrix, in Logistic or Bernoulli conditional model. The goal there is to predict directly the quantized values, or equivalently, to reconstruct the sign matrix, instead of the underlying real values, therefore the non-identifiability issue could be avoided.

We next turn to a detailed comparison of our results for 1-bit matrix completion to those obtained in Davenport et al. (2012), also for approximately low-rank matrices. Using the trace-norm as

a proxy to rank, Davenport et al. (2012) have studied 1-bit matrix completion under the *uniform sampling distribution* over the parameter space $K_*(\alpha, r)$ as given in (11), for some $\alpha > 0$ and $r \leq \min\{d_1, d_2\}$ is a positive integer. To recover the unknown $M^* \in K_*(\alpha, r)$, given a collection of observations Y_S where S follows a Bernoulli model, that is, every entry $(k, l) \in [d_1] \times [d_2]$ is observed independently with equal probability $\frac{n}{d_1 d_2}$, they propose the following trace-norm constrained MLE

$$\widehat{M}_{\text{tr}} = \underset{M \in K_*(\alpha, r)}{\operatorname{argmin}} \ell_S(M; Y)$$

and prove that for a sample size $n \geq d \log(d)$, $d = d_1 + d_2$, with high probability,

$$\frac{1}{d_1 d_2} \|\widehat{M}_{\text{tr}} - M^*\|_F^2 \lesssim \beta_\alpha L_\alpha \alpha \sqrt{\frac{rd}{n}}. \tag{25}$$

Comparing to (18) with $R = \alpha\sqrt{r}$, it is easy to see that under the uniform sampling model, the error bounds in (rescaled) Frobenius norm for the two estimates \widehat{M}_{max} and \widehat{M}_{tr} are of the same order. Moreover, Theorem 3 in Davenport et al. (2012) and Theorem 5, respectively, provide lower bounds showing that both \widehat{M}_{tr} and \widehat{M}_{max} achieve the minimax rate of convergence for recovering approximately low-rank matrices over the parameter spaces $K_*(\alpha, r)$ and $K_{\text{max}}(\alpha, R)$ respectively.

As mentioned in the introduction, the uniform sampling distribution assumption is restrictive and not valid in many applications including the well-known Netflix problem. When the sampling distribution is non-uniform, it was shown in Salakhutdinov and Srebro (2010) that the standard trace-norm regularized method might fail, specifically in the setting where the row and column marginal distributions are such that certain rows or columns are sampled with very high probabilities. Moreover, it was proposed to use a weighted variant of the trace-norm, which incorporates the knowledge of the true sampling distribution in its construction, and showed experimentally that this variant indeed leads to superior performance. Using this weighted trace-norm, Negahban and Wainwright (2012) provided theoretical guarantees on approximate low-rank matrix completion in general sampling case while assuming that each row and column is sampled with positive probability (see condition (17)). In addition, requiring that the probabilities to observe an element from any row or column are of order $O((d_1 \wedge d_2)^{-1})$, Klopp (2012) analyzed the performance of the trace-norm penalized estimators, and provided near-optimal (up to a logarithmic factor) bounds which are similar to the bounds in this paper.

Next we provide an analysis of the performance of the weighted trace-norm in 1-bit matrix completion. Given the knowledge of the true sampling distribution, we establish an upper bound on the error in recovering M^* , which comparing to (25), includes an additional $\log^{1/2}(d)$ factor. We do not rule out the possibility that this logarithmic factor might be an artifact of the technical tools used in proof described below. The proof in Davenport et al. (2012) for the trace-norm regularization in uniform sampling case may also be extended to the weighted trace-norm method under the general sampling model, by using the matrix Bernstein inequality instead of Seginer’s theorem. The extra logarithmic factor, however, is still inevitable based on this argument. We will not pursue the details in this paper.

Given a sampling distribution $\Pi = \{\pi_{kl}\}$ on $[d_1] \times [d_2]$, define its row- and column-marginals as

$$\pi_k = \sum_{l=1}^{d_2} \pi_{kl} \quad \text{and} \quad \pi_l = \sum_{k=1}^{d_1} \pi_{kl},$$

respectively. Under the condition (17), we have

$$\pi_{k\cdot} \geq \frac{1}{\mu d_1}, \quad \pi_{\cdot l} \geq \frac{1}{\mu d_2}, \quad \text{for all } (k, l) \in [d_1] \times [d_2]. \quad (26)$$

As in Salakhutdinov and Srebro (2010), consider the following weighted trace-norm with respect to the distribution Π :

$$\|M\|_{w,*} := \|M_w\|_* = \left\| \text{diag}(\sqrt{\pi_{1\cdot}}, \dots, \sqrt{\pi_{d_1\cdot}}) \cdot M \cdot \text{diag}(\sqrt{\pi_{\cdot 1}}, \dots, \sqrt{\pi_{\cdot d_2}}) \right\|_*, \quad (27)$$

where $(M_w)_{k,l} := \sqrt{\pi_{k\cdot} \pi_{\cdot l}} M_{k,l}$. Notice that if M has rank at most r and $\|M\|_\infty \leq \alpha$, then

$$\|M\|_{w,*} \leq \sqrt{r} \|M\|_F = \sqrt{r} \left(\sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \pi_{k\cdot} \pi_{\cdot l} M_{k,l}^2 \right)^{1/2} \leq \alpha \sqrt{r}.$$

Analogously to the previous studied class $K_*(\alpha, r)$, as given in (11), containing the low trace-norm matrices, define

$$K_{\Pi,*} \equiv K_{\Pi,*}(r, \alpha) = \left\{ M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_{w,*} \leq \alpha \sqrt{r}, \|M\|_\infty \leq \alpha \right\}$$

and consider estimating the unknown $M^* \in K_{\Pi,*}$ by solving the following optimization problem:

$$\widehat{M}_{w,tr} = \arg \min_{M \in K_{\Pi,*}} \ell_S(M; Y). \quad (28)$$

The following theorem states that the weighted trace-norm regularized approach can be nearly as good as the max-norm regularized estimator (up to logarithmic and constant factors), under a general sampling distribution that is not too far from uniform. The theoretical performance of the weighted trace-norm is first studied by Foygel et al. (2011) in the standard matrix completion problems under arbitrary sampling distributions.

Theorem 7 *Suppose that Condition U holds but with $M^* \in K_{\Pi,*}$, and assume that the training set S follows a general weighted sampling model according to the distribution Π satisfying (17). Then there exists an absolute constant $C > 0$ such that, for a sample size $n \geq \mu \min\{d_1, d_2\} \log(d)$ and any $\delta > 0$, the minimizer $\widehat{M}_{w,tr}$ of the optimization program (28) satisfies*

$$\|\widehat{M}_{w,tr} - M^*\|_{\Pi}^2 \leq C \beta_\alpha \left\{ L_\alpha \alpha \sqrt{\frac{\mu r d \log(d)}{n}} + U_\alpha \sqrt{\frac{\log(4/\delta)}{n}} \right\}, \quad (29)$$

with probability at least $1 - \delta$.

Since the construction of weighted trace-norm $\|\cdot\|_{w,*}$ highly depends on the underlying sampling distribution which is typically unknown in practice, the constraint $M^* \in K_{\Pi,*}$ seems to be artificial. The max-norm constrained approach, on the contrary, does not require the knowledge of the exact sampling distribution and the error bound in weighted Frobenius norm, as shown in (16), holds even without prior assumption on Π , for example, condition (17). Moreover, to ensure that the weighted trace-norm regularized method performs well, it is necessary that the marginals are not too small or equivalently that

$$\mu = \max \left\{ \frac{1}{d_1 \pi_{k\cdot}} \vee \frac{1}{d_2 \pi_{\cdot l}} : k, l \in [d_1] \times [d_2] \right\}$$

is not too large. Otherwise, both the error bounds in (29) and the sample complexity $\mu \min\{d_1, d_2\} \log(d)$ would grow larger with μ when the marginals were far from uniform. We conjecture that the factor μ would also appear in the results that are extended from those in Davenport et al. (2012). As evident in Theorem 3, using the max-norm based regularization does not lead to a deterioration in either the error bounds or the sample complexity when the sampling distribution was far from uniform.

To clarify the major difference between the principles behind (25) and (29), we remark that one of the key technical tools used in Davenport et al. (2012) is a bound of Seginer (2000) on the spectral norm of a random matrix with i.i.d. zero mean entries (corresponding to the uniform sampling distribution), that is, for any $h \leq 2 \log(\max\{d_1, d_2\})$,

$$\mathbb{E}[\|A\|^h] \leq K^h \left(\mathbb{E} \left[\max_{k=1, \dots, d_1} \|a_{k \cdot}\|_2^h \right] + \mathbb{E} \left[\max_{j=1, \dots, d_2} \|a_{\cdot j}\|_2^h \right] \right),$$

where $a_{k \cdot}$ (resp. $a_{\cdot j}$) denote the rows (resp. columns) of A and K is a universal constant. Under the non-uniform sampling model, we will deal with a matrix with independent entries that are not necessarily identically distributed, to which case an alternative result of Latala (2005) can be applied, that is,

$$\mathbb{E}[\|A\|] \leq K' \left(\max_{k=1, \dots, d_1} \mathbb{E} \|a_{k \cdot}\|_2 + \max_{j=1, \dots, d_2} \mathbb{E} \|a_{\cdot j}\|_2 + \left(\sum_{k,l} \mathbb{E} a_{kl}^4 \right)^{1/4} \right),$$

or instead, resorting to the matrix Bernstein inequality. Using either inequality would thus bring an additional logarithmic factor, appeared in (29).

It is also worth noticing that though the sampling distribution is not known exactly in practice, its empirical analogues are expected to be stable enough as an alternative. According to Foygel et al. (2011), given a random sample $S = \{(i_t, j_t)\}_{t=1}^n$, consider the empirical marginals

$$\hat{\pi}^r(i) = \frac{\#\{t : i_t = i\}}{n}, \quad \hat{\pi}^c(j) = \frac{\#\{t : j_t = j\}}{n} \quad \text{and} \quad \hat{\pi}_{ij} = \hat{\pi}^r(i) \hat{\pi}^c(j),$$

as well as the smoothed empirical marginals

$$\check{\pi}^r(i) = \frac{1}{2} (\hat{\pi}^r(i) + 1/d_1), \quad \check{\pi}^c(j) = \frac{1}{2} (\hat{\pi}^c(j) + 1/d_2) \quad \text{and} \quad \check{\pi}_{ij} = \check{\pi}^r(i) \check{\pi}^c(j).$$

The smoothed empirically-weighted trace-norm $\|\cdot\|_{\check{\pi},*}$ can be defined in the same spirit as in the definition (27) of weighted trace-norm, only with $\{\pi_{ij}\}$ replaced by $\{\check{\pi}_{ij}\}$. Then the unknown matrix can be estimated via regularization on the $\check{\pi}$ -weighted trace-norm, that is,

$$\check{M}_{\check{\pi},tr} = \arg \min \{ \ell_S(M; Y) : \|M\|_\infty \leq \alpha, \|M\|_{\check{\pi},*} \leq \alpha \sqrt{r} \}.$$

Adopting Theorem 4 in Foygel et al. (2011) to the current 1-bit problem will lead to a learning guarantee similar to (29).

4. Computational Algorithm

Problems of the form (13) can now be solved using a variety of algorithms, including interior point method (Srebro et al., 2005), Frank-Wolfe-type algorithm (Jaggi, 2013) and projected gradient

method (Lee et al., 2010). The first two are convex methods with guaranteed convergence rates to the global optimum, though can be slow in practice and might not scale to matrices with hundreds of rows or columns. We describe in this section a simple first order method due to Lee et al. (2010), which is a special case of a projected gradient algorithm for solving large-scale convex programs involving the max-norm. This method is non-convex, but as long as the size of the problem is large enough, it is guaranteed that each local minimum is also a global optimum, due to Burer and Monteiro (2003).

We start from rewriting the original problem as an optimization over factorizations of a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ into two terms $M = UV^T$, where $U \in \mathbb{R}^{d_1 \times k}$ and $V \in \mathbb{R}^{d_2 \times k}$ for some $1 \leq k \leq d = d_1 + d_2$. More specifically, for any $1 \leq k \leq d$ fixed, define

$$\mathcal{M}_k(R) := \left\{ UV^T : U \in \mathbb{R}^{d_1 \times k}, V \in \mathbb{R}^{d_2 \times k}, \max\{\|U\|_{2,\infty}^2, \|V\|_{2,\infty}^2\} \leq R \right\}.$$

Then the global optimum of (13) is equal to that of

$$\begin{aligned} & \text{minimize} && \ell(M; Y) \\ & \text{subject to} && M \in \mathcal{M}_k(R), \quad \|M\|_\infty \leq \alpha. \end{aligned} \tag{30}$$

Here we write $\ell(M; Y) = \frac{1}{|S|} \ell_S(M; Y)$ for brevity. This problem is non-convex, come with no guaranteed convergence rates to the global optimum. A surprising fact is that when $k \geq 1$ is large enough, this problem has no local minimum (Burer and Monteiro, 2003). Notice that $\ell(\cdot; Y)$ is differentiable with respect to the first argument, then (30) can be solved iteratively via the following updates:

$$\begin{bmatrix} U(\tau) \\ V(\tau) \end{bmatrix} = \begin{bmatrix} U^t - \frac{\tau}{\sqrt{t}} \cdot \nabla f(U^t (V^t)^T; Y) V^t \\ V^t - \frac{\tau}{\sqrt{t}} \cdot \nabla f(U^t (V^t)^T; Y)^T U^t \end{bmatrix},$$

where $\tau > 0$ is a stepsize parameter and $t = 0, 1, 2, \dots$. Next, we project $(U(\tau), V(\tau))$ onto $\mathcal{M}_k(R)$ according to

$$\begin{bmatrix} \tilde{U}^{t+1} \\ \tilde{V}^{t+1} \end{bmatrix} = \mathcal{P}_R \left(\begin{bmatrix} U(\tau) \\ V(\tau) \end{bmatrix} \right).$$

This orthogonal projection can be computed by re-scaling the rows of the current iterate whose ℓ_2 -norms exceed R so that their norms become exactly R , while rows with norms already less than R remain unchanged. If $\|\tilde{U}^{t+1} (\tilde{V}^{t+1})^T\|_\infty > \alpha$, we replace

$$\begin{bmatrix} \tilde{U}^{t+1} \\ \tilde{V}^{t+1} \end{bmatrix} \quad \text{with} \quad \frac{\sqrt{\alpha}}{\|\tilde{U}^{t+1} (\tilde{V}^{t+1})^T\|_\infty^{1/2}} \begin{bmatrix} \tilde{U}^{t+1} \\ \tilde{V}^{t+1} \end{bmatrix},$$

otherwise we keep it still. The resulting update is then denoted by (U^{t+1}, V^{t+1}) .

It is important to note that the choice of k must be large enough, at least as big as the rank of M^* . Suppose that, before solving (13), we know that the target matrix M^* has rank at most r^* . Then it is best to solve (30) for $k = r^* + 1$ in the sense that, if we choose $k \leq r^*$, then (30) is not equivalent to (13), and if we take $k > r^* + 1$, then we would be solving a larger program than necessary. In practice, we do not know the exact value of r^* in advance. Nevertheless, motivated by Burer and Monteiro (2003), we suggest the following scheme to solve the problem which avoids solving (30) for $r \gg r^*$:

- (1) Choose an initial small k and compute a local minimum (U, V) of (30), using above projected gradient method.
- (2) Use an optimization technique to determine whether the injections \widehat{U} of U into $\mathbb{R}^{d_1 \times (k+1)}$ and \widehat{V} of V into $\mathbb{R}^{d_2 \times (k+1)}$ comprise a local minimum of (30) with the size increased to $k + 1$.
- (3) If $(\widehat{U}, \widehat{V})$ is a local minimum, then we can take $M = UV^T$ as the final solution; otherwise compute a better local minimum $(\widetilde{U}, \widetilde{V})$ of (30) with size $k + 1$ and repeat step (2) with $(U, V) = (\widetilde{U}, \widetilde{V})$ and $k = k + 1$.

It was also suggested in Lee et al. (2010) that when dealing with extremely large data sets with S consisting of hundreds of millions of index pairs, one may consider using a stochastic gradient method based on the following decomposition for ℓ , that is,

$$\begin{aligned} \ell(UV^T; Y) &= \frac{1}{|S|} \sum_{(i,j) \in S} g(u_i^T v_j; Y_{i,j}) \quad \text{with} \\ g(t; y) &= \mathbf{1}_{\{y=1\}} \log \left(\frac{1}{F(t)} \right) + \mathbf{1}_{\{y=-1\}} \log \left(\frac{1}{1-F(t)} \right), \end{aligned}$$

where $S \subset [d_1] \times [d_2]$ is a training set of row-column indices, u_i and v_j denote the i -th row of U and j -th row of V , respectively. The stochastic gradient method says that at t -th iteration, we only need to pick one training pair (i_t, j_t) at random from S , then update $g(u_{i_t}^T v_{j_t}; Y_{i_t, j_t})$ via the previous procedure. More precisely, if $\|u_{i_t}\|_2^2 > R$, we project it back so that $\|u_{i_t}\|_2^2 = R$, otherwise we do not make any change (do the same for v_{j_t}). Next, if $|u_{i_t}^T v_{j_t}| > \alpha$, replace u_{i_t} and v_{j_t} with $\sqrt{\alpha} u_{i_t} / |u_{i_t}^T v_{j_t}|^{1/2}$ and $\sqrt{\alpha} v_{j_t} / |u_{i_t}^T v_{j_t}|^{1/2}$ respectively, otherwise we keep everything still. At the t -th iteration, we do not need to consider any other rows of U and V . This simple algorithm could be computationally as efficient as optimization with the trace-norm.

5. Numerical Results

In this section, we report the simulation results for low-rank matrix recovery based on 1-bit observations. In all cases presented below, we solved the convex program (30) by using our implementation in MATLAB of the projected gradient algorithm proposed in Section 4 for a wide range of values of the step-size parameter τ .

We first consider a rank-2, $d \times d$ target matrix M^* with eigenvalues $\{d/\sqrt{2}, d/\sqrt{2}, 0, \dots, 0\}$, so that $\|M^*\|_F/d = 1$. We choose to work with the Gaussian conditional model under uniform sampling. Let Y_S be the noisy binary observations with $S = \{(i_1, j_1), \dots, (i_t, j_t)\}$, that is, for $(i, j) \in S$,

$$Y_{i,j} = \begin{cases} +1, & \text{with probability } \Phi(M_{i,j}^*/\sigma), \\ -1, & \text{with probability } 1 - \Phi(M_{i,j}^*/\sigma), \end{cases}$$

and the objective function is given by

$$\ell_S(M; Y) = \frac{1}{|S|} \left\{ \sum_{(i,j) \in \Omega^+} \log \left[\frac{1}{\Phi(M_{i,j}/\sigma)} \right] + \sum_{(i,j) \in \Omega^-} \log \left[\frac{1}{1 - \Phi(M_{i,j}/\sigma)} \right] \right\},$$

where $\Omega^+ = \{(i, j) \in S : Y_{i,j} = 1\}$ and $\Omega^- = \{(i, j) \in S : Y_{i,j} = -1\}$. In Figure 1, averaging the results over 20 repetitions, we plot the squared Frobenius norm of the error (normalized by the

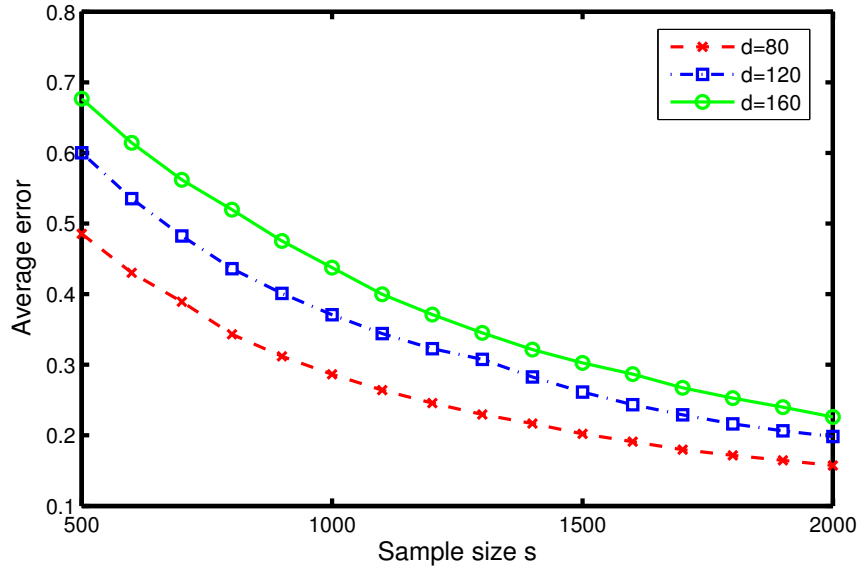


Figure 1: Plot of the average Frobenius error $\|\widehat{M} - M^*\|_F^2/d^2$ versus the sample size s for three different matrix sizes $d \in \{80, 120, 160\}$, all with rank $r = 2$.

dimension) $\|\widehat{M} - M^*\|_F^2/d^2$ versus a range of sample sizes $s = |S|$, with the noise level σ taken to be $\alpha/2$, for three different matrix sizes, $d \in \{80, 120, 160\}$. Naturally, in each case, the Frobenius error decays as s increases, although larger matrices require larger sample sizes, as reflected by the upward shift of the curves as d is increased.

Next, we compare the performance of the max-norm based regularization with that of the trace-norm using the same criterion as in Davenport et al. (2012). More specifically, the target matrix M^* is constructed at random by generating $M = LR^T$, where L and R are $d \times r$ matrices with i.i.d. entries drawn from Uniform $[-1/2, 1/2]$, so that $\text{rank}(M^*) = r$. It is then scaled such that $\|M^*\|_\infty = 1$, while in the last case, M^* is formed such that $\|M^*\|_F/d = 1$. As before, we focus on the Gaussian conditional model but with noise level σ varies from 10^{-3} to 10, and set $d = 500$, $r = 1$ and $s = 0.15d^2$, which is exactly the same case studied in Davenport et al. (2012). We plot in Figure 2 the squared Frobenius norm of the error (normalized by the norm of the underlying matrix M^*) over a range of different values of noise level σ on a logarithmic scale. As evident in Figure 2, the max-norm based regularization performs slightly but consistently better than the trace-norm, except on the one point where $\sigma = \log_{10}(0.25)$. Also, we see that for both methods, the performance is poor when the noise is either too little or too much.

In the third experiment, we consider matrices with dimension $d = 200$ and choose a moderate level of noise, that is, $\sigma = \log_{10}(-0.75)$, according to previous experiences. Figure 3 plots the relative Frobenius norm of the error versus the sample size s for three different matrix ranks, $r \in \{3, 5, 10\}$. Indeed, larger rank means larger intrinsic dimension of the problem, and thus increases the difficulty of any reconstruction procedure.

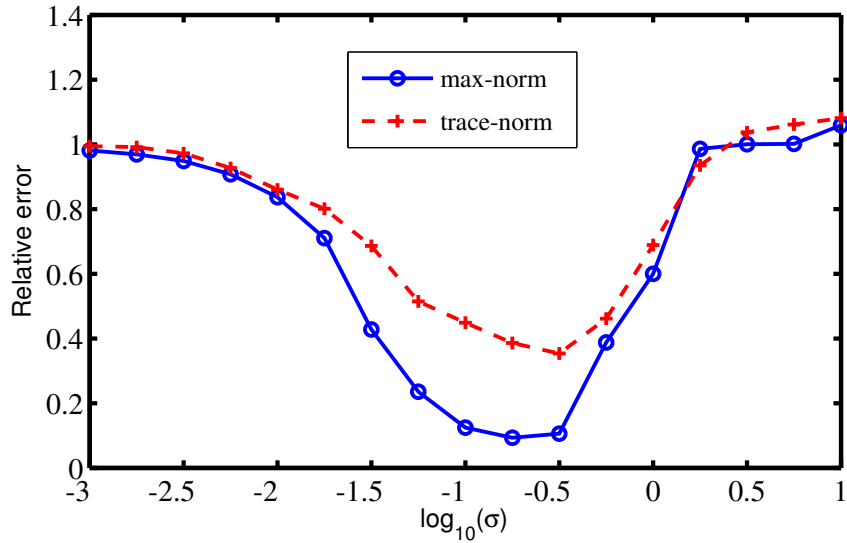


Figure 2: Plot of the relative Frobenius error $\|\widehat{M} - M^*\|_F^2 / \|M^*\|_F^2$ versus the noise level σ on a logarithmic scale, with rank $r = 1$, using both max-norm and trace-norm constrained methods.

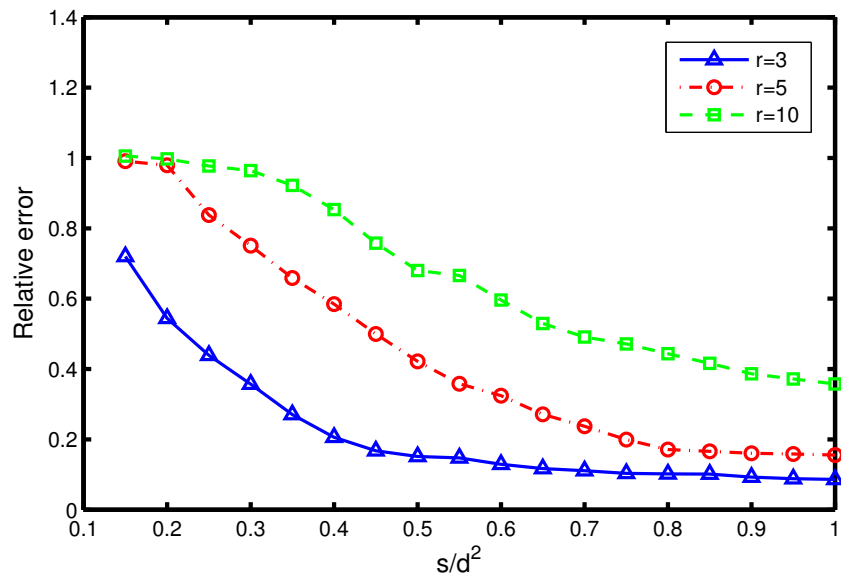


Figure 3: Plot of the relative Frobenius error versus the rescaled sample size s/d^2 for three different ranks $r \in \{3, 5, 10\}$, all with matrix size $d = 200$.

6. Discussion

This paper studies the problem of recovering a low-rank matrix based on highly quantized (to a single bit) noisy observation of a subset of entries. The problem was first formulated and analyzed by Davenport et al. (2012), where the authors consider approximately low-rank matrices in terms that the singular values belong to a scaled Schatten-1 ball. When the infinity norm of the unknown matrix M^* is bounded by a constant and its entries are observed uniformly in random, they show that M^* can be recovered from binary measurements accurately and efficiently.

Our theory, on the other hand, focuses on approximately low-rank matrices in the sense that unknown matrix belongs to certain max-norm ball. The unit max-norm ball is nearly the convex hull of rank-1 matrices whose entries are bounded in magnitude by 1, thus is a natural convex relaxation of low-rank matrices, particularly with bounded infinity norm. Allowing for non-uniform sampling, we show that the max-norm constrained maximum likelihood estimation is rate-optimal up to a constant factor, and that the corresponding convex program may be solved efficiently in polynomial time. An interesting question naturally arises that whether it is possible to push the theory further to cover exact low-rank matrix completion from noisy binary measurements.

The numerical study in Section 5 provides some evidence of the efficiency of the max-norm constraint approach in 1-bit matrix completion problem. More extensive experimental studies, applications to real data, and numerical comparisons with empirically weighted trace-norm method in a non-uniform scenario will be left as future work.

In our previous work (Cai and Zhou, 2013), we suggest to use max-norm constrained least square estimation to study standard matrix completion (from observations where additive noise is present) under a general sampling model. Similar error bounds are obtained, which are tight to within a constant. Comparing both results in the case of Gaussian noise demonstrates that as long as the signal-to-noise ratio remains constant, almost nothing is lost by quantizing to a single bit.

7. Proofs

We provide the proofs of the main results in this section.

7.1 Proof of Theorem 3

The proof of Theorem 3 is based on general excess risk bounds developed in Bartlett and Mendelson (2002) for empirical risk minimization when the loss function is Lipschitz. We regard matrix recovery as a prediction problem, that is, consider a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ as a function: $[d_1] \times [d_2] \rightarrow \mathbb{R}$, that is, $M(k, l) = M_{k,l}$. Moreover, define a function $g(x; y) \mathbb{R} \times \{\pm 1\} \mapsto \mathbb{R}$, which can be seen as a loss function:

$$g(x; y) = \mathbf{1}_{\{y=1\}} \log \left(\frac{1}{F(x)} \right) + \mathbf{1}_{\{y=-1\}} \log \left(\frac{1}{1-F(x)} \right).$$

For a subset $S = \{(i_1, j_1), \dots, (i_n, j_n)\} \subseteq ([d_1] \times [d_2])^n$ of the observed entries of Y , let $\mathcal{D}_S(M; Y) = \frac{1}{n} \sum_{i=1}^n g(M_{i,j_i}; Y_{i,j_i}) = \frac{1}{n} \ell_S(M; Y)$ be the average empirical likelihood function, where the training set S is drawn i.i.d. according to Π (with replacement) on $[d_1] \times [d_2]$. Then we have

$$\mathcal{D}_\Pi(M; Y) := \mathbb{E}_{S \sim \Pi} [g(M_{i,j_i}; Y_{i,j_i})] = \sum_{(k,l) \in [d_1] \times [d_2]} \pi_{kl} \cdot g(M_{k,l}; Y_{k,l}).$$

Under condition (U3), we can consider g as a function: $[-\alpha, \alpha] \times \{\pm 1\} \rightarrow \mathbb{R}$, such that for any $y \in \{\pm 1\}$ fixed, $g(\cdot; y)$ is essentially an L_α -Lipschitz loss function. Also notice that in the current case, $Y_{i,j}$ take ± 1 values and appear only in indicator functions, $\mathbf{1}\{Y_{i,j} = 1\}$ and $\mathbf{1}\{Y_{i,j} = -1\}$. Therefore, a combination of Theorem 8, (4) of Theorem 12 from Bartlett and Mendelson (2002) as well as the upper bound (6) on the Rademacher complexity of the unit max-norm ball yields that, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over choosing a training set S of $2 < n \leq d_1 d_2$ index pairs according to Π :

$$\begin{aligned} & \sup_{M \in K_{\max}(\alpha, R)} (\mathbb{E}_Y \mathcal{D}_\Pi(M; Y) - \mathbb{E}_Y \mathcal{D}_S(M; Y)) \\ & \leq 17L_\alpha R \sqrt{\frac{d}{n}} + U_\alpha \sqrt{\frac{8 \log(2/\delta)}{n}} := R_n(\alpha, r; \delta). \end{aligned} \quad (31)$$

Since \widehat{M}_{\max} is optimal and M^* is feasible to the optimization problem (13), we have

$$\mathcal{D}_S(\widehat{M}_{\max}; Y) \leq \mathcal{D}_S(M^*; Y) = \frac{1}{n} \sum_{t=1}^n g(M_{i_t, j_t}^*; Y_{i_t, j_t}).$$

Because M^* has a fixed value which does not depend on S , the empirical likelihood term $\mathcal{D}_S(M^*; Y)$ is an unbiased estimator of $\mathcal{D}_\Pi(M^*; Y)$, that is,

$$\mathbb{E}_{S \sim \Pi}[\mathcal{D}_S(M^*; Y)] = \mathcal{D}_\Pi(M^*; Y).$$

Next, we will derive an upper bound on the deviation $\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y)$ that holds with high probability. To do this, let A_1, \dots, A_n be independent random variables taking values in $[d_1] \times [d_2]$ according to Π , that is, $\mathbb{P}[A_t = (k, l)] = \pi_{kl}$, $t = 1, \dots, n$, such that $\mathcal{D}_S(M^*; Y) = \frac{1}{n} \sum_{t=1}^n g(M_{A_t}^*; Y_{A_t})$ and

$$\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y) = \frac{1}{n} \sum_{t=1}^n (g(M_{A_t}^*; Y_{A_t}) - \mathbb{E}[g(M_{A_t}^*; Y_{A_t})]).$$

Then we apply the Hoeffding's inequality to the random variables $Z_{A_t} := g(M_{A_t}^*; Y_{A_t}) - \mathbb{E}[g(M_{A_t}^*; Y_{A_t})]$, conditionally on Y . Observe that $0 \leq g(M_{A_t}^*; Y_{A_t}) \leq U_\alpha$ almost surely for all $1 \leq t \leq n$, therefore for any $u > 0$, we have

$$\mathbb{P}_{S \sim \Pi} \{ \mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y) > u \} \leq \exp\left(-\frac{2nu^2}{U_\alpha^2}\right), \quad (32)$$

which in turn implies that that with probability at least $1 - \delta$ over choosing a subset S according to Π ,

$$\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y) \leq U_\alpha \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (33)$$

Putting pieces together, we get

$$\begin{aligned} & \mathbb{E}_Y [\mathcal{D}_\Pi(\widehat{M}_{\max}; Y) - \mathcal{D}_\Pi(M^*; Y)] \\ & = \mathbb{E}_Y [\mathcal{D}_\Pi(\widehat{M}_{\max}; Y) - \mathcal{D}_S(M^*; Y)] + \mathbb{E}_Y [\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y)] \\ & \leq \mathbb{E}_Y [\mathcal{D}_\Pi(\widehat{M}_{\max}; Y) - \mathcal{D}_S(\widehat{M}_{\max}; Y)] + \mathbb{E}_Y [\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y)] \\ & \leq \sup_{M \in K_{\max}(\alpha, R)} \{ \mathbb{E}_Y [\mathcal{D}_\Pi(M; Y)] - \mathbb{E}_Y [\mathcal{D}_S(M; Y)] \} \\ & \quad + \mathbb{E}_Y [\mathcal{D}_S(M^*; Y) - \mathcal{D}_\Pi(M^*; Y)]. \end{aligned} \quad (34)$$

Moreover, observe that the left-hand side of (34) is equal to

$$\begin{aligned} & \mathbb{E}_Y [\mathcal{D}_\Pi(\widehat{M}_{\max}; Y) - \mathcal{D}_\Pi(M^*; Y)] \\ &= \sum_{(k,l) \in [d_1] \times [d_2]} \pi_{kl} \left[F(M_{k,l}^*) \log \left(\frac{F(M_{k,l}^*)}{F((\widehat{M}_{\max})_{k,l})} \right) + (\bar{F}(M_{k,l}^*)) \log \left(\frac{\bar{F}(M_{k,l}^*)}{\bar{F}((\widehat{M}_{\max})_{k,l})} \right) \right], \end{aligned}$$

which is the weighted Kullback-Leibler divergence between matrices $F(M)$ and $F(\widehat{M}_{\max})$, denoted by $\mathbb{K}_\Pi(F(M) \| F(\widehat{M}_{\max}))$, where

$$\bar{F}(\cdot) := 1 - F(\cdot) \quad \text{and} \quad F(M) := (F(M_{k,l}))_{d_1 \times d_2}.$$

This, combined with (31), (33) and (34) imply that for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over S :

$$\mathbb{K}_\Pi(F(M^*) \| F(\widehat{M}_{\max})) \leq R_n(\alpha, r; \delta/2) + U_\alpha \sqrt{\frac{\log(2/\delta)}{2n}}. \quad (35)$$

Together, (8), (35) and Lemma 8 below establish (16).

Lemma 8 (Lemma 2, Davenport et al., 2012) *Let F be an arbitrary differentiable function, and s, t are two real numbers satisfying $|s|, |t| \leq \alpha$. Then*

$$d_H^2(F(s); F(t)) \geq \inf_{|x| \leq \alpha} \frac{(F'(x))^2}{8F(x)(1-F(x))} \cdot (s-t)^2$$

The proof of Theorem 3 is now completed. ■

7.2 Proof of Theorem 5

The proof for the lower bound follows an information-theoretic method based on Fano's inequality (Cover and Thomas, 1991), as used in the proof of Theorem 3 in Davenport et al. (2012). To begin with, we have the following lemma which guarantees the existence of a suitably large packing set for $K_{\max}(\alpha, R)$ in the Frobenius norm. The proof follows from Lemma 3 of Davenport et al. (2012) with a simple modification, see, for example, the proof of Lemma 3.1 in Cai and Zhou (2013).

Lemma 9 *Let $r = (R/\alpha)^2$ and $\gamma \leq 1$ be such that $r \leq \gamma^2 \min(d_1, d_2)$ is an integer. There exists a subset $\mathcal{S}(\alpha, \gamma) \subseteq K_{\max}(\alpha, R)$ with cardinality*

$$|\mathcal{S}(\alpha, \gamma)| = \left\lceil \exp \left(\frac{r \max(d_1, d_2)}{16\gamma^2} \right) \right\rceil + 1$$

and with the following properties:

- (i) For any $N \in \mathcal{S}(\alpha, \gamma)$, $\text{rank}(N) \leq \frac{r}{\gamma^2}$ and $N_{k,l} \in \{\pm \gamma\alpha/2\}$, such that

$$\|N\|_\infty = \frac{\gamma\alpha}{2}, \quad \frac{1}{d_1 d_2} \|N\|_F^2 = \frac{\gamma^2 \alpha^2}{4}.$$

(ii) For any two distinct $N^k, N^l \in \mathcal{S}(\alpha, \gamma)$,

$$\frac{1}{d_1 d_2} \|N^k - N^l\|_F^2 > \frac{\gamma^2 \alpha^2}{8}.$$

Then we construct the packing set \mathcal{M} by letting

$$\mathcal{M} = \left\{ N + \alpha(1 - \gamma/2)E_{d_1, d_2} : N \in \mathcal{S}(\alpha, \gamma) \right\}, \quad (36)$$

where $E_{d_1, d_2} \in \mathbb{R}^{d_1 \times d_2}$ is such that the $(d_1, d_2)^{th}$ entry equals one and others are zero. Clearly, $|\mathcal{M}| = |\mathcal{S}(\alpha, \gamma)|$. Moreover, for any $M \in \mathcal{M}$, $M_{k,l} \in \{\alpha, (1 - \gamma)\alpha\}$ by the construction of $\mathcal{S}(\alpha, \gamma)$ and (36), and

$$\|M\|_{\max} = \|N + \alpha(1 - \gamma/2)E_{d_1, d_2}\|_{\max} \leq \frac{\alpha\sqrt{r}}{2} + \alpha(1 - \gamma/2) \leq \alpha\sqrt{r},$$

provided that $r \geq 4$. Therefore, \mathcal{M} is indeed a δ -packing of $K_{\max}(\alpha, R)$ in the Frobenius metric with

$$\delta^2 = \frac{\alpha^2 \gamma^2 d_1 d_2}{8},$$

that is, for any two distinct $M, M' \in \mathcal{M}$, we have $\|M - M'\|_F \geq \delta$.

Next, a standard argument (Yang and Barron, 1999; Yu, 1997) yields a lower bound on the $\|\cdot\|_F$ -risk in terms of the error in a multi-way hypothesis testing problem. More concretely,

$$\inf_{\tilde{M}} \max_{M \in K_{\max}(\alpha, R)} \mathbb{E} \|\tilde{M} - M\|_F^2 \geq \frac{\delta^2}{4} \min_{\tilde{M}} \mathbb{P}(\tilde{M} \neq M^*),$$

where the random variable $M^* \in \mathbb{R}^{d_1 \times d_2}$ is uniformly distributed over the packing set \mathcal{M} , and the minimum is carried out over all estimators \tilde{M} taking values in \mathcal{M} . Applying Fano's inequality (Cover and Thomas, 1991) gives the lower bound

$$\mathbb{P}(\tilde{M} \neq M^*) \geq 1 - \frac{I(M^*; Y_S) + \log 2}{\log |\mathcal{M}|}, \quad (37)$$

where $I(M^*; Y_S)$ denotes the mutual information between the random parameter M^* in \mathcal{M} and the observation matrix Y_S . Following the proof of Theorem 3 in Davenport et al. (2012), we could bound $I(M^*; Y_S)$ as follows:

$$\begin{aligned} I(M^*; Y_S) &\leq \max_{M, M' \in \mathcal{M}, M \neq M'} \mathbb{K}(Y_S | M \| Y_S | M') \\ &= \max_{M, M' \in \mathcal{M}, M \neq M'} \sum_{(k,l) \in \mathcal{S}} \mathbb{K}(Y_{k,l} | M_{k,l} \| Y_{k,l} | M'_{k,l}) \\ &\leq \frac{n[F(\alpha) - F((1 - \gamma)\alpha)]^2}{F((1 - \gamma)\alpha)[1 - F((1 - \gamma)\alpha)]} \leq \frac{n\alpha^2\gamma^2}{\beta_{(1-\gamma)\alpha}}, \end{aligned}$$

where the last inequality holds provided that $F'(x)$ is decreasing on $(0, \infty)$. Substituting this into the Fano's inequality (37) yields

$$\mathbb{P}(\tilde{M} \neq M^*) \geq 1 - \left(\frac{n\alpha^2\gamma^2}{\beta_{(1-\gamma)\alpha}} + \log 2 \right) / \left(\frac{r(d_1 \vee d_2)}{16\gamma^2} \right)$$

Recall that $\gamma^* > 0$ solves the equation (22):

$$\gamma^* = \min \left\{ \frac{1}{2}, \frac{R^{1/2}}{\alpha} \left(\frac{\beta_{(1-\gamma^*)\alpha}(d_1 \vee d_2)}{32n} \right)^{1/4} \right\}.$$

Requiring

$$\frac{64 \log(2)(\gamma^*)^2}{d_1 \vee d_2} \leq r \leq (d_1 \wedge d_2)(\gamma^*)^2,$$

which is guaranteed by (23), to ensure that this probability is least 1/4. Consequently, we have

$$\inf_{\widehat{M}} \max_{M \in K_{\max}(\alpha, R)} \mathbb{E} \|\widehat{M} - M\|_F^2 \geq \frac{\alpha^2 (\gamma^*)^2 d_1 d_2}{128},$$

which in turn implies (24). ■

7.3 Proof of Theorem 7

The proof of Theorem 7 modifies the proof of Theorem 3, therefore we only summarize the key steps in the following. Let $\{A_1, \dots, A_n\} = \{(i_1, j_1), \dots, (i_n, j_n)\}$ be independent random variables taking values in $[d_1] \times [d_2]$ according to Π , and recall that

$$\ell_S(M; Y) = \sum_{t=1}^s \left[\mathbf{1}_{\{Y_{A_t}=1\}} \log \left(\frac{1}{F(M_{A_t})} \right) + \mathbf{1}_{\{Y_{A_t}=-1\}} \log \left(\frac{1}{1 - F(M_{A_t})} \right) \right].$$

According to Srebro et al. (2005) and the proof of Theorem 3, it suffices to derive an upper bound on

$$\Delta := \mathbb{E} \left[\sup_{M \in K_{\Pi, s}} \sum_{t=1}^n \frac{\varepsilon_t}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} (M_w)_{A_t} \right] = \mathbb{E} \left[\sup_{M \in K_s(\alpha, r)} \sum_{t=1}^n \frac{\varepsilon_t}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} M_{A_t} \right],$$

where ε_t are i.i.d. Rademacher random variables. Then it follows from (5) that

$$\begin{aligned} \Delta &\leq \alpha \sqrt{r} \cdot \mathbb{E} \left[\sup_{\|u\|_2 = \|v\|_2 = 1} \sum_{t=1}^n \frac{\varepsilon_t}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} u_{i_t} v_{j_t} \right] \\ &= \alpha \sqrt{r} \cdot \mathbb{E} \left[\sup_{\|u\|_2 = \|v\|_2 = 1} \sum_{i, j} \left(\sum_{t: (i_t, j_t) = (i, j)} \frac{\varepsilon_t}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} \right) u_i v_j \right] \\ &= \alpha \sqrt{r} \cdot \mathbb{E} \left[\left\| \sum_{t=1}^n \varepsilon_t \frac{e_{i_t} e_{j_t}^T}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} \right\| \right]. \end{aligned}$$

An upper bound on the above spectral norm has been derived in Foygel et al. (2011) using a recent result of Tropp (2012). Let $Q_t = \varepsilon_t \frac{e_{i_t} e_{j_t}^T}{\sqrt{\pi_{i_t} \cdot \pi_{j_t}}} \in \mathbb{R}^{d_1 \times d_2}$ be i.i.d. random matrices with zero-mean, then the problem reduces to estimate $\mathbb{E} \left\| \sum_{t=1}^s Q_t \right\|$. Following the the proof of Foygel et al. (2011), we see that, under condition (26)

$$\mathbb{E} \left\| \sum_{t=1}^n Q_t \right\| \leq C \left(\sigma_1 \sqrt{\log(d)} + \sigma_2 \log(d) \right)$$

with

$$\begin{aligned}\sigma_1 &= n \cdot \max \left\{ \max_k \sum_l \frac{\pi_{kl}}{\pi_k \pi_l}, \max_l \sum_k \frac{\pi_{kl}}{\pi_k \pi_l} \right\} \leq \mu n \max\{d_1, d_2\}, \\ \sigma_2 &= \max_{k,l} \frac{1}{\sqrt{\pi_k \pi_l}} \leq \mu \sqrt{d_1 d_2}.\end{aligned}$$

Putting above estimates together, we conclude that

$$\Delta \leq C\alpha\sqrt{r} \left(\sqrt{\mu n \max\{d_1, d_2\} \log(d)} + \mu \sqrt{d_1 d_2} \log(d) \right),$$

which in turn yields that for any $\delta \in (0, 1)$, inequality

$$\begin{aligned}\mathbb{K}_\Pi(F(M^*) \| F(\widehat{M}_{w,lr})) \\ \leq C \left\{ L_\alpha \alpha \sqrt{\frac{\mu r \max\{d_1, d_2\} \log(d)}{n}} + U_\alpha \sqrt{\frac{\log(4/\delta)}{n}} \right\}\end{aligned}$$

holds with probability at least $1 - \delta$, provided that $n \geq \mu \min\{d_1, d_2\} \log(d)$. \blacksquare

7.4 An Extension to Sampling Without Replacement

In this paper, we have focused on sampling with replacement. We shall show here that in the uniform sampling setting, the results obtained in this paper continue to hold if the (binary) entries are sampled without replacement. Recall that in the proof of Theorem 3, we let A_1, \dots, A_n be random variables taking values in $[d_1] \times [d_2]$, $S = \{A_1, \dots, A_n\}$ and assume the A_t 's are distributed uniformly and independently, that is, $S \sim \Pi = \{\pi_{kl}\}$ with $\pi_{kl} \equiv \frac{1}{d_1 d_2}$. The purpose now is to prove that the arguments remain valid when the A_t 's are selected without replacement, denoted by $S \sim \Pi_0$. In this notation, we have

$$\mathcal{D}_S = \frac{1}{n} \sum_{(i,j) \in S} g(M_{i,j}; Y_{i,j}) \quad \text{and} \quad \mathcal{D}_{\Pi_0} = \mathbb{E}_{S \sim \Pi_0}[\mathcal{D}_S] = \frac{1}{d_1 d_2} \sum_{(k,l)} g(M_{k,l}; Y_{k,l}).$$

By Lemma 3 in Foygel and Srebro (2011) and (31), for any $\delta > 0$,

$$\sup_{M \in \mathcal{K}_{\max}(\alpha, R)} \left(\mathbb{E}_Y \mathcal{D}_{\Pi_0}(M; Y) - \mathbb{E}_Y \mathcal{D}_S(M; Y) \right) \leq 17L_\alpha R \sqrt{\frac{d}{n}} + U_\alpha \sqrt{\frac{8(\log(4n) + \log(2/\delta))}{n}}$$

holds with probability at least $1 - \delta$ over choosing a training set S of $2 < n \leq d_1 d_2$ index pairs according to Π_0 . Next, observe that the large deviation bound (32) for the sum of independent bounded random variables is a direct consequence of Hoeffding's inequality. To see how inequality (32) may be extended to the current case, we start with a more general problem. Let \mathcal{C} be a finite set with cardinality N . For $1 \leq n \leq N$, let X_1, \dots, X_n be independent random variables taking values in \mathcal{C} uniformly at random, such that (X_1, \dots, X_n) is a \mathcal{C}^n -valued random vector modeling sampling with replacement from \mathcal{C} . On the other hand, let (Y_1, \dots, Y_n) be a \mathcal{C}^n -valued random vector sampled uniformly without replacement. Assume that X_i is centered and bounded, and write $S_X = \sum_{i=1}^n X_i$, $S_Y = \sum_{i=1}^n Y_i$. Then a large deviation bound holds for S_X by Hoeffding's inequality. In the proof, the tail probability is bounded from above in terms of the moment-generating function, say, $m_X(\lambda) =$

$\mathbb{E} \exp(\lambda S_X)$. According to the notion of negative association (Joag-Dev and Proschan, 1983), it is well-known that $m_Y(\lambda) = \mathbb{E} \exp(\lambda S_Y) \leq m_X(\lambda)$, which in turn gives a similar large deviation bound for S_Y . Therefore, inequalities (32) and (33) are still valid if Π is replaced by Π_0 . Keep all other arguments the same, we get the desired result.

Acknowledgments

The authors would like to thank Yaniv Plan for helpful discussions and for pointing out the importance of allowing non-uniform sampling. We also thank the Associate Editor and two referees for thorough and useful comments which have helped to improve the presentation of the paper. The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973, NSF Grant DMS-1208982, and NIH Grant R01 CA127334. A part of this work was done when the second author was visiting the Wharton Statistics Department of the University of Pennsylvania. He wishes to thank the institution and particularly Tony Cai for their hospitality.

References

- A. Ai, A. Lapanowski, Y. Plan and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and Its Applications*, forthcoming, 2013.
- P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, **3**:463–482, 2002.
- P. Biswas, T. C. Lian, T. C. Wang and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks*, **2**(2):188–220, 2006.
- P. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pages 16–21, Princeton, NJ, March 2008.
- S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, **95**(2):329–357, February 2003.
- T. T. Cai and W.-X. Zhou. Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*, 2013.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- M. A. Davenport, Y. Plan, E. van den Berg and M. Wooters. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2012.
- R. Foygel, R. Salakhudinov, O. Shamir and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in Neural Information Processing Systems (NIPS)*, **24**, pages 2133–2141, 2011.
- R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. *JMLR: Workshop and Conference Proceedings*, **19**:315–339, 2011.

- D. Goldberg, D. Nichols, B. M. Oki and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, **35**(12):61–70, December 1992.
- P. E. Green and Y. Wind. *Multiattribute Decisions in Marketing: A Measurement Approach*. Dryden Press, Hinsdale, IL, 1973.
- D. Gross and V. Neshveyev. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.
- L. Jacques, J. N. Laska, P. T. Boufounos and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *arXiv preprint arXiv:1104.3160*, 2011.
- M. Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization. *JMLR: Workshop and Conference Proceedings*, **28**(1):427–435, 2013.
- G. J. O. Jameson. *Summing and Nuclear Norms in Banach Space Theory*. London Mathematical Society Student Texts, 8. Cambridge University Press, Cambridge, 1987.
- K. Joag-Dev and F. Proschan. Negative association of random variables with applications. *Annals of Statistics*, **11**(1):286–295, 1983.
- O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, forthcoming, 2012.
- V. Koltchinskii, K. Lounici and A. B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39**(5):2302–2329, 2011.
- J. N. Laska and R. G. Baraniuk. Regime change: bit-depth versus measurement-rate in compressive sensing. *IEEE Transactions on Signal Processing*, **60**(7):3496–3505, 2012.
- R. Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, **133**(5):1273–1282, 2005.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes*. Springer-Verlag, Berlin, 2011.
- J. Lee, B. Recht, R. Salakhutdinov, N. Srebro and J. Tropp. Practical large-scale optimization for max-norm regularization. *Advances in Neural Information Processing Systems (NIPS)*, **23**, 2010.
- N. Linial, S. Mendelson, G. Schechtman and A. Shraibman. Complexity measures of sign measures. *Combinatorica*, **27**(4):439–463, 2007.
- S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *Journal of Machine Learning Research*, **13**:1665–1697, 2012.
- Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Transactions on Information Theory*, **59**(1):482–494, January, 2013a.
- Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, **66**(8):1275–1297, August, 2013b.

- R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: learning with the weighted trace norm. *Advances in Neural Information Processing Systems (NIPS)*, **23**, December, 2010.
- Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, **9**(2):149–166, March, 2000.
- I. Spence and D. W. Domoney. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika*, **39**(4):469–490, December, 1974.
- N. Srebro, J. Rennie and T. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems (NIPS)*, **17**, pages 1329–1336. MIT Press, 2005.
- N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Learning theory, Proceedings of COLT-2005. Lecture Notes in Computer Science*, **3559**:545–560, Springer, Berlin, 2005.
- N. Srebro, K. Sridharan and A. Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, **12**(4):389–434, August, 2012.
- Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, **27**(5):1564–1599, 1999.
- B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam*, pages 423–435.