Electronic Journal of Statistics
Vol. 10 (2016) 1493–1525
ISSN: 1935-7524
DOI: 10.1214/16-EJS1147

Matrix completion via max-norm constrained optimization

T. Tony Cai*

Department of Statistics, The Wharton School
University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: tcai@wharton.upenn.edu

and

Wen-Xin Zhou†

Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544, USA
e-mail: wenxinz@princeton.edu

Abstract: Matrix completion has been well studied under the uniform sampling model and the trace-norm regularized methods perform well both theoretically and numerically in such a setting. However, the uniform sampling model is unrealistic for a range of applications and the standard trace-norm relaxation can behave very poorly when the underlying sampling scheme is non-uniform.

In this paper we propose and analyze a max-norm constrained empirical risk minimization method for noisy matrix completion under a general sampling model. The optimal rate of convergence is established under the Frobenius norm loss in the context of approximately low-rank matrix reconstruction. It is shown that the max-norm constrained method is minimax rate-optimal and yields a unified and robust approximate recovery guarantee, with respect to the sampling distributions. The computational effectiveness of this method is also discussed, based on first-order algorithms for solving convex optimizations involving max-norm regularization.

MSC 2010 subject classifications: Primary 62H12, 62J99; secondary 15A83.
Keywords and phrases: Compressed sensing, low-rank matrix, matrix completion, max-norm constrained minimization, minimax optimality, non-uniform sampling, sparsity.

Received December 2015.

Contents

*Supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.
†Supported in part by NIH Grant R01-GM100474-5.

1. Introduction

The problem of recovering a low-rank matrix from a subset of its entries, also known as *matrix completion*, has been an active topic of recent research with a range of applications including collaborative filtering (the Netflix problem) (Goldberg et al., 1992), multi-task learning (Argyriou, Evgeniou and Pontil, 2008), system identification (Liu and Vandenberghe, 2009), and sensor localization (Singer and Cucuringu, 2010; Candés and Plan, 2010), among many others. We refer to Candés and Plan (2010) for detailed discussions of the aforementioned applications. Another noteworthy example is the structure-from-motion problem in computer vision (Tomasi and Kanade, 1992; Chen and Suter, 2004). Let f and d be the number of frames and feature points, respectively. The data are stacked into a low-rank matrix of trajectories, say $M \in \mathbb{R}^{2f\times d}$, such that every element of M corresponds to an image coordinate from a feature point of a rigid moving object at a given frame. Due to objects occlusions, errors on the tracking or variable out of range (i.e. images beyond the camera field of view), missing data are inevitable in real-life applications and are represented as empty entries in the matrix. Therefore, accurate and effective matrix completion methods, which fill in missing entries by suitable estimates, are required.

Because a direct search for the lowest-rank matrix satisfying the equality constraints is NP-hard, most previous work on matrix completion has focused on using the trace-norm, which is defined to be the sum of the singular values of the matrix, as a convex relaxation for the rank. This can be viewed as an

analogue to relaxing the *sparsity* of a vector to its ℓ_1-norm, which has been shown to be effective both empirically and theoretically in compressed sensing. Several recent papers proved in different settings that a generic $d \times d$ rank-r matrix can be exactly and efficiently recovered from $O\{rd\,\mathrm{poly}(\log d)\}$ randomly chosen entries (Candés and Recht, 2009; Candés and Tao, 2010; Gross, 2011; Recht, 2011). These results thus provide theoretical guarantees for the constrained trace-norm minimization method. In the case of recovering approximately low-rank matrices based on noisy observations, different types of trace-norm based estimators, which are akin to the Lasso and Dantzig selector used in sparse signal recovery, were proposed and well-studied. See, for example, Candés and Plan (2010), Keshavan and Montanari (2010), Rohde and Tsybakov (2011), Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2012), Koltchinskii (2011) and Klopp (2011, 2014), among others.

It is, however, unclear that whether the trace-norm is the best convex relaxation for the rank, especially when the underlying sampling scheme is non-uniform, and more importantly, is unknown. A matrix $M \in \mathbb{R}^{d_1 \times d_2}$ can be viewed as an operator mapping from $\mathbb{R}^{d_2}$ to $\mathbb{R}^{d_1}$, its rank can be alternatively expressed as the smallest integer k such that the matrix M can be decomposed as $M = UV^\intercal$ for some $U \in \mathbb{R}^{d_1 \times k}$ and $V \in \mathbb{R}^{d_2 \times k}$. In view of the matrix factorization $M = UV^\intercal$, by enforcing U and V to have a small number of columns we obtain a low-rank M. 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., 2004), defined by

$$\|M\|_{\max} = \min_{M=UV^\intercal} \|U\|_{2,\infty}\|V\|_{2,\infty}, \tag{1.1}$$

where the infimum is carried out over all factorizations $M = UV^\intercal$ with $\|U\|_{2,\infty}$ denoting the operator norm of $U : \ell_2^k \mapsto \ell_\infty^{d_1}$ and $\|V\|_{2,\infty}$ the operator norm of $V : \ell_2^k \mapsto \ell_\infty^{d_2}$ (or, equivalently, $V^\intercal : \ell_1^{d_2} \mapsto \ell_2^k$) and $k = 1, \ldots, \min(d_1, d_2)$. Note that $\|U\|_{2,\infty}$ is also the maximum ℓ_2 row norm of U. Since ℓ_2 is a Hilbert space, the factorization constant $\|\cdot\|_{\max}$ indeed defines a norm on the space of operators between $\ell_1^{d_2}$ and $\ell_\infty^{d_1}$.

The max-norm was recently proposed as an alternative convex surrogate to the rank of the matrix. For collaborative filtering problems, the max-norm has been shown to be empirically superior to the trace-norm Srebro, Rennie and Jaakkola (2004). Foygel and Srebro (2011) 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, Sridharan and Tewari, 2010). Further, a max-norm constrained maximum likelihood method was considered by Cai and Zhou (2013) for one-bit matrix completion, where instead of observing real-valued entries of an unknown matrix one is only able to see binary outputs, i.e. yes/no, true/false, agree/disagree (Davenport et al., 2014). Theoretical guarantees are obtained in general non-uniform sampling models, and numerical studies show that the max-norm based approach is comparable to and sometimes slightly outperform the corresponding trace-norm method.

Matrix completion has been well analyzed in the uniform sampling model, where observed entries are assumed to be sampled randomly and uniformly. In such a setting, the trace-norm regularized approach has been shown to have good theoretical and numerical performance. However, in some applications such as collaborative filtering, the uniform sampling model is unrealistic. For example, in the Netflix problem, the uniform sampling model is equivalent to assuming all users are equally likely to rate each movie and all movies are equally likely to be rated by any user. From a practical point of view, invariably some users are more active than others and some movies are more popular and thus rated more frequently. Hence, the sampling distribution is in fact non-uniform in the real world. In such a setting, Salakhutdinov and Srebro (2010) showed that the standard trace-norm relaxation can sometimes behave poorly, and suggested a weighted trace-norm penalty, which incorporates the knowledge of true sampling distribution in its construction. 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, a practically available method relies on the empirically-weighted trace-norm (Foygel et al., 2011). It is also worth noticing that, when the sampling probabilities are bounded from below and above, the trace-norm penalized estimator is minimax optimal up to a logarithmic factor (Klopp, 2014). We refer to Fang et al. (2015b) for further numerical evaluations of the trace-norm regularized method under various non-uniform sampling schemes.

In this paper, we employ the max-norm as a convex relaxation for the rank to study matrix completion based on noisy observations in a general, unspecified sampling model. The rate of convergence for the max-norm constrained least squares estimator is obtained. Information-theoretical methods are used to establish a matching minimax lower bound in the general non-uniform sampling model. Together, the minimax upper and lower bounds yield the optimal rate of convergence for the Frobenius norm loss. It is shown that the max-norm regularized approach indeed provides a unified and robust approximate recovery guarantee with respect to sampling schemes. In the uniform sampling model as a special case, our results also show that the extra logarithmic factors appeared in the error rates obtained by Srebro, Sridharan and Tewari (2010) and Foygel and Srebro (2011) could be avoided after a careful analysis to match the minimax lower bound with the upper bound (see Theorems 3.1 and 3.3 and the discussions in Section 3).

The max-norm constrained minimization problem is a convex program. To solve general convex programs that involve either a max-norm constraint or a max-norm penalization, a first-order algorithm was proposed by Lee et al. (2010), which is computationally effective and outperforms the semi-definite programming (SDP) method of Srebro, Rennie and Jaakkola (2004). In principle, the method of Lee et al. (2010) is based on nonconvex relaxations. Therefore, their algorithm is only guaranteed to find a stationary point, and statistical properties of such solutions are difficult to analyze. Recently, Fang et al. (2015b) proposed a scalable algorithm based on the alternating direction of multipliers method to efficiently solve the max-norm constrained optimiza-

tion problem with guaranteed rate of convergence to the global optimum. In summary, the max-norm constrained empirical risk minimization problem can indeed be implemented in polynomial time as a function of the sample size and matrix dimensions.

The remainder of the paper is organized as follows. After introducing basic notation and definitions, Section 2 collects a few useful results on the max-norm, trace-norm and Rademacher complexity that will be needed in the rest of the paper. Section 3 introduces the model and the estimation procedure and then investigates the theoretical properties of the estimator. Both minimax upper and lower bounds are given. The results show that the max-norm constrained minimization method achieves the optimal rate of convergence over the parameter space. Comparison with past work is also given. Computation and implementation issues are discussed in Section 4. A brief discussion is given in Section 5, and the proofs of the main results and key technical lemmas are placed in Section 6.

2. Notations and preliminaries

In this section, we begin with some notation that will be used throughout the paper, and then collect some known results on the max-norm, trace-norm and Rademacher complexity that will be applied repeatedly later.

For any positive integer d, we use $[d]$ to denote the collection of integers $\{1, 2, \ldots, d\}$. For any set S, denote by S^c its complement, and $|S|$ its cardinality. For a vector $u \in \mathbb{R}^d$ and $1 \le p < \infty$, define its ℓ_p-norm by $\|u\|_p = \big(\sum_{i=1}^d |u_i|^p\big)^{1/p}$. In particular, $\|u\|_\infty = \max_{i=1,\ldots,d} |u_i|$ is the ℓ_∞-norm. For any $d_1 \times d_2$ matrix $M = (M_{k\ell})_{1\le k\le d_1, 1\le \ell\le d_2}$, let $\|M\|_F = \sqrt{\sum_{k=1}^{d_1}\sum_{\ell=1}^{d_2} M_{k\ell}^2}$ be the Frobenius norm and let $\|M\|_\infty = \max_{(k,\ell)\in[d_1]\times[d_2]} |M_{k\ell}|$ 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 $\|M\|_{p,q} = \sup_{\|u\|_p=1} \|Mu\|_q$. It is easy to verify that $\|M\|_{p,q} = \|M^\intercal\|_{q^*,p^*}$, where (p, p^*) and (q, q^*) are conjugate pairs; that is, $1/p + 1/p^* = 1/q + 1/q^* = 1$. In particular, $\|M\| = \|M\|_{2,2}$ is the spectral norm; $\|M\|_{2,\infty} = \max_{k=1,\ldots,d_1} \sqrt{\sum_{\ell=1}^{d_2} M_{k\ell}^2}$ is also known as the maximum row norm of M. Moreover, for two real numbers a and b, we write for ease of presentation that $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

2.1. Max-norm and trace-norm

For a matrix $M \in \mathbb{R}^{d_1\times d_2}$, the trace-norm (also known as the Schatten 1-norm) $\|M\|_1$ is defined as the sum of all singular values of M, or equivalently,

$$\|M\|_1 = \inf\left\{\sum_{j=1}^{d_1\wedge d_2} |\sigma_j| : M = \sum_{j=1}^{d_1\wedge d_2} \sigma_j u_j v_j^\intercal,\ u_j \in \mathbb{R}^{d_1}, v_j \in \mathbb{R}^{d_2},\ \|u_j\|_2 = \|v_j\|_2 = 1\right\}.$$

In other words, the trace-norm promotes low-rank decompositions with factors in ℓ_2. Similarly, using Grothendiek's inequality (Jameson, 1987), the max-norm defined in (1.1) has the following analogous representation in terms of factors in ℓ_∞:

$$\|M\|_{\max} \approx \inf\left\{\sum_{j=1}^{d_1\wedge d_2}|\sigma_j| : M = \sum_{j=1}^{d_1\wedge d_2}\sigma_j u_j v_j^{\intercal},\ \|u_j\|_\infty = \|v_j\|_\infty = 1\right\}.$$

The factor of equivalence is the Grothendieck's constant $K_G \in (1.67, 1.79)$. Based on these properties, the max-norm regularization is expected to be more effective when dealing with uniformly bounded data (Lee et al., 2010).

Of the same spirit as the definition of the max-norm in (1.1), the trace-norm has the following equivalent characterization in terms of matrix factorizations:

$$\|M\|_1 = \min_{U,V:M=UV^{\intercal}}\|U\|_F\|V\|_F = \frac{1}{2}\min_{U,V:M=UV^{\intercal}}\left(\|U\|_F^2 + \|V\|_F^2\right).$$

See, for example, Srebro and Shraibman (2005). It is easy to see that

$$\frac{1}{\sqrt{d_1 d_2}}\|M\|_1 \le \|M\|_{\max}, \tag{2.1}$$

which in turn implies that any low max-norm approximation is also a low trace-norm approximation. As pointed out by Srebro and Shraibman (2005), there can be a large gap between $\frac{1}{\sqrt{d_1 d_2}}\|\cdot\|_1$ and $\|\cdot\|_{\max}$. The following relation between the trace-norm and Frobenius norm is well-known: $\|M\|_F \le \|M\|_1 \le \sqrt{\operatorname{rank}(M)}\cdot\|M\|_F$. An analogous bound holds for the max-norm, in connection with the element-wise ℓ_∞-norm (Linial et al., 2004):

$$\|M\|_\infty \le \|M\|_{\max} \le \sqrt{\operatorname{rank}(M)}\cdot\|M\|_{1,\infty} \le \sqrt{\operatorname{rank}(M)}\cdot\|M\|_\infty. \tag{2.2}$$

For any $R > 0$, let

$$\mathbb{B}_{\max}(R) = \left\{M \in \mathbb{R}^{d_1\times d_2} : \|M\|_{\max} \le R\right\}$$
$$\text{and}\quad \mathbb{B}_{\mathrm{tr}}(R) = \left\{M \in \mathbb{R}^{d_1\times d_2} : \|M\|_1 \le R\right\}$$

be the max-norm and trace-norm ball with radius R, respectively. It is now well-known (Srebro and Shraibman, 2005) that $\mathbb{B}_{\max}(1)$ can be bounded, from both below and above, by the convex hull of rank-one sign matrices $\mathcal{M}_\pm = \{M \in \{\pm 1\}^{d_1\times d_2} : \operatorname{rank}(M) = 1\}$, i.e.

$$\operatorname{conv}\mathcal{M}_\pm \subseteq \mathbb{B}_{\max}(1) \subseteq K_G\cdot\operatorname{conv}\mathcal{M}_\pm \tag{2.3}$$

with $K_G \in (1.67, 1.79)$ denoting the Grothendieck's constant. Moreover, $\mathcal{M}_\pm$ is a finite class with cardinality $|\mathcal{M}_\pm| = 2^{d-1}$, where $d = d_1 + d_2$.

2.2. Rademacher complexity

A technical tool used in our analysis involves data-dependent estimates of the Rademacher and Gaussian complexities of a function class. We refer to Bartlett and Mendelson (2002) and Srebro and Shraibman (2005) for a detailed introduction of these concepts.

Definition 2.1. *For a class $\mathcal{F}$ of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, its empirical Rademacher complexity over a specific sample $S = (x_1, x_2, \ldots, x_n) \subseteq \mathcal{X}$ is given by*

$$\widehat{R}_S(\mathcal{F}) = \frac{2}{|S|}\mathbb{E}_{\boldsymbol{\varepsilon}}\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right\},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^{\mathsf{T}}$ is a Rademacher sequence. The Rademacher complexity with respect to a distribution $\mathcal{P}$ is the expectation, over an independent and identically distributed (i.i.d.) sample of $|S|$ points drawn from $\mathcal{P}$, denoted by

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

Replacing $\varepsilon_1, \ldots, \varepsilon_n$ with independent standard normal variables $g_1, \ldots, g_n$ leads to the definition of (empirical) Gaussian complexity.

Considering a matrix as a function from the index pairs to the entry values, Srebro and Shraibman (2005) obtained upper bounds on the Rademacher complexity of the unit balls under both the trace-norm and the max-norm. Specifically, for any $d_1, d_2 > 2$ and any sample of size $2 < |S| < d_1 d_2$, the empirical Rademacher complexity of the max-norm unit ball is bounded by

$$\widehat{R}_S\left(\mathbb{B}_{\max}(1)\right) \le 12\sqrt{\frac{d_1+d_2}{|S|}}. \tag{2.4}$$

3. Max-norm constrained empirical risk minimization

3.1. The statistical model

We now consider matrix completion under a general random sampling model. Let $M^* \in \mathbb{R}^{d_1\times d_2}$ be the unknown target matrix. Suppose that a random sample

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

of the index set is drawn independently according to a general sampling distribution $\Pi = \{\pi_{k\ell}\}_{1\le k\le d_1, 1\le \ell\le d_2}$ on $[d_1]\times[d_2]$, with replacement; that is, $\mathbb{P}\{(i_t, j_t) = (k, \ell)\} = \pi_{k\ell}$ for all $t = 1, \ldots, n$ and $(k, \ell) \in [d_1]\times[d_2]$. Given a random index subset $S = \{(i_1, j_1), \ldots, (i_n, j_n)\}$ of size n, we observe noisy entries $\{Y_{i_t j_t}\}_{t=1}^n$ indexed by S, i.e.

$$Y_{i_t j_t} = M^*_{i_t j_t} + \sigma\xi_t, \quad t = 1, \ldots, n, \tag{3.1}$$

for some $\sigma > 0$. The noise variables ξ_t are independent with zero mean and unit variance. By expressing the model as in (3.1), it is implicitly assumed that the noise on the entry is drawn independently each time.

Instead of assuming the uniform sampling distribution, we consider a general sampling distribution Π here. Since $\sum_{k=1}^{d_1}\sum_{\ell=1}^{d_2}\pi_{k\ell}=1$, we have $\max_{k,\ell}\pi_{k\ell}\geq (d_1d_2)^{-1}$. Motivated by some applications, to ensure that each entry is observed with a positive probability, it is sometimes natural to assume that there exists a positive constant $\nu\geq 1$ such that

$$\pi_{k\ell}\geq \frac{1}{\nu\, d_1d_2} \tag{3.2}$$

holds for all $(k,\ell)\in[d_1]\times[d_2]$. We write hereafter $d=d_1+d_2$ for brevity. Clearly, $\max(d_1,d_2)\leq d\leq 2\max(d_1,d_2)$.

The rescaled Frobenius norm $(d_1d_2)^{-1}\|\cdot\|_F^2$ is typically used in the literature as a natural measure of the estimation accuracy. Now that the sampling distribution Π is arbitrary, we use instead the weighted Frobenius norm with respect to Π to measure the estimation error. For any $A=(A_{k\ell})\in\mathbb{R}^{d_1\times d_2}$, define

$$\|A\|_\Pi^2=\mathbb{E}_{(i,j)\sim\Pi}A_{ij}^2=\sum_{k=1}^{d_1}\sum_{\ell=1}^{d_2}\pi_{k\ell}A_{k\ell}^2. \tag{3.3}$$

When Π corresponds to the uniform distribution, $\|A\|_\Pi=(d_1d_2)^{-1/2}\|A\|_F$.

The preceding work on matrix completion has mainly focused on the case of exact low-rank matrices. Here we allow a relaxation of this assumption and consider the more general setting of approximately low-rank matrices. Specifically, we consider recovery of matrices with ℓ_∞-norm and max-norm constraints defined by

$$\mathcal{K}(\alpha,R):=\big\{M\in\mathbb{R}^{d_1\times d_2}:\|M\|_\infty\leq\alpha,\|M\|_{\max}\leq R\big\}. \tag{3.4}$$

Here both α and R are free parameters to be determined. If the matrix M^* is of rank at most r and $\|M^*\|_\infty\leq\alpha$, then by (2.2) we have $M^*\in\mathbb{B}_{\max}(\alpha\sqrt{r})$ and hence $M^*\in\mathcal{K}(\alpha,\alpha\sqrt{r})$.

3.2. Max-norm constrained least squares estimator

Given a collection of observations $Y_S=\{Y_{i_tj_t}\}_{t=1}^n$ from the observation model (3.1), we estimate the unknown $M^*\in\mathcal{K}(\alpha,R)$ for some $\alpha,R>0$ by the minimizer of the empirical risk with respect the quadratic loss function

$$\widehat{\mathcal{L}}_n(M;Y)=\frac{1}{n}\sum_{t=1}^n(Y_{i_tj_t}-M_{i_tj_t})^2.$$

That is,

$$\widehat{M}_{\max}:=\operatorname*{arg\,min}_{M\in\mathcal{K}(\alpha,R)}\widehat{\mathcal{L}}_n(M;Y). \tag{3.5}$$

The minimization procedure requires that all the entries of M^* are bounded in magnitude by a prespecified constant α. This condition enforces that M^* should not be too "spiky", and a too large bound may jeopardize exactness of the estimation. See, for example, Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2012) and Klopp (2014). On the other hand, as argued in Lee et al. (2010), the max-norm regularization is expected to be more effective particularly for uniformly bounded data, which is our main motivation for using the max-norm constrained estimator.

Although the max-norm constrained minimization problem (3.5) is a convex program, fast and efficient algorithms for solving large-scale optimization problems that incorporate the max-norm have only been developed recently in Lee et al. (2010) and Fang et al. (2015b). We will show in Section 4 that the convex optimization problem (3.5) can be implemented in polynomial time as a function of the sample size n and dimensions d_1 and d_2.

3.3. Upper bounds

In this section, we state our main results regarding the recovery of an approximately low-rank (low-max-norm) matrix M^* using max-norm constrained empirical risk minimization.

Theorem 3.1. *Suppose that the noise sequence $\{\xi_t\}_{t=1}^n$ are independent sub-exponential random variables; that is, there is a constant $K > 0$ such that*

$$\max_{1\le t\le n} \mathbb{E}\{\exp(|\xi_t|/K)\} \le e. \tag{3.6}$$

The parameters $\alpha, R > 0$ are such that $M^* \in \mathcal{K}(\alpha, R)$. Then, for a sample size n satisfying $d \le n \le d_1 d_2$,

$$\|\widehat{M}_{\max} - M^*\|_\Pi^2 \le C(\alpha \vee K\sigma)R\sqrt{\frac{d}{n}}, \tag{3.7}$$

with probability greater than $1 - 2e^{-d}$, where $C > 0$ is an absolute constant. If, in addition, assumption (3.2) is satisfied, then for a sample size n with $d \le n \le d_1 d_2$,

$$\frac{1}{d_1 d_2}\|\widehat{M}_{\max} - M^*\|_F^2 \le C\,\nu(\alpha \vee K\sigma)R\sqrt{\frac{d}{n}} \tag{3.8}$$

holds with probability at least $1 - 2e^{-d}$.

Remark 3.1.

(1) It is worth noticing that the general result on approximate reconstruction guarantee (3.7) holds without any prior information on the sampling distribution Π, in particular the lower bound assumption (3.2). In fact, it is reflected in the result that for every location index (k, ℓ), the smaller the sampling probability $\pi_{k\ell}$ is, the more difficult it will be to recovery the entry at this location.

(2) The upper bounds given in Theorem 3.1 hold with high probability. The rate of convergence under expectation can be obtained as a direct consequence. More specifically, for a sample size n with $d \le n \le d_1 d_2$, we have

$$\sup_{M^* \in \mathcal{K}(\alpha, R)} \frac{1}{d_1 d_2} \mathbb{E} \|\widehat{M}_{\max} - M^*\|_F^2 \le C\, \nu(\alpha \vee \sigma) R \sqrt{\frac{d}{n}}. \tag{3.9}$$

In view of the upper bound in (6.1), when the noise level σ is comparable to or dominated by α, the rate is of order $\alpha R\,(\frac{d}{n})^{1/2}$. To fully understand how the random noise affects the estimation accuracy particularly when σ is much smaller than α, we provide a complementary result in Theorem 3.2 which generalizes Theorem 9 in Foygel and Srebro (2011) to the general non-uniform sampling model.

Theorem 3.2. *Assume that the conditions of Theorem 3.1 are satisfied and $\sigma \le \alpha$. Then,*

$$\begin{aligned} &\|\widehat{M}_{\max} - M^*\|_\Pi^2 \\ &\le C \left\{ \sigma \sqrt{(\log n)^3 \frac{R^2 d}{n} + (\log n)^{3/2} \frac{\alpha^2}{n}} + (\log n)^3 \frac{R^2 d}{n} + (\log n)^{3/2} \frac{\alpha^2}{n} \right\} \end{aligned} \tag{3.10}$$

holds with probability at least $1 - 2n^{-1}$ over a random sample of size n satisfying $d \le n \le d_1 d_2$, where $C > 1$ is a constant.

An interesting consequence of Theorem 3.2 is that, in the noiseless case where $\sigma = 0$ and a random subset of the entries of M^* are perfectly observed, then for any prespecified tolerance level $\epsilon > 0$, the target matrix M^* can be approximately recovered in the sense that $\|\widehat{M}_{\max} - M^*\|_\Pi^2 \le \epsilon$ whenever the sample size $n \gtrsim \max\{\frac{R^2 d}{\epsilon}(\log n)^3, \frac{\alpha^2}{\epsilon}(\log n)^{3/2}\}$.

3.4. Information-theoretic lower bounds

Theorem 3.1 gives the rate of convergence for the max-norm constrained least squares estimator $\widehat{M}_{\max}$. In this section we shall use information-theoretical methods to establish a minimax lower bound for *non-uniform sampling at random* matrix completion on the max-norm ball. The minimax lower bound matches the rate of convergence given in (3.8) when the sampling distribution Π satisfies $\frac{1}{\nu\, d_1 d_2} \le \min_{k,\ell} \pi_{k\ell} \le \max_{k,\ell} \pi_{k\ell} \le \frac{\mu}{d_1 d_2}$ for some constants ν and μ. The results show that the max-norm constrained least-squares estimator is indeed rate-optimal in such a setting.

To derive the lower bound, we assume that the sampling distribution Π satisfies

$$\max_{k,\ell} \pi_{k\ell} \le \frac{\mu}{d_1 d_2} \tag{3.11}$$

for a positive constant $\mu \ge 1$. Clearly, when $\mu = 1$, it amounts to say that the sampling distribution is uniform.

Theorem 3.3. *Suppose that the noise sequence* $\{\xi_t\}_{t=1}^n$ *are i.i.d. standard normal random variables, the sampling distribution* Π *satisfies the condition* (3.11) *and the quintuple* (n, d_1, d_2, α, R) *satisfies*

$$\frac{48\alpha^2}{d_1 \vee d_2} \le R^2 \le \frac{\sigma^2(d_1 \wedge d_2)d_1 d_2}{128\mu n}. \tag{3.12}$$

Then the minimax $\|\cdot\|_F$*-risk is lower bounded as*

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{K}(\alpha,R)} \frac{1}{d_1 d_2}\mathbb{E}\|\widehat{M} - M\|_F^2 \ge \min\left\{\frac{\alpha^2}{16}, \frac{\sigma}{256}R\sqrt{\frac{d}{\mu n}}\right\}. \tag{3.13}$$

In particular, for a sample size $n \ge \frac{1}{\alpha^2\mu}R^2 d$,

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{K}(\alpha,R)} \frac{1}{d_1 d_2}\mathbb{E}\|\widehat{M} - M\|_F^2 \ge \frac{1}{256}(\alpha \wedge \sigma)R\sqrt{\frac{d}{\mu n}}. \tag{3.14}$$

Assume that both ν and μ, respectively appeared in (3.2) and (3.11), are bounded above by universal constants, then comparing the lower bound (3.14) with the upper bound (3.9) shows that if the sample size $n > (\frac{R}{\alpha})^2 d$, the optimal rate of convergence is $R\sqrt{d/n}$; that is,

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{K}(\alpha,R)} \frac{1}{d_1 d_2}\mathbb{E}\|\widehat{M} - M\|_F^2 \asymp R\sqrt{\frac{d_1 + d_2}{n}}, \tag{3.15}$$

and the max-norm constrained least-squares estimator (3.5) is rate-optimal. The requirement here on the sample size $n > (\frac{R}{\alpha})^2(d_1 + d_2)$ is weak. If, in addition, $d_1 = d_2$, condition (3.12) is reduced to $\alpha^2 d^{-1} \lesssim R^2 \lesssim \sigma\alpha d$, which is a mild constraint since R^2 is of order $\alpha^2 r_0$ in the exact low-rank case where $r_0 = \operatorname{rank}(M^*)$.

The proof of Theorem 3.3 uses information-theoretic methods. A key technical tool for the proof is the following lemma which guarantees the existence of a suitably large packing set for $\mathcal{K}(\alpha, R)$ in the Frobenius norm.

Lemma 3.1. *Let* $r = (\frac{R}{\alpha})^2$ *and let* $\gamma \le 1$ *be such that* $r \le \gamma^2(d_1 \wedge d_2)$ *is an integer. Then, there exists a subset* $\mathcal{M} \subseteq \mathcal{K}(\alpha, R)$ *with cardinality*

$$|\mathcal{M}| = \left\lfloor \exp\left\{\frac{r(d_1 \vee d_2)}{16\gamma^2}\right\}\right\rfloor + 1$$

and with the following properties:

(i) *For any* $M = (M_{k\ell}) \in \mathcal{M}$, $\operatorname{rank}(M) \le r/\gamma^2$ and $M_{k\ell} \in \{\pm\gamma\alpha\}$, such that

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

(ii) For any two distinct $M^i, M^j \in \mathcal{M}$,

$$\frac{1}{d_1 d_2}\|M^i - M^j\|_F^2 > \frac{\gamma^2\alpha^2}{2}.$$

The proof of Lemma 3.1 is based on an adaptation of the arguments used to prove Lemma 3 in Davenport et al. (2014), which for self-containment, is given in Section 6.4.

3.5. Comparison to past work

We now compare the results established in this section with those known in the literature for matrix completion under uniform or general sampling schemes.

3.5.1. Approximate/non-exact low-rank recoveries

It is now well-known that the exact recovery of a low-rank matrix in the noiseless case requires the "incoherence conditions" on the target matrix M^* (Candés and Recht, 2009; Candés and Tao, 2010; Recht, 2011; Gross, 2011). In this paper, we consider instead a general setting of approximately low-rank matrices, and prove that approximate recovery is still possible without enforcing exact structural assumptions.

Our results are directly comparable to those of Koltchinskii, Lounici and Tsybakov (2011) and Negahban and Wainwright (2012), in which the trace-norm was used as a proxy to the rank. Taking the latter as an example to illustrate, Negahban and Wainwright (2012) considered the setup where the sampling distribution is a *product distribution*, i.e. for all $(k, \ell) \in [d_1] \times [d_2]$,

$$\pi_{k\ell} = \pi_{k\cdot}\pi_{\cdot\ell},$$

where $\pi_{k\cdot}$ and $\pi_{\cdot\ell}$ are marginals that satisfy

$$\pi_{k\cdot} \geq \frac{1}{\sqrt{\nu} d_1}, \quad \pi_{\cdot\ell} \geq \frac{1}{\sqrt{\nu} d_2} \quad \text{for some } \nu \geq 1. \tag{3.16}$$

Accordingly, define the weighted norms as

$$\|M\|_{w(\dagger)} := \left\|\sqrt{W_{\mathrm{r}}} M \sqrt{W_{\mathrm{c}}}\right\|_{\dagger}, \quad \dagger \in \{F, 1, \infty\},$$

where $W_{\mathrm{r}} = d_1 \cdot \mathrm{diag}(\pi_{1\cdot}, \ldots, \pi_{d_1\cdot})$ and $W_{\mathrm{c}} = d_2 \cdot \mathrm{diag}(\pi_{\cdot 1}, \ldots, \pi_{\cdot d_2})$

Based on a collection of observations

$$Y_{i_t j_t} = \varepsilon_t M^*_{i_t j_t} + \sigma\xi_t, \quad t = 1, \ldots, n,$$

where (i_t, j_t) are i.i.d. according to $\mathbb{P}\{(i_t, j_t) = (k, \ell)\} = \pi_{k\ell}$ and $\varepsilon_t \in \{-1, +1\}$ are i.i.d. random signs, and under the assumption that the unknown matrix M^* satisfies

$$\|M^*\|_{w(1)} \leq R\sqrt{d_1 d_2}, \quad \|M^*\|_{w(F)} \leq \sqrt{d_1 d_2} \quad \text{and} \quad \frac{\|M^*\|_{w(\infty)}}{\|M^*\|_{w(F)}} \leq \frac{\alpha}{\sqrt{d_1 d_2}}, \tag{3.17}$$

Negahban and Wainwright (2012) proposed the following estimator of M^* based on the trace-norm penalized minimization:

$$\widehat{M}_{\mathrm{tr}} \in \underset{\|M\|_{w(\infty)} \le \alpha}{\arg\min} \left\{ \frac{1}{n} \sum_{t=1}^{n} (Y_{i_t j_t} - \varepsilon_t M_{i_t j_t})^2 + \lambda_n \|M\|_{w(1)} \right\}. \tag{3.18}$$

In the context of low-trace-norm (approximately low-rank) matrix recovery where the true matrix M^* satisfies (3.17), they proved that for properly chosen λ_n depending on σ (see, e.g. Corollary 2 therein), there exist absolute positive constants c_1–c_3 such that

$$\frac{1}{d_1 d_2} \|\widehat{M}_{\mathrm{tr}} - M^*\|_F^2 \le c_1 \nu \left\{ (\sigma \vee \nu) \alpha R \sqrt{\frac{d \log d}{n}} + \frac{\nu \alpha^2}{n} \right\}, \tag{3.19}$$

holds with probability at least $1 - c_2 \exp(-c_3 \log d)$.

First, the product distribution assumption can be fairly restrictive in practice and is not valid in many applications. For example, in the case of the Netflix problem, this assumption would imply that conditional on any movie, it will be rated by all users with the same probability. Second, the constraint on M^* highly depends on the true sampling distribution which is really unknown in practice and can only be estimated based on the empirical frequencies, i.e. for any pair $(k, \ell) \in [d_1] \times [d_2]$,

$$\widehat{\pi}_{k\cdot} = \frac{1}{n} \sum_{t=1}^{n} 1\{i_t = k\}, \quad \widehat{\pi}_{\cdot \ell} = \frac{1}{n} \sum_{t=1}^{n} 1\{j_t = \ell\}.$$

Since only a relatively small sample of the entries of M^* is observed, these estimates may not be accurate enough. The max-norm constrained minimization approach, on the other hand, is proved (Theorem 3.1) to be effective in the presence of non-uniform sampling distributions. The method does not require either a product distribution or the knowledge of the exact true sampling distribution. From this point of view, the max-norm constrained method indeed yields a more robust approximate recovery guarantee, with respect to the sampling distributions.

We now turn to the special case of uniform sampling. The "spikeness" assumption in Negahban and Wainwright (2012) can actually be reduced to a single constraint on the ℓ_∞-norm (Klopp, 2014). Let $\mathbb{B}_\infty(\alpha) = \{M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_\infty \le \alpha\}$ be the ℓ_∞-norm ball with radius α. Define the class of matrices

$$\mathcal{K}_{\mathrm{tr}}(\alpha, R) := \big\{ M \in \mathbb{B}_\infty(\alpha) : (d_1 d_2)^{-1/2} \|M\|_1 \le R \big\}. \tag{3.20}$$

It can be seen from (2.1) and (2.2) that $\{M \in \mathbb{B}_\infty(\alpha) : \mathrm{rank}(M) \le r\} \subsetneq \mathcal{K}(\alpha, \alpha\sqrt{r}) \subsetneq \mathcal{K}_{\mathrm{tr}}(\alpha, \alpha\sqrt{r})$. The following results provide upper bounds on the accuracy of both the max- and trace-norm regularized estimators under the Frobenius norm.

Corollary 3.1. *Suppose that the noise sequence* $\{\xi_t\}_{t=1}^n$ *are i.i.d.* $N(0,1)$ *random variables and the sampling distribution* Π *is uniform on* $[d_1] \times [d_2]$. *Then the following inequalities hold with probability at least* $1 - 3d^{-1}$*:*

(i) *The optimum* $\widehat{M}_{\max}$ *to the convex program* (3.5) *satisfies*

$$\sup_{M^* \in \mathcal{K}(\alpha,R)} \frac{1}{d_1 d_2} \|\widehat{M}_{\max} - M^*\|_F^2 \lesssim (\sigma \vee \alpha) R \sqrt{\frac{d}{n}} + \alpha^2 \frac{\log d}{n}. \tag{3.21}$$

(ii) *The minimum* $\widehat{M}_{\mathrm{tr}}$ *to the SDP* (3.18) *with all weighted norms replaced by the standard ones and with a properly chosen* λ_n *satisfies*

$$\sup_{M^* \in \mathcal{K}_{\mathrm{tr}}(\alpha,R)} \frac{1}{d_1 d_2} \|\widehat{M}_{\mathrm{tr}} - M^*\|_F^2 \lesssim (\sigma \vee \alpha) R \sqrt{\frac{d \log d}{n}} + \alpha^2 \frac{\log d}{n}. \tag{3.22}$$

The upper bound (3.21) follows immediately from (6.1) in Theorem 3.1, and (3.22) is a straightforward extension of Theorem 7 in Klopp (2014) on exact low-rank matrix recovery to the case of low-trace-norm matrix reconstruction. The proof is essentially the same and thus is omitted.

Foygel and Srebro (2011) analyzed the recovery guarantee for $\widehat{M}_{\max}$ based on an excess risk bound for empirical risk minimization with a smooth loss function recently developed in Srebro, Sridharan and Tewari (2010). Specifically, assuming a uniform sampling model with sub-exponential noise and that the target matrix $M^* \in \mathcal{K}(\alpha, R)$, they proved that with high probability,

$$\frac{1}{d_1 d_2} \|Y - \widehat{M}_{\max}\|_F^2 - \widehat{\sigma}^2 \lesssim (\log n)^{3/2} \widehat{\sigma} \sqrt{\frac{R^2 d}{n}} + (\log n)^3 \frac{R^2 d}{n}, \tag{3.23}$$

where $Y = M^* + Z$ with $Z = (\xi_{k\ell})_{1 \le k \le d_1, 1 \le \ell \le d_2}$, and $\widehat{\sigma}^2 := \frac{1}{d_1 d_2} \sum_{k=1}^{d_1} \sum_{\ell=1}^{d_2} \xi_{k\ell}^2$ denotes the average noise level which is concentrated around σ^2 with high probability.

After a more delicate analysis, our result shows that the additional logarithmic factors in (3.23) purely arise from an artifact of the proof technique and thus can be avoided. Moreover, in view of the lower bounds given in Theorem 3.3, we see that the max-norm constrained least square estimator $\widehat{M}_{\max}$ achieves the optimal rate of convergence for recovering approximately low-rank matrices over the parameter space $\mathcal{K}(\alpha, R)$ under the Frobenius norm loss. To our knowledge, the best known rate for the trace-norm regularized estimator given in (3.22) is near-optimal up to logarithmic factors in a minimax sense, over a larger parameter space $\mathcal{K}_{\mathrm{tr}}(\alpha, R)$.

3.5.2. Uniform/non-uniform sampling distributions

We now provide further insight into the rationale behind the phenomenon that the max-norm regularized/constrained method is more robust with respect to

the sampling distribution. As before, we focus on the setting with a product sampling distribution $\pi_{k\ell} = \pi_{k\cdot}\pi_{\cdot\ell}$ for $(k,\ell) \in [d_1] \times [d_2]$.

Motivated by Salakhutdinov and Srebro (2010), Negahban and Wainwright (2012) studied the weighted trace-norm penalized estimator $\widehat{M}_{\mathrm{tr}}$ given at (3.18), where for any matrix $M \in \mathbb{R}^{d_1 \times d_2}$,

$$\|M\|_{w(1)} = \sqrt{d_1 d_2}\, \left\|\mathrm{diag}(\sqrt{\pi_{1\cdot}}, \ldots, \sqrt{\pi_{d_1\cdot}})\, M\, \mathrm{diag}(\sqrt{\pi_{\cdot 1}}, \ldots, \sqrt{\pi_{\cdot d_2}})\right\|_1 . \quad (3.24)$$

However, the "true" form of the sampling distribution is ambiguous and even if it is a product distribution, the marginal probabilities $\pi_{k\cdot}$ and $\pi_{\cdot\ell}$ are typically unknown. Therefore, the weighted trace-norm $\|\cdot\|_{w(1)}$ can not be used in practice.

For the max-norm, a useful equivalent definition is that for any $M \in \mathbb{R}^{d_1 \times d_2}$,

$$\|M\|_{\max} = \max_{u \in \mathbb{R}^{d_1}, v \in \mathbb{R}^{d_2} : \|u\|_2 = \|v\|_2 = 1} \|\mathrm{diag}(u)\, M\, \mathrm{diag}(v)\|_1 .$$

See, for example, Theorem 9 in Lee, Shraibman and Spalek (2008). As a result, by considering a max-norm penalized estimator that solves

$$\min_{\|M\|_\infty \le \alpha} \left\{ \frac{1}{n} \sum_{t=1}^{n} (Y_{i_t j_t} - \varepsilon_t M_{i_t j_t})^2 + \lambda_n \|M\|_{\max} \right\},$$

all the possible marginal probabilities are taken into account, and therefore the solution is expected to be more robust with respect to the unknown sampling distributions.

Although the sampling distribution is not known exactly in practice, its estimated version is expected to be stable enough as an alternative. According to Foygel et al. (2011), given a random sample $S = \{(i_t, j_t)\}_{t=1}^n$, we can estimate $\pi_{k\ell}$ by $\widehat{\pi}_{k\ell} = \widehat{\pi}_{k\cdot}\widehat{\pi}_{\cdot\ell}$ with empirical marginals $\widehat{\pi}_{k\cdot} = n^{-1}\sum_{t=1}^n 1\{i_t = k\}$ and $\widehat{\pi}_{\cdot\ell} = n^{-1}\sum_{t=1}^n 1\{j_t = \ell\}$, or by $\widetilde{\pi}_{ij} = \widetilde{\pi}_{k\cdot}\widetilde{\pi}_{\cdot\ell}$ with smoothed empirical marginals

$$\widetilde{\pi}_{k\cdot} = \frac{1}{2}\left(\widehat{\pi}_{k\cdot} + d_1^{-1}\right), \quad \widetilde{\pi}_{\cdot\ell} = \frac{1}{2}\left(\widehat{\pi}_{\cdot\ell} + d_2^{-1}\right).$$

The empirically-weighted trace-norm $\|\cdot\|_{\widehat{w}(1)}$ can be defined in the same spirit as in (3.24) for the weighted trace-norm, only with $\pi_{k\ell}$ replaced by $\widehat{\pi}_{k\ell}$. Then the unknown matrix can be estimated via penalization on the $\widehat{\pi}$-weighted trace-norm, i.e.

$$\min_{\|M\|_\infty \le \alpha} \left\{ \frac{1}{n} \sum_{t=1}^{n} (Y_{i_t j_t} - \varepsilon_t M_{i_t j_t})^2 + \lambda_n \|M\|_{\widehat{w}(1)} \right\}.$$

Foygel et al. (2011) proved the error bound for the excess risk of the empirically-weighted trace-norm constrained estimator when the loss function is Lipschitz. It is interesting to investigate whether the results similar to those in Negahban and Wainwright (2012) hold for the empirically-weighted trace-norm constrained and penalized estimators when the quadratic loss function is used.

It is also worth noting that, under condition (3.6) and when the sampling distribution is nearly uniform in the sense that

$$\min_{(k,\ell)\in[d_1]\times[d_2]} \pi_{k\ell} \geq \frac{1}{\nu\, d_1 d_2} \quad \text{and} \quad \max\left(\sum_{k=1}^{d_1}\pi_{k\ell}, \sum_{\ell=1}^{d_2}\pi_{k\ell}\right) \leq \frac{L}{\min(d_1,d_2)} \tag{3.25}$$

for some constants $\nu, L \geq 1$, Klopp (2014) showed that the trace-norm penalized estimator

$$\widehat{M}_{\mathrm{tr}}(\lambda) \in \underset{\|M\|_\infty \leq \alpha}{\arg\min} \left\{ \frac{1}{n}\sum_{t=1}^{n}(Y_{i_t j_t} - \varepsilon_t M_{i_t j_t})^2 + \lambda \|M\|_1 \right\}$$

satisfies

$$\frac{1}{d_1 d_2}\|\widehat{M}_{\mathrm{tr}}(\lambda) - M^*\|_F^2 \lesssim (\sigma \vee \alpha)^2 \nu^2 L \frac{r_0\, d \log d}{n} + \nu\alpha^2\sqrt{\frac{\log d}{n}}$$

with probability greater than $1 - 3d^{-1}$, provided that $\|M^*\|_\infty \leq \alpha$ and $\lambda = \lambda_n \asymp \sigma(\frac{L\log d}{nd})^{1/2}$. In the case of Gaussian errors and under condition (3.11), the above rate of convergence is minimax optimal, up to a logarithmic factor, for the class of exact low-rank matrices $\{M \in \mathbb{R}^{d_1\times d_2} : \|M\|_\infty \leq \alpha, \mathrm{rank}(M) \leq r_0\}$ (Koltchinskii, Lounici and Tsybakov, 2011). An interesting and challenging open problem is that in the context of exact low-rank matrix recovery and when the sampling probabilities satisfy (3.25), whether the optimal recovery guarantee can be achieved using the max-norm constrained method. Also, to the best of our knowledge, there are no theoretical guarantees for exactly recovering a low-rank matrix when the sampling distribution is non-uniform and unspecified.

4. Computational algorithms

Although Theorem 3.1 presents theoretical guarantees that hold uniformly for any global minimizer, it does not provide guidance on how to approximate such a global minimizer using a polynomial-time algorithm. A parallel line of work has studied computationally efficient algorithms for solving problems with the trace-norm constraint or penalization. See Lin et al. (2009), Mazumber, Hastie and Tibshirani (2010) and Nesterov (2013), among others. Here we restrict our attention to the less-studied max-norm oriented approach. We discuss two different types of algorithms which are particularly designed to solve large scale optimization problems that incorporate the max-norm as a semidefinite relaxation of the rank. The first one is a fast first-order algorithm developed in Lee et al. (2010) based on nonconvex relaxation. The problem of interest to us is the optimization program (3.5) with both the max-norm and the element-wise ℓ_∞-norm constraints, in which case the algorithm introduced in Lee et al. (2010) can be applied after suitable modifications as described in Section 4.1. The second one, on the other hand, is a convex approach proposed by Fang et al. (2015b) using the alternating direction of multipliers method with guaranteed convergence to the global optimum since it deals with the convex problem (4.1) directly.

4.1. A projected gradient method

Due to Srebro, Rennie and Jaakkola (2004), the max-norm of a $d_1 \times d_2$ matrix M can be computed via a semi-definite program:

$$\|M\|_{\max} = \min R \quad \text{s.t.} \quad \begin{pmatrix} W_1 & M \\ M^\intercal & W_2 \end{pmatrix} \succeq 0, \quad \operatorname{diag}(W_1) \le R, \quad \operatorname{diag}(W_2) \le R.$$

Correspondingly, we can reformulate (3.5) as the following SDP problem

$$\begin{aligned} &\min f(M) \\ &\text{s.t.} \begin{pmatrix} W_1 & M \\ M^\intercal & W_2 \end{pmatrix} \succeq 0, \quad \operatorname{diag}(W_1) \le R, \quad \operatorname{diag}(W_2) \le R, \quad \|M\|_\infty \le \alpha, \end{aligned}$$

where the objective function f is given by

$$f(M) = f(M; Y) = \widehat{\mathcal{L}}_n(M; Y).$$

This SDP can be solved using standard interior-point methods, though are fairly slow and do not scale to matrices with large dimensions. For large-scale problems, an alternative factorization method based on (1.1), as described below, is preferred (Lee et al., 2010).

We begin by introducing dummy variables $U \in \mathbb{R}^{d_1 \times k}$, $V \in \mathbb{R}^{d_2 \times k}$ for some $1 \le k \le d_1 + d_2$ and let $M = UV^\intercal$. If the optimal solution $\widehat{M}_{\max}$ is known to have rank at most r, we can take $U \in \mathbb{R}^{d_1 \times (r+1)}$, $V \in \mathbb{R}^{d_2 \times (r+1)}$. In practice, without a known guarantee on the rank of $\widehat{M}_{\max}$, we alternatively truncate the number of columns k to some reasonably high value less that $d_1 + d_2$. Then, we rewrite the original problem (3.5) in the factored form as follows:

$$\begin{aligned} \text{minimize} \quad & f(UV^\intercal) = \frac{1}{n} \sum_{t=1}^{n} (U_{i_t}^\intercal V_{j_t} - Y_{i_t j_t})^2 \\ \text{subject to} \quad & \|U\|_{2,\infty}^2 \le R, \ \|V\|_{2,\infty}^2 \le R, \ \max_{(k,\ell) \in [d_1] \times [d_2]} |U_k^\intercal V_\ell| \le \alpha, \end{aligned} \tag{4.1}$$

where $\{(i_1, j_1), \ldots, (i_n, j_n)\} \subseteq ([d_1] \times [d_2])^n$ is a training set of row-column indices, U_i and V_j denote the ith row of U and the jth row of V, respectively. This problem, however, is non-convex since it involves a constraint on all product factorizations $UV^\intercal$. When the size of the problem k is large enough, Burer and Choi (2006) proved that this reformulated problem has no local minima. To solve this problem fast and efficiently, Lee et al. (2010) suggested the following first-order method.

Notice that $f(M) = \widehat{\mathcal{L}}_n(M; Y)$ is a smooth function $\mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}$. The projected gradient descent method generates a sequence of iterates $\{(U^t, V^t), t = 0, 1, 2, \ldots\}$ by the recursion: First, define an intermediate iterate

$$\begin{bmatrix} \widetilde{U}^{t+1} \\ \widetilde{V}^{t+1} \end{bmatrix} = \begin{bmatrix} U^t - \frac{\tau}{\sqrt{t}} \cdot \nabla f(U^t (V^t)^\intercal; Y) V^t \\ V^t - \frac{\tau}{\sqrt{t}} \cdot \nabla f(U^t (V^t)^\intercal; Y)^\intercal U^t \end{bmatrix} \quad \text{for } t = 0, 1, 2, \ldots,$$

where $\tau > 0$ is a stepsize parameter. If $\|\widetilde{U}^{t+1}(\widetilde{V}^{t+1})^{\intercal}\|_{\infty} > \alpha$, we replace

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

otherwise we keep it still. Next, compute updates according to

$$\begin{bmatrix} U^{t+1} \\ V^{t+1} \end{bmatrix} = \Pi_R \left(\begin{bmatrix} \widetilde{U}^{t+1} \\ \widetilde{V}^{t+1} \end{bmatrix} \right),$$

where Π_R is the Euclidean projection onto $\{(U, V) : \|U\|_{2,\infty}^2 \vee \|V\|_{2,\infty}^2 \leq R\}$. This 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.

4.2. An alternating direction method of multipliers based approach

The first-order algorithm described in Section 4.1 is computationally efficient and fast. However, (4.1) is in principle a non-convex optimization problem and thus the algorithm is only guaranteed to find a stationary point. Recently, an alternating direction method of multipliers (ADMM) based approached was proposed by Fang et al. (2015b) to solve the convex program (3.5) efficiently with strong theoretical guarantee. Furthermore, it was shown in Fang et al. (2015a) that the worst-case rate of convergence of the ADMM method is of order $1/t$, where t denotes the iteration counter. We briefly summarize this ADMM approach here for the sake of readability.

Define the class of matrices

$$\mathcal{P} = \big\{W \in \mathcal{S}^d : \mathrm{diag}(W) \geq 0,\ \|W_{11}\|_{\infty} \leq R,\ \|W_{22}\|_{\infty} \leq R,\ \|W_{12}\|_{\infty} \leq \alpha\big\},$$

where $d = d_1 + d_2$, $\mathcal{S}^d$ denotes the class of all symmetric matrices in $\mathbb{R}^{d \times d}$ and for every $W \in \mathcal{S}^d$, we write

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{12}^{\intercal} & W_{22} \end{pmatrix} \text{ with } W_{11} \in \mathbb{R}^{d_1 \times d_2},\ W_{22} \in \mathbb{R}^{d_2 \times d_2} \text{ and } W_{12} \in \mathbb{R}^{d_1 \times d_2}.$$

In this notation, the problem (4.1) can be equivalently formulated as

$$\min_{W, X \in \mathbb{R}^{d \times d}} f(W_{12}) \quad \text{s.t.} \quad W \in \mathcal{P},\ \ X \succeq 0,\ \ W - X = 0, \tag{4.2}$$

where as before, the function $f : \mathbb{R}^{d_1 \times d_2} \mapsto \mathbb{R}$ is given by $f(M) = \widehat{L}_n(M; Y)$. As pointed out by Fang et al. (2015b), the rationale of reformulating the problem into (4.2) is to divide the complexity of the feasible set in (4.1), which consists of a positive semidefinite constraint and ℓ_{∞}-norm constraints, into two parts. Then, by using an iterative method, we only need to control the ℓ_{∞}-norm of W

and project X into the positive semidefinite cone in each step. The additional constraint $W - X = 0$ ensures the feasibility of both W and X.

More specifically, consider the augmented Lagrangian function of (4.2) that is given by

$$F(W, X, Z) = f(W_{12}) + 2\langle W - X, Z\rangle + \rho\|W - X\|_F^2$$

for $W \in \mathcal{P}$ and $X \in \mathcal{S}_+^d = \{S \in \mathcal{S}^d : S \succeq 0\}$, where Z denotes the dual variable and $\rho > 0$ is prespecified. The ADMM is used to solve (4.2) iteratively as follows: Initialize (W^0, X^0, Z^0) and $\rho > 0$; at the $(t+1)$-th iteration, update (W, X, Z) according to

$$\begin{aligned} X^{t+1} &= \arg\min_X F(W^t, X^t, Z^t) = \Pi_{\mathcal{S}_+^d}(W^t - \rho^{-1}Z^t), \\ W^{t+1} &= \arg\min_{W\in\mathcal{P}} \left\{ f(W_{12}) + \rho\|W - X^{t+1} - \rho^{-1}Z^t\|_F^2 \right\}, \\ Z^{t+1} &= Z^t + \rho(X^{t+1} - W^{t+1}), \end{aligned} \tag{4.3}$$

where $\Pi_{\mathcal{S}_+^d} : \mathbb{R}^{d\times d} \mapsto \mathcal{S}_+^d$ is the map that projects a matrix into the semidefinite cone $\mathcal{S}_+^d$. The second step of (4.3) has an explicit solution given by (Fang et al., 2015b)

$$W_{k\ell}^{t+1} = \begin{cases} \Pi_{[-\alpha,\alpha]}(\rho^{-1}Y_{k\ell} + W_{k\ell}^t), & \text{if } (k,\ell) \in ([d_1]\times[d]\setminus[d_1]) \cap S, \\ \Pi_{[-\alpha,\alpha]}(W_{k\ell}^t), & \text{if } (k,\ell) \in ([d_1]\times[d]\setminus[d_1]) \setminus S, \\ \Pi_{[-R,R]}(W_{k\ell}^t), & \text{if } (k,\ell) \in [d_1]\times[d_1],\ k \neq \ell, \\ \Pi_{[0,R]}(W_{k\ell}^t), & \text{if } (k,\ell) \in [d_1]\times[d_1],\ k = \ell, \\ \Pi_{[-R,R]}(W_{k\ell}^t), & \text{if } (k,\ell) \in [d]\setminus[d_1]\times[d]\setminus[d_1],\ k \neq \ell, \\ \Pi_{[0,R]}(W_{k\ell}^t), & \text{if } (k,\ell) \in [d]\setminus[d_1]\times[d]\setminus[d_1],\ k = \ell, \end{cases}$$

where $S = \{(i_t, j_t)\}_{t=1}^n$ is the index set of observed entries and $\Pi_{[a,b]}(x) = \min\{b, \max(a, x)\}$ is the projection function from $\mathbb{R}$ to $[a, b]$.

4.3. Implementation

Before the max-norm constraint approach can be actually implemented in practice to generate a full matrix by filling in missing entries, additional prior knowledge of the unknown true matrix is needed to avoid deviated results. As before, let $M^* \in \mathbb{R}^{d_1\times d_2}$ be the true underlying matrix. Suitable upper bounds on the following key quantities are needed in advance:

$$\alpha_0 = \|M^*\|_\infty, \quad R_0 = \|M^*\|_{\max} \quad \text{and} \quad r_0 = \operatorname{rank}(M^*). \tag{4.1}$$

In order to estimate R_0 directly from a missing data matrix, it can be seen from (2.2) that $\alpha_0\sqrt{r_0}$ is a sharp upper bound on R_0 and is more amenable to estimation. Fortunately, it is possible to convincingly specify α_0 beforehand in many real-life applications. When dealing with the Netflix data, for instance,

α_0 can be chosen as the highest rating index; in the structure-from-motion problem, α_0 depends on the range of the camera field of view, which in most cases is sufficiently large to capture the feature point trajectories. In case where the percentage of missing entries is low, the largest magnitude of the observed entries can be used as an alternative for α_0.

As for r_0, we recommend the rank estimation approach recently developed in Juliá et al. (2011), which was shown to be effective in computer vision problems. Recall that in the structure-from-motion problem, each column of the data matrix corresponds a trajectory along the frames of a given feature point, and can be regarded as a signal vector with missing coordinates. Due to the rigidity of the moving objects, it was noted in Juliá et al. (2011) that the behavior of observed and missing data is the same and thus they both generate an analogous (frequency) spectral representation. Motivated by this observation, the proposed approach is based on the study of changes in frequency spectra on the initial matrix after missing entries are recovered.

In general, choosing the tuning parameter $R > 0$ in (3.5) adaptively is a difficult problem. In the regression case, it can be done by the Scaled LASSO method (Sun and Zhang, 2012). It is unclear whether a similar approach would work for matrix completion problems. By convexity and strong duality, the optimization program in (3.5) is equivalent to

$$\min_{M\in\mathbb{R}^{d_1\times d_2}:\|M\|_\infty\le\alpha}\left\{\frac{1}{n}\sum_{t=1}^{n}(Y_{i_tj_t}-M_{i_tj_t})^2+\lambda\|M\|_{\max}\right\} \tag{4.2}$$

for a properly chosen λ. In fact, for any $R > 0$ specified in (3.5), there exists a $\lambda > 0$ such that the solutions to the two problems (3.5) and (4.2) coincide. In practice, we suggest to solve (4.2) using the ADMM method described in Section 4.2 with λ obtained via cross-validation, in a way similarly to that for LASSO or the trace-norm penalized M-estimator studied in Negahban and Wainwright (2011).

Next we describe an implementation of the max-norm constrained matrix completion procedure, which incorporates the rank estimation approach in Juliá et al. (2011). Assume without loss of generality that α_0 is known.

(1) Given the observed partial matrix M_S, the initial matrix M_{ini} is obtained by adding the average of the corresponding column to the missing entries of M_S. Applying the Fast Fourier Transform (FFT) to the columns of M_{ini} and taking its modulus, i.e. $F := |\text{FFT}(M_{\text{ini}})|$.
(2) Set an initial rank $r = 2$ and an upper bound $r_{\max}$. Clearly, $r_{\max} \le \min(d_1, d_2)$ and it can be computed automatically by adding a criteria for stopping the iteration.
(3) For the current value of r, using the computational algorithms given in Section 4 with $R = \alpha_0\sqrt{r}$ to solve the max-norm constraint optimization (3.5). The resulting estimated full matrix is denoted by $\widehat{M}_r$.
(4) Apply the FFT to $\widehat{M}_r$ as in step 1. Write $F_r = |\text{FFT}(\widehat{M}_r)|$ and compute the error $e(r) = \|F - F_r\|_F$.
(5) If $r < r_{\max}$, set $r = r + 1$ and go to step 3.

Finally, let

$$r^* = \underset{2\le r\le r_{\max}}{\arg\min}\ e(r)$$

and the corresponding $\widehat{M}_{r^*}$ is the final estimate of M^*. Clearly, the above procedure can be modified by replacing the rank r with the max-norm R. A suitable initial value for the max-norm is $R = \alpha_0\sqrt{2}$ and at each iteration, increase $R = R + \delta$ with a fixed step size $\delta > 0$. An upper-bound $R_{\max}$ could be automatically computed by adding some criteria for stopping the iteration.

5. Discussions

This paper considers the approximate recovery of approximately low-rank matrices, in particular low-max-norm matrices in contrary to low-trace-norm matrices. The max-norm ball with radius 1 is nearly equivalent to the convex hull of rank-1 matrices, and therefore is an alternative convex surrogate for the rank. A max-norm constrained empirical risk minimization method is proposed and its theoretical properties are studied along with computational algorithms. Allowing for *unknown non-uniform sampling* which is an important relaxation of the uniform assumption in practice, it is shown that the method is rate-optimal and can be solved efficiently in polynomial time.

When the underlying matrix has exactly rank r, it is known that using the trace-norm based approach leads to a mean square error of order $O\{rd(\log d)/n\}$ (Keshavan and Montanari, 2010; Koltchinskii, Lounici and Tsybakov, 2011; Negahban and Wainwright, 2012; Klopp, 2014), where $d = d_1 + d_2$. In the ideal uniform sampling model, the trace-norm regularized method is arguably the mostly preferable one as it achieves optimal rate of convergence (up to a logarithmic factor) and is computationally feasible. The sampling scheme considered in this paper is unspecified and is allowed to be highly non-uniform, which brings additional randomness and uncertainty to the recovery problem. Therefore, we are essentially dealing with a much more complex model, and the max-norm constraint is not only introduced as a convex relaxation for low-rankness according to (2.2) but also takes into account the effect of non-uniform sampling.

6. Proofs

We prove the main results, Theorems 3.1 and 3.3, in this section. The proofs of a few key technical lemmas including Lemma 3.1 are also given.

6.1. Proof of Theorem 3.1

For ease of exposition, we write $\widehat{M} = \widehat{M}_{\max}$ as long as there is no ambiguity. To illustrate the main idea, we first consider the case where $\xi_1, \ldots, \xi_n$ are i.i.d. normal random variables and prove that there exists an absolute constant C

such that for any $t \in (0,1)$ and a sample size n satisfying $2 < n \le d_1 d_2$,

$$\|\widehat{M}_{\max} - M^*\|_\Pi^2 \le C\left\{(\alpha \vee \sigma)R\sqrt{\frac{d}{n}} + \alpha^2 \frac{\log(2/t)}{n}\right\} \tag{6.1}$$

holds with probability greater than $1 - t - e^{-d}$. The case of sub-exponential noise can be obtained via a straightforward adaptation of the arguments for Gaussian noise.

To begin with, noting that $\widehat{M}$ is optimal and M^* is feasible for the convex optimization problem (3.5), we thus have the basic inequality that

$$\frac{1}{n}\sum_{t=1}^n (Y_{i_t j_t} - \widehat{M}_{i_t j_t})^2 \le \frac{1}{n}\sum_{t=1}^n (Y_{i_t j_t} - M^*_{i_t j_t})^2.$$

This, combined with our model assumption $Y_{i_t j_t} = M^*_{i_t j_t} + \sigma\xi_t$ yields that

$$\frac{1}{n}\sum_{t=1}^n \widehat{\Delta}^2_{i_t j_t} = \frac{1}{n}\sum_{t=1}^n (\widehat{M}_{i_t j_t} - M^*_{i_t j_t})^2 \le \frac{2\sigma}{n}\sum_{t=1}^n \xi_t \widehat{\Delta}_{i_t j_t}, \tag{6.2}$$

where $\widehat{\Delta} = \widehat{M} - M^* \in \mathcal{K}(2\alpha, 2R)$ is the error matrix. By (6.2), the major challenges in proving Theorem 3.1 consist of two parts, bounding the left-hand side of (6.2) from below in a uniform sense and the right-hand side of (6.2) from above.

Step 1. (Upper bound). Recalling that $\{\xi_t\}_{t=1}^n$ is a sequence of $N(0,1)$ random variables and that $S = \{(i_1, j_1), \ldots, (i_n, j_n)\}$ is drawn i.i.d. according to Π on $[d_1] \times [d_2]$, we define

$$\widehat{\mathcal{R}}_n(\alpha, R) := \sup_{M \in \mathcal{K}(\alpha,R)} \left|\frac{1}{n}\sum_{t=1}^n \xi_t M_{i_t j_t}\right|. \tag{6.3}$$

Due to Pisier (1989), we obtain that for any realization of the training set S and for any $\delta > 0$, with probability at least $1 - \delta$ over $\boldsymbol{\xi} = \{\xi_t\}_{t=1}^n$,

$$\begin{aligned}
&\sup_{M \in \mathcal{K}(\alpha,R)} \left|\frac{1}{n}\sum_{t=1}^n \xi_t M_{i_t j_t}\right| \\
&\le \mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{M \in \mathcal{K}(\alpha,R)} \left|\frac{1}{n}\sum_{t=1}^n \xi_t M_{i_t j_t}\right|\right\} + \pi\sqrt{\frac{\log(1/\delta)\sup_{M \in \mathcal{K}(\alpha,R)} \sum_{t=1}^n M^2_{i_t j_t}}{2n^2}} \\
&\le \mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{M \in \mathcal{K}(\alpha,R)} \left|\frac{1}{n}\sum_{t=1}^n \xi_t M_{i_t j_t}\right|\right\} + \pi(\alpha \wedge R)\sqrt{\frac{\log(1/\delta)}{2n}}.
\end{aligned} \tag{6.4}$$

Thus it remains to estimate the following expectation over the class of matrices $\mathcal{K}(\alpha, R)$:

$$\mathcal{R}_n := \mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{M \in \mathcal{K}(\alpha,R)} \left|\frac{1}{n}\sum_{t=1}^n \xi_t M_{i_t j_t}\right|\right\}.$$

As a direct consequence of (2.3), we have

$$\mathcal{R}_n \leq K_G \cdot R \cdot \mathbb{E}_{\boldsymbol{\xi}} \left(\max_{M \in \mathcal{M}_\pm} \left| \frac{1}{n} \sum_{t=1}^{n} \xi_t M_{i_t j_t} \right| \right), \tag{6.5}$$

where $\mathcal{M}_\pm$ contains rank-one sign matrices with cardinality $|\mathcal{M}_\pm| = 2^{d-1}$. For each $M \in \mathcal{M}_\pm$, $\sum_{t=1}^{n} \xi_t M_{i_t j_t}$ is a Gaussian random variable with mean zero and variance n. Then, the expectation of the Gaussian maximum in (6.5) can be bounded by

$$2\sqrt{n \log(|\mathcal{M}_\pm|)} \leq 2\sqrt{\log 2}\sqrt{nd}.$$

Substituting this into (6.5) gives

$$\mathcal{R}_n \leq 2K_G\sqrt{\log 2} \cdot R\sqrt{nd}.$$

Since this upper bound holds uniformly over all realizations of S, we conclude that with probability at least $1 - \delta$ over both the random samples S and the noise $\boldsymbol{\xi} = \{\xi_t\}_{t=1}^n$,

$$\widehat{\mathcal{R}}_n(\alpha, R) \leq 3 \left\{ R\sqrt{\frac{d}{n}} + (\alpha \wedge R)\sqrt{\frac{\log(1/\delta)}{n}} \right\}. \tag{6.6}$$

In the case of sub-exponential noise, i.e. $\{\xi_t\}_{t=1}^n$ satisfies the assumption (3.6), it follows from (2.3) that

$$\widehat{\mathcal{R}}_n(\alpha, R) \leq K_G \cdot R \cdot \sup_{M \in \mathcal{M}_\pm} \left| \frac{1}{n} \sum_{t=1}^{n} \xi_t M_{i_t j_t} \right| \quad \text{with} \quad |\mathcal{M}_\pm| = 2^{d-1}.$$

For any realization of the training set $S = \{(i_1, j_1), \ldots, (i_n, j_n)\}$ and for any $M \in \mathcal{M}_\pm$ fixed, it follows from a Bernstein-type inequality for sub-exponential random variables (Vershynin, 2012) that

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{t=1}^{n} \xi_t M_{i_t j_t} \right| \geq t \right) \leq 2 \exp\left\{ -c \cdot \min\left(\frac{nt^2}{K^2}, \frac{nt}{K} \right) \right\},$$

where $c > 0$ is an absolute constant. By the union bound, it can be easily verified that for a sample size $n \geq d$,

$$\widehat{\mathcal{R}}_n(\alpha, R) \leq CKR\sqrt{\frac{d}{n}} \tag{6.7}$$

holds with probability at least $1 - e^{-d}$ for some absolute constant $C > 0$.

Step 2. (Lower bound). For the given sampling distribution Π, note that

$$\|M\|_\Pi^2 = \sum_{k,\ell} \pi_{k\ell} M_{k\ell}^2 = \frac{1}{|S|} \mathbb{E}_{S \sim \Pi} \|M_S\|_2^2,$$

where $M_S = (M_{i_1 j_1}, \ldots, M_{i_n j_n})^\intercal \in \mathbb{R}^n$ for any training set $S = \{(i_t, j_t)\}_{t=1}^n$ of size n. For $\beta \geq 1$ and $\delta > 0$, consider the following subset

$$\mathcal{C}(\beta, \delta) := \big\{M \in \mathcal{K}(1, \beta) : \|M\|_\Pi^2 \geq \delta\big\}.$$

Here, δ can be regarded as a tolerance parameter. The goal is to show that there exists some function f_β such that with high probability, the following inequality

$$\frac{1}{n}\|M_S\|_2^2 \geq \frac{1}{2}\|M\|_\Pi^2 - f_\beta(n, d_1, d_2) \tag{6.8}$$

holds uniformly over $M \in \mathcal{C}(\beta, \delta)$.

Proof of (6.8). Instead, we will prove a stronger result that with exponentially high probability,

$$\left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right| \leq \frac{1}{2}\|M\|_\Pi^2 + f_\beta(n, d_1, d_2)$$

holds for all $M \in \mathcal{C}(\beta, \delta)$, based on a straightforward adaptation of the peeling argument used in Negahban and Wainwright (2012). Taking $\varrho = \frac{3}{2}$, define a sequence of subsets

$$\mathcal{C}_\ell(\beta, \delta) := \big\{M \in \mathcal{C}(\beta, \delta) : \varrho^{\ell-1}\delta \leq \|M\|_\Pi^2 \leq \varrho^\ell \delta\big\}$$

for $\ell = 1, 2, \ldots,$ and for any radius $D > 0$, set

$$\mathcal{B}(D) := \big\{M \in \mathcal{C}(\beta, \delta) : \|M\|_\Pi^2 \leq D\big\}. \tag{6.9}$$

In fact, if there exists some $M \in \mathcal{C}(\beta, \delta)$ satisfying

$$\left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right| > \frac{1}{2}\|M\|_\Pi^2 + f_\beta(n, d_1, d_2),$$

then there corresponds an $\ell \geq 1$ such that, $M \in \mathcal{C}_\ell(\beta, \delta) \subseteq \mathcal{B}(\varrho^\ell \delta)$ and

$$\left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right| > \frac{1}{3}\varrho^\ell \delta + f_\beta(n, d_1, d_2).$$

Therefore, the main task is to show that the latter event occurs with small probability. To this end, define the maximum deviation for each $S \subseteq ([d_1] \times [d_2])^n$ that

$$\Delta_D(S) = \sup_{M \in \mathcal{B}(D)} \left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right|. \tag{6.10}$$

The following lemma shows that $n^{-1}\|M_S\|_2^2$ does not deviate far from its expectation *uniformly* for all $M \in \mathcal{B}(D)$.

Lemma 6.1 (Concentration)**.** *There exists a universal positive constant C_1 such that, for any $D > 0$,*

$$\mathbb{P}\left\{\Delta_D(S) > \frac{D}{3} + C_1\beta\sqrt{\frac{d}{n}}\right\} \leq e^{-nD/26}. \tag{6.11}$$

In view of the above lemma, we take $f_\beta(n, d_1, d_2) = C_1\beta\sqrt{d/n}$ and consider the following sequence of events

$$\mathcal{E}_\ell = \left\{\Delta_{\varrho^\ell\delta}(S) > \tfrac{1}{3}\varrho^\ell\delta + f_\beta(n, d_1, d_2)\right\} \quad \text{for } \ell = 1, 2, \ldots.$$

Because $\mathcal{C}(\beta, \delta) = \cup_{\ell\geq 1}\mathcal{C}_\ell(\beta, \delta)$, using the union bound we have

$$\begin{aligned}
&\mathbb{P}\left\{\exists M \in \mathcal{C}(\beta,\delta),\ \text{s.t.}\ \left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right| > \frac{1}{2}\|M\|_\Pi^2 + f_\beta(n, d_1, d_2)\right\} \\
&\leq \sum_{\ell=1}^{\infty}\mathbb{P}\left\{\exists M \in \mathcal{C}_\ell(\beta,\delta),\ \text{s.t.}\ \left|\frac{1}{n}\|M_S\|_2^2 - \|M\|_\Pi^2\right| > \frac{1}{2}\|M\|_\Pi^2 + f_\beta(n, d_1, d_2)\right\} \\
&\leq \sum_{\ell=1}^{\infty} P\left(\mathcal{E}_\ell^c\right) \\
&\leq \sum_{\ell=1}^{\infty}\exp(-n\varrho^\ell\delta/26) \\
&\leq \sum_{\ell=1}^{\infty}\exp\{-\log(\varrho)\ell n\delta/26\} \leq \frac{\exp(-c_0 n\delta)}{1-\exp(-c_0 n\delta)} \qquad (6.12)
\end{aligned}$$

with $c_0 = \log(3/2)/26$, where we used the elementary inequality that

$$\varrho^\ell = \exp\{\ell\log(\varrho)\} \geq \ell\log(\varrho).$$

Consequently, for a sample size $n \leq d_1 d_2$ satisfying $\exp(-c_0 n\delta) \leq \frac{1}{2}$, or equivalently, $n > (c_0\delta)^{-1}\log 2$, we obtain that with probability greater than $1 - 2\exp(-c_0 n\delta)$,

$$\frac{1}{n}\|M_S\|_2^2 \geq \frac{1}{2}\|M\|_\Pi^2 - C_1\beta\sqrt{\frac{d}{n}} \qquad (6.13)$$

holds for all $M \in \mathcal{C}(\beta, \delta)$.

Step 3. Now we combine the results in *Step 1* and *Step 2* to finish the proof. On one hand, it follows from (6.6) that for a sample size $2 < n \leq d_1 d_2$,

$$\frac{1}{n}\sum_{t=1}^{n}\xi_t\widehat{\Delta}(i_t, j_t) \leq \widehat{\mathcal{R}}_n(2\alpha, 2R) \leq 12R\sqrt{\frac{d}{n}}$$

holds with probability at least $1 - e^{-d}$. On the other hand, set $\widetilde{\Delta} = \widehat{\Delta}/(2\alpha)$ such that $\|\widetilde{\Delta}\|_\infty \leq 1$ and $\|\widetilde{\Delta}\|_{\max} \leq R/\alpha := \beta$, or equivalently, $\widetilde{\Delta} \in \mathcal{K}(1, \beta)$. For any $0 < t < 1$, applying (6.13) with $\delta = \frac{\log(2/t)}{c_0 n}$ implies that for a sample size n with $2 < n \leq d_1 d_2$,

$$\|\widetilde{\Delta}\|_\Pi^2 \leq \max\left\{\frac{\log(2/t)}{c_0 n}, \frac{2}{n}\|\widetilde{\Delta}_S\|_2^2 + 2\beta C_1\sqrt{\frac{d}{n}}\right\}$$

holds with probability at least $1 - t$. The last two displays, joint with the basic inequality (6.2) lead to the final conclusion (6.1) after a simple rescaling. Similarly, using the upper bound (6.7), instead of (6.6), together with the lower bound (6.13) proves (3.7) in the case of sub-exponential noise. □

6.1.1. Proof of Lemma 6.1

Here, we prove the concentration inequality given in Lemma 6.1. The argument is based on some basic techniques of probability in Banach spaces, including symmetrization, contraction inequality and Bousquet's version of Talagrand concentration inequality as well as the upper bound (2.4) on the empirical Rademacher complexity of the max-norm ball.

Regarding the matrix $M \in \mathbb{R}^{d_1 \times d_2}$ as a function: $[d_1] \times [d_2] \mapsto \mathbb{R}$, i.e. $M(k,\ell) = M_{k\ell}$, we are interested in the following empirical process indexed by $\mathcal{B}(D)$:

$$\Delta_D(S) = \sup_{f_M : M \in \mathcal{B}(D)} \left| \frac{1}{n} \sum_{t=1}^{n} f_M(i_t, j_t) - \mathbb{E}\{f_M(i_t, j_t)\} \right| \quad \text{with} \quad f_M(\cdot) = \{M(\cdot)\}^2.$$

Recall that $|M_{k\ell}| \le \|M\|_\infty \le 1$ for all pairs (k,ℓ), we have

$$\sup_{M \in \mathcal{B}(D)} \mathrm{Var}\{f_M(i_1, j_1)\} \le \sup_{M \in \mathcal{B}(D)} \|M\|_\infty^2 \|M\|_\Pi^2 \le D.$$

We first bound $\mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\}$, and then show that $\Delta_D(S)$ is concentrated around its expectation. A standard symmetrization argument Ledoux and Talagrand (1991) yields

$$\mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\} \le 2\mathbb{E}_{S\sim\Pi}\left[\mathbb{E}_\varepsilon \left\{ \sup_{M \in \mathcal{B}(D)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i M_{i_t j_t}^2 \right| \right\} \right],$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence, independent of S. Given an index set $S = \{(i_1, j_1), \ldots, (i_n, j_n)\}$, since $|M_{i_t j_t}| \le 1$, using Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) implies that for $d = d_1 + d_2$,

$$\begin{aligned} &\mathbb{E}_\varepsilon \left\{ \sup_{M \in \mathcal{B}(D)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i M_{i_t j_t}^2 \right| \right\} \le 4\mathbb{E}_\varepsilon \left\{ \sup_{M \in \mathcal{B}(D)} \left| \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t M_{i_t j_t} \right| \right\} \\ &\le 4\mathbb{E}_\varepsilon \left(\sup_{\|M\|_{\max} \le \beta} \left| \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t M_{i_t j_t} \right| \right) \le 48\beta \sqrt{\frac{d}{n}}, \end{aligned}$$

where we used inequality (2.4) in the last step. Since the "worst-case" Rademacher complexity is uniformly bounded, we have

$$\mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\} \le 96\beta \sqrt{\frac{d}{n}}. \tag{6.14}$$

Next, applying Bousquet's version of Talagrand's concentration inequality for empirical processes indexed by bounded functions (Bousquet, 2003) yields that for every $t > 0$,

$$\Delta_D(S) \le \mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\} + \sqrt{\frac{2tD}{n} + \mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\}\frac{4t}{n}} + \frac{t}{3n}$$

$$\le \mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\} + 2\sqrt{\mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\}\frac{t}{n}} + \sqrt{\frac{2tD}{n}} + \frac{t}{3n}$$
$$\le 2\mathbb{E}_{S\sim\Pi}\{\Delta_D(S)\} + \sqrt{\frac{2tD}{n}} + \frac{4t}{3n}$$

with probability at least $1 - e^{-t}$. The conclusion (6.11) thus follows by taking $t = nD/26$. □

6.2. Proof of Theorem 3.2

The proof is based on a general result in Srebro, Sridharan and Tewari (2010) on excess risk bounds for learning with a smooth loss. Recall that the noisy response is of the form $Y_{i_t j_t} = M^*_{i_t j_t} + \xi_t$ for $t = 1, 2, \ldots$, where the location (i_t, j_t) of the entry is drawn from $[d_1] \times [d_2]$ according to Π and the noise ξ_t on the entry is drawn independently each time. For every $d_1 \times d_2$ matrix M, define the quadratic loss function

$$\mathcal{L}(M) = \mathbb{E}_{\substack{(i_t,j_t)\sim\Pi \\ \xi_t\sim N(0,1)}} (M - Y)^2_{i_t j_t}$$
$$= \mathbb{E}_{\substack{(i_t,j_t)\sim\Pi \\ \xi_t\sim N(0,1)}} \{(M^* - M)_{i_t j_t} + \xi_t\}^2 = \|M - M^*\|^2_\Pi + \sigma^2,$$

and its empirical counterpart $\widehat{\mathcal{L}}(M) = \frac{1}{n}\sum_{t=1}^n (M_{i_t j_t} - Y_{i_t j_t})^2 + \sigma^2$ for a given i.i.d. sample $\{(i_t, j_t), Y_{i_t j_t} = M^*_{i_t j_t} + \xi_t\}_{t=1}^n$. In this notation, our estimator $\widehat{M}_{\max}$ can be written as $\widehat{M}_{\max} = \arg\min_{M\in\mathcal{K}(\alpha,R)} \widehat{\mathcal{L}}(M)$.

In view of Definition 2.1, define the worst-case Rademacher complexity as

$$R_n(\mathcal{K}) = \sup_{\{(i_t,j_t)\}_{t=1}^n \in ([d_1]\times[d_2])^n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left\{ \sup_{M\in\mathcal{K}} \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i M(i_t, j_t) \right| \right\}$$
$$= \sup_{\{(i_t,j_t)\}_{t=1}^n \in ([d_1]\times[d_2])^n} \mathbb{E}_{\boldsymbol{\varepsilon}} \left(\sup_{M\in\mathcal{K}} \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i M_{i_t j_t} \right| \right),$$

where $\mathcal{K} = \mathcal{K}(\alpha, R)$.

For any $B > 0$, let $\mathcal{E}_B$ be the event that $\max_{1\le t\le n} |\xi_t| \le B$ holds. On $\mathcal{E}_B$, applying Theorem 1 in Srebro, Sridharan and Tewari (2010) by taking $H = 2$ and $b = 5\alpha^2 + 4\alpha\sigma B$ that, for any $0 < \delta < 1$,

$$\mathcal{L}(\widehat{M}_{\max}) - \min_{M\in\mathcal{K}(\alpha,R)} \mathcal{L}(M)$$
$$\le C_1 \left[\sqrt{\min_{M\in\mathcal{K}(\alpha,R)} \mathcal{L}(M) \left\{ (\log n)^3 R_n^2(\mathcal{K}) + \frac{B\log(1/\delta)}{n} \right\}} \right.$$
$$\left. + (\log n)^3 R_n^2(\mathcal{K}) + \frac{B\log(1/\delta)}{n} \right]$$

holds with probability at least $1 - \delta$ over a random sample $\{(i_t, j_t)\}_{t=1}^n$ of size n, where $C_1 > 0$ is an absolute constant. By (2.4), the worst-case Rademacher

complexity $R_n(\mathcal{K})$ is bounded by $6R\sqrt{d/n}$. Moreover, note that $\min_{M\in\mathcal{K}(\alpha,R)} \mathcal{L}(M) = \mathcal{L}(M^*) = \sigma^2$ and

$$\mathcal{L}(\widehat{M}_{\max}) = \|\widehat{M}_{\max} - M^*\|_\Pi^2 + \sigma^2.$$

Putting the above calculations together, we obtain that on the event $\mathcal{E}_B$,

$$\begin{aligned}&\|\widehat{M}_{\max} - M^*\|_\Pi^2 \\ &\le C_2\,\sigma\left[\sqrt{\left\{(\log n)^3\frac{R^2 d}{n} + \frac{B\log(1/\delta)}{n}\right\}} + (\log n)^3\frac{R^2 d}{n} + \frac{B\log(1/\delta)}{n}\right] \end{aligned} \tag{6.15}$$

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

Finally, it follows from Borell's inequality that for every $t>0$,

$$\mathbb{P}\left\{\max_{1\le t\le n}|\xi_t| \ge \mathbb{E}\left(\max_{1\le t\le n}|\xi_t|\right) + t\right\} \le e^{-t^2/2}.$$

A standard result on Gaussian maximum gives $\mathbb{E}(\max_{1\le t\le n}|\xi_t|) \le 2\sqrt{\log n}$. Together with the last display, this implies that with probability at least $1-\delta$,

$$\max_{1\le t\le n}|\xi_t| \le 2\sqrt{\log n} + \sqrt{2\log(1/\delta)}. \tag{6.16}$$

In particular, taking $\delta = n^{-1}$ in both (6.15) and (6.16) proves (3.10). □

6.3. *Proof of Theorem 3.3*

By construction in Lemma 3.1, setting $\delta = \gamma\alpha\sqrt{d_1 d_2/2}$ we see that $\mathcal{M}$ is a δ-packing set of $\mathcal{K}(\alpha, R)$ in the Frobenius norm. Next, a standard argument (Yang and Barro, 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 specifically,

$$\inf_{\widehat{M}}\max_{M\in\mathcal{K}(\alpha,R)}\mathbb{E}\|\widehat{M} - M\|_F^2 \ge \frac{\delta^2}{4}\min_{\widetilde{M}}\mathbb{P}(\widetilde{M}\ne M^*),$$

where the random variable $M^* \in \mathbb{R}^{d_1\times d_2}$ is uniformly distributed over the packing set $\mathcal{M}$. Conditional on $S = \{(i_1,j_1),\dots,(i_n,j_n)\}$, a variant of Fano's inequality (Cover and Thomas, 1991) leads to the lower bound

$$\mathbb{P}(\widetilde{M}\ne M^*|S) \ge 1 - \frac{\binom{|\mathcal{M}|}{2}^{-1}\sum_{i\ne j}K(M^i\|M^j) + \log 2}{\log|\mathcal{M}|}, \tag{6.17}$$

where $K(M^i\|M^i)$ denotes the Kullback-Leibler divergence between distributions $(Y_S|M^i)$ and $(Y_S|M^j)$. For the observation model (3.1) with i.i.d. Gaussian noise, we have

$$K(M^i\|M^j) = \frac{1}{2\sigma^2}\sum_{t=1}^n (M^i - M^j)_{i_t j_t}^2$$

and

$$\mathbb{E}_{S\sim\Pi}\{K(M^i\|M^j)\} = \frac{n}{2\sigma^2}\|M^i - M^j\|_\Pi^2, \tag{6.18}$$

where $\|\cdot\|_\Pi$ is the weighted Frobenius norm as in (3.3). For any two distinct $M^i, M^j \in \mathcal{M}$, $\|M^i - M^j\|_F^2 \le 4d_1d_2\gamma^2$, which together with (6.17), (6.18) and the assumption $\max_{k,\ell}\pi_{k\ell} \le \frac{\mu}{d_1d_2}$ implies that

$$\begin{aligned}
&\mathbb{P}(\widetilde{M} \neq M^*)\\
&\ge 1 - \frac{\binom{|\mathcal{M}|}{2}^{-1}\sum_{i\neq j}\mathbb{E}_{S\sim\Pi}\{K(M^i\|M^j)\} + \log 2}{\log|\mathcal{M}|}\\
&\ge 1 - \frac{\frac{32\mu\gamma^4\alpha^2 n}{\sigma^2} + 12\gamma^2}{r(d_1\vee d_2)} \ge 1 - \frac{32\mu\gamma^4\alpha^2 n}{\sigma^2 r(d_1\vee d_2)} - \frac{12}{r(d_1\vee d_2)} \ge \frac{1}{2},
\end{aligned} \tag{6.19}$$

provided that $r(d_1\vee d_2)\ge 48$ and $\gamma^4 \le \frac{\sigma^2}{128\alpha^2}\frac{r(d_1\vee d_2)}{\mu n}$. If $\frac{\sigma^2}{128\alpha^2}\frac{r(d_1\vee d_2)}{\mu n} > 1$, we choose $\gamma = 1$ so that

$$\inf_{\widehat{M}}\max_{M\in\mathcal{K}(\alpha,r)}\frac{1}{d_1d_2}\mathbb{E}\|\widehat{M} - M\|_F^2 \ge \frac{\alpha^2}{16}.$$

Otherwise, as long as the parameters (n, d_1, d_2, α, R) satisfy (3.12), taking

$$\gamma^2 = \frac{\sigma}{8\sqrt{2}\,\alpha}\sqrt{\frac{r(d_1\vee d_2)}{\mu n}}$$

yields

$$\inf_{\widehat{M}}\max_{M\in\mathbb{B}_{\max}(R)}\frac{1}{d_1d_2}\mathbb{E}\|\widehat{M} - M\|_F^2 \ge \frac{\sigma\alpha}{128\sqrt{2}}\sqrt{\frac{r(d_1\vee d_2)}{\mu n}} \ge \frac{\sigma R}{256}\sqrt{\frac{d}{\mu n}},$$

as desired. □

6.4. Proof of Lemma 3.1

We proceed via a probabilistic method. Assume without loss of generality that $d_2 \ge d_1$. Let $N = \exp(\frac{rd_2}{16\gamma^2})$, $B = \frac{r}{\gamma^2}$, and for each $i = 1,\ldots,N$, we draw a random matrix $M^i \in \mathbb{R}^{d_1\times d_2}$ as follows: The matrix M^i consists of i.i.d. blocks of dimensions $B\times d_2$, stacked from top to bottom, with the entries of the first block being i.i.d. symmetric random variables taking values $\pm\alpha\gamma$, such that

$$M^i_{k\ell} := M^i_{k'\ell}, \quad k' = k(\mathrm{mod}\ B) + 1.$$

Next, we show that above random procedure succeeds in generating a set having all desired properties, with non-zero probability. For $1\le i\le N$, it is easy to see that

$$\|M^i\|_\infty = \alpha\gamma \le \alpha, \quad \frac{1}{d_1d_2}\|M^i\|_F^2 = \alpha^2\gamma^2$$

and because $\operatorname{rank}(M^i) \le B$,

$$\|M^i\|_{\max} \le \sqrt{B}\,\|M^i\|_\infty = \sqrt{\frac{r}{\gamma^2}}\,\alpha\gamma = \alpha\sqrt{r} = R.$$

Consequently, $M^i \in \mathcal{K}(\alpha, R)$ and it remains to show that the set $\{M^i\}_{i=1}^N$ satisfies property (ii). In fact, for any $1 \le i \ne j \le N$,

$$\begin{aligned}\|M^i - M^j\|_F^2 &= \sum_{k,\ell}(M^i_{k\ell} - M^j_{k\ell})^2 \\ &\ge \left\lfloor \frac{d_1}{B} \right\rfloor \sum_{k=1}^{B}\sum_{\ell=1}^{d_2}(M^i_{k\ell} - M^j_{k\ell})^2 = 4\alpha^2\gamma^2 \left\lfloor \frac{d_1}{B} \right\rfloor \sum_{k=1}^{B}\sum_{\ell=1}^{d_2}\delta_{kl},\end{aligned}$$

where δ_{kl} are independent 0/1 Bernoulli random variables with mean 1/2. Using Hoeffding's inequality gives

$$\mathbb{P}\left(\sum_{k=1}^{B}\sum_{\ell=1}^{d_2}\delta_{k\ell} \ge \frac{Bd_2}{4}\right) \le \exp(-Bd_2/8).$$

Because there are less than $N^2/2$ such index pairs in total, the above inequality, together with the union bound implies that with probability at least $1 - \frac{N^2}{2}\exp(-Bd_2/8) \ge 1/2$,

$$\|M^i - M^j\|_F^2 > \alpha^2\gamma^2 \left\lfloor \frac{d_1}{B} \right\rfloor Bd_2 \ge \frac{\alpha^2\gamma^2 d_1 d_2}{2}$$

holds for all $i \ne j$. This completes the proof of Lemma 3.1. □

Acknowledgements

We thank the editors and an anonymous referee for their careful reviews and constructive comments.

References

Argyriou, A., Evgeniou, T. and Pontil, M. (2008). Convex multi-task feature learning. *Mach. Learn.* **73**, 243–272.

Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482. MR1984026

Bousquet, O. (2003). Concentration inequalities for sub-addtive functions using the entropy method. In *Stochastic Inequalities and Applications. Progress in Probability* **56**, 213–247. Birkhäuser, Basel. MR2073435

Burer, S. and Choi, C. (2006). Computational enhancements in low-rank semidefinite programming. *Optim. Method Softw.* **21**, 493–512. MR2197509

Cai, T. T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.* **13**, 3619–3647. MR3159403

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

Candès, E. and Plan, Y. (2011). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory* **57**, 2342–2359. MR2809094

Candès, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**, 717–772. MR2565240

Candès, E. and Tao, T. (2010). The power of convex relaxations: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56**, 2053–2080. MR2723472

Chen, P. and Suter, D. (2004). Recovering the missing components in a large noisy low-rank matrix: Application to SFM. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 1051–1063.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* John Wiley and Sons, New York. MR1122806

Davenport, M. A., Plan, Y., van den Berg, E. and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA* **3**, 189–223. MR3311452

Fang, E. X., He, B., Liu, H. and Yuan, X. (2015a). Generalized alternating direction method of multipliers: New theoretical insights and applications. *Math. Prog. Comp.* **7**, 149–187. MR3347621

Fang, E. X., Liu, H., Toh, K.-C. and Zhou, W.-X. (2015b). Max-norm optimization for robust matrix recovery. Technical Report.

Foygel, R., Salakhutdinov, R., Shamir, R. and Srebro, N. (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in Neural Information Processing Systems* **24**, 2133–2141.

Foygel, R. and Srebro, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. *JMLR: Workshop and Conference Proceedings* **19**, 315–339.

Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Comm. ACM* **35**, 61–70. MR1279415

Green, P. and Wind, Y. (1973). *Multiattribute Decisions in Marketing: A Measurement Approach.* Dryden Press, Hinsdale, IL.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57**, 1548–1566. MR2815834

Jameson, G. J. O. (1987). *Summing and Nuclear Norms in Banach Space Theory.* London Mathematical Society Student Texts, 8. Cambridge University Press, Cambridge. MR0902804

Julià, C., Sappa, A. D., Lumbreras, F., Serrat, J. and López, A. (2011). Rank estimation in missing data matrix problems. *J. Math. Imaging Vis.* **39**, 140–160. MR2788516

Keshavan, R., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11**, 2057–2078. MR2678022

Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.* **5**, 1161–1183. MR2842903

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20**, 282–303. MR3160583

Koltchinskii, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.* **39**, 2936–2973. MR3012397

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Statist.* **39**, 2302–2329. MR2906869

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York. MR1102015

Lee, J., Recht, B., Salakhutdinov, R., Srebro, N. and Tropp, J. (2010). Practical large-scale optimization for max-norm regularization. *Advances in Neural Information Processing Systems* **23**, 1297–1305.

Lee, T., Shraibman, A. and Špalek, R. (2008). A direct product theorem for discrepancy. In *Proceedings of the 23rd Annual IEEE Conference on Computational Complexity*, 71–80.

Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M. and Ma, Y. (2009). Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Aruba, Dutch Antilles.

Linial, N., Mendelson, S., Schechtman, G. and Shraibman, A. (2004). Complexity measures of sign measures. *Combinatorica* **27**, 439–463. MR2359826

Liu, Z. and Vandenberghe, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* **31**, 1235–1256. MR2558821

Mazumber, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11**, 2287–2322. MR2719857

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 1069–1097. MR2816348

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13**, 1665–1697. MR2930649

Nesterov, Y. (2013). Gradient methods for minimizing composite objective function. *Math. Program.* **140**, 125–161. MR3071865

Pisier, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, Cambridge. MR1036275

Recht, B. (2011). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12**, 3413–3430. MR2877360

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 887–930. MR2816342

Salakhutdinov, R. and Srebro, N. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. *Advances in Neural Information Processing Systems* **23**, 2056–2064.

Singer, A. and Cucuringu, M. (2010). Uniqueness of low-rank matrix completion by rigidity theory. *SIAM J. Matrix Anal. Appl.* **31**, 1621–1641. MR2595541

Srebro, N., Rennie, J. and Jaakkola. T. (2004). Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems* **17**, 1329–1336.

Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory, Proceedings of COLT-2005. Lecture Notes in Comput. Sci.* **3559**, 545–560. Springer, Berlin. MR2203286

Srebro, N., Sridharan, K. and Tewari, A. (2010). Optimistic rates for learning with a smooth loss. *Advances in Neural Information Processing Systems* **23**, 2199–2207.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898. MR2999166

Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams under orthography: A factorization method. *Int. J. Comput. Vis.* **9**, 137–154.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12**, 389–434. MR2946459

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge University Press, Cambridge. MR2963170

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564–1599. MR1742500

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York. MR1462963