Journal of Machine Learning Research 17 (2016) 1-34 Submitted 6/15; Revised 12/15; Published 4/16

Probabilistic Low-Rank Matrix Completion from Quantized Measurements

Sonia A. Bhaskar SBHASKAR@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Editor: Benjamin Recht

Abstract

We consider the recovery of a low rank real-valued matrix M given a subset of noisy discrete (or quantized) measurements. Such problems arise in several applications such as collaborative filtering, learning and content analytics, and sensor network localization. We consider constrained maximum likelihood estimation of M, under a constraint on the entry-wise infinity-norm of M and an exact rank constraint. We provide upper bounds on the Frobenius norm of matrix estimation error under this model. Previous theoretical investigations have focused on binary (1-bit) quantizers, and been based on convex relaxation of the rank. Compared to the existing binary results, our performance upper bound has faster convergence rate with matrix dimensions when the fraction of revealed observations is fixed. We also propose a globally convergent optimization algorithm based on low rank factorization of M and validate the method on synthetic and real data, with improved performance over previous methods.

Keywords: constrained maximum likelihood, quantization, matrix completion, collaborative filtering, convex optimization

1. Introduction

Recovery of a low-rank matrix from a subset of its entries is known as the matrix completion problem. This problem arises in many applications, including collaborative filtering (Rennie and Srebro, 2005; Koren et al., 2009), sensor network localization (Shang et al., 2004; Karbasi and Oh, 2013), learning and content analytics (Lan et al., 2014c,b), rank aggregation (Gleich and Lim, 2011), and manifold learning (Tenenbaum et al., 2000; Saul and Roweis, 2003). In many of these applications, the entries of the matrix are not real-valued, but discrete or quantized, e.g., binary-valued or multiple-valued. For example, in the Netflix problem where a subset of the users' ratings is observed, the ratings take integer values between 1 and 5. Classical matrix completion has treated these values as real-valued with good results, however, performance improvement can be achieved when the observations are treated as discrete (Davenport et al., 2014; Lan et al., 2014a).

We consider the problem of completing a matrix from a subset of its entries, but instead of assuming the observed entries are real-valued, we observe subset of quantized measurements. These observations are related to the underlying matrix M via a probabilistic model, as follows. Given $M \in \mathbb{R}^{m\times n}$, a subset of indices $\Omega \subseteq [m] \times [n]$, and a twice differentiable function $f_\ell : \mathbb{R} \to [0,1]$, with $\ell \in [K]$, $K \geq 2$, we observe

$$Y_{ij} = \ell \text{ with probability } f_\ell(M_{ij}) \text{ for } (i,j) \in \Omega, \tag{1}$$

where $\sum_{\ell=1}^{K} f_\ell(M_{ij}) = 1$. One important application of this model is the K-level quantization of noisy $M_{ij} + Z_{ij}$, where Y_{ij} is given by (Lan et al., 2014a)

$$Y_{ij} = \mathcal{Q}(M_{ij} + Z_{ij}), \quad (i,j) \in \Omega, \tag{2}$$

where the noise matrix Z has i.i.d. entries with cumulative distribution function (CDF) $\Phi(z)$, and the function $\mathcal{Q} : \mathbb{R} \to [K]$ corresponds to a scalar quantizer that maps a real number to one of the K ordered labels according to

$$\mathcal{Q}(x) = \ell \text{ if } \omega_{\ell-1} < x \leq \omega_\ell, \ \ell \in [K], \tag{3}$$

where $\omega_0 < \omega_1 < \cdots < \omega_K$ are the quantization bin boundaries. We will take $\omega_0 = -\infty$ and $\omega_K = \infty$. This quantization model was first considered in McCullagh (1980) for regression applications.

It then follows that

$$\begin{aligned} f_\ell(M_{ij}) &= P(Y_{ij} = \ell | M_{ij}) \\ &= \Phi(\omega_\ell - M_{ij}) - \Phi(\omega_{\ell-1} - M_{ij}). \end{aligned} \tag{4}$$

This observation model arises in many applications. In connectivity-based sensor network localization (Shang et al., 2004; Karbasi and Oh, 2013), M is a matrix of distances between sensors, and Y_{ij} takes binary values based on whether sensor i and sensor j are within a specified radius of each other. In learning and content analytics (Lan et al., 2014c,b), M governs the learners' responses to questions, and in recommender systems (Rennie and Srebro, 2005; Koren et al., 2009), M can represent the true underlying preferences of users. The matrix Y is then the matrix of ratings $\ell \in [K]$, which may represent quantization of some underlying real-valued user preference. Hence, the model (1)-(3) accounts for finer ordering of users' true preferences which then are quantized to discrete values dictated by the rating system. It is known that Netflix uses such real-valued predictions to order movie recommendations when generating recommendations for a user.

The probabilistic model described in (1)-(3) was first introduced by Davenport et al. (2014) for the case of binary ($K = 2$), or 1-bit, observations and has been studied in depth in the literature. This case corresponds to (2) with $\omega_1 = 0$ when the quantization model is considered. Under the assumption that M is low-rank, Davenport et al. (2014) and Lafond et al. (2014) proposed a convex program using maximum likelihood estimation and a nuclear (or trace) norm to promote a low-rank solution. Both works present theoretical recovery guarantees for the estimate, with the latter improving the convergence rate of the upper bound on the error. In Cai and Zhou (2013), a constrained maximum likelihood estimator was also considered but with the max-norm in place of the nuclear norm. Upper and lower bounds on the error norm of the solution to the resulting convex program were also given of the same order as Davenport et al. (2014). The binary model is also investigated in Soni et al. (2014) for sparse factor models using maximum likelihood estimation with an exact low-rank constraint; their results apply to non-sparse models also.

The theoretical recovery guarantee for the estimate given in Soni et al. (2014) is in the form of an upper bound on the expectation of the error norm, in contrast to Davenport et al. (2014), Lafond et al. (2014) and Cai and Zhou (2013), where the (high probability) upper bounds on the error norm itself are given. The bounds presented in this paper are also on the error norm, not on its expectation.

The extension to multi-level observations ($K \geq 2$) was introduced in Lan et al. (2014a), with a focus on the quantized observation model given in (2). A constrained maximum likelihood estimator, similar to that of Davenport et al. (2014), was proposed and validated through numerical experiments, but no theoretical results were given. An extension to multi-level observations was also proposed in Lafond et al. (2014). In contrast to the quantization observation model of Lan et al. (2014a), which involves just one M, the observation model of Lafond et al. (2014) related the matrix Y of K level observations to a vector of $K-1$ underlying matrices $(M^j)_{j=1}^{K-1}$. An upper bound on the error norm was given for a penalized maximum likelihood estimate of this vector of matrices, of the same order as established for the binary case. Recently Cao and Xie (2015) also investigated matrix completion for categorical data and extended the results of Davenport et al. (2014) to multi-level observations. The error bounds of Cao and Xie (2015) are of the same order as that of Davenport et al. (2014) for the binary case. The multi-level observation model of Cao and Xie (2015) does not include the quantized observation model given in (2).

Generalized performance bounds for a generic regularized convex program with arbitrary regularizer were given in Gunasekar et al. (2014) and Lafond (2015), which can be applied to the observation model (1) for $K \geq 2$ when the the link function f comes from the exponential family. Hence, this theoretical guarantee does not apply when considering a quantization observation model, given in (2) when $K > 2$. In this paper, we will allow for an arbitrary log-concave link function in our observation model and theoretical results, which will allow for applications such as noisy K-level quantization.

Another line of work in the context of collaborative filtering has been concerned with probabilistic matrix factorization (PMF) models (Salakhutdinov and Mnih, 2008; Gopalan et al., 2014), some of which can handle integer-valued observations. In an item ratings context, for an $m \times n$ ratings matrix M, one writes $M = UV^\top$ where the factors $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$ represent latent users and item feature matrices. A Gaussian model for the observations, parametrized by these factors, is used in Salakhutdinov and Mnih (2008), and a Poisson model is used in Gopalan et al. (2014) allowing for integer-valued observations. The item and user feature vectors are assigned priors, and hyperparameters and feature vectors are estimated by maximizing the log-posterior in Salakhutdinov and Mnih (2008) and minimizing the Kullback-Leibler divergence in Gopalan et al. (2014). To our knowledge, there are no theoretical recovery guarantees regarding the performance of these PMF models. Also in the context of collaborative filtering, Koren and Sill (2011, 2013) proposed an ordinal model for predicting missing rating distributions from revealed multi-level numerical ratings. The model of Koren and Sill (2011, 2013) is a quantized observation model similar to that in Lan et al. (2014a), and as in Lan et al. (2014a), no theoretical results were given. Modeling of ordinal data with Gaussian restricted Boltzmann machines for both vector-variates and matrix-variates has been investigated in Tran et al. (2012), where a quantized observation model is also considered. No theoretical results were given Tran et al. (2012).

Aside from the PMF models, all prior works have considered convex programs which use a convex relaxation of matrix rank as a surrogate for promoting low rank. While this may be advantageous in cases where the matrix is approximately low rank and also because it results in a convex problem, often in applications the rank is known (as in sensor network localization), or can be reliably estimated. One question is, if we consider an exact rank constraint, can performance guarantees be proved and performance improvement be achieved, and can we find an algorithm to lead us to a global solution?

In this paper, we extend the theory of 1-bit matrix completion to that of multi-level discrete measurements, with an emphasis on the application to quantization. We consider maximum likelihood (ML) estimation of M from multi-level quantized observations, and establish upper bounds on the estimation error norm for this problem, which has a faster rate of convergence than previously established upper bounds for the binary case. In contrast to Gunasekar et al. (2014) and Lafond (2015), we do not restrict the likelihood (i.e., the link function f) to come from an exponential family distribution. We allow the likelihood to come from any strictly log-concave distribution, which includes distributions of bounded discrete random variables from the exponential family. We furthermore focus on the application to a quantization observation model similar to that of Lan et al. (2014a), where the likelihood is not from the exponential family when the number of levels is greater than two (that is, when $K > 2$). Rather than using a convex relaxation to promote a low-rank solution as in previous works, we use an exact rank constraint. We present several algorithms based on matrix factorization for solving our optimization problem, one of which is globally convergent. Our method outperforms some of the existing low-rank matrix completion methods on both synthetic and real world data.

In a preliminary short version of this paper (Bhaskar, 2015), we presented Algorithms 1 and 2 (for known bin boundaries) and stated a preliminary version of our performance upper bound without any proof. The present paper provides a more comprehensive upper bound of the Frobenius norm of the error with the complete proof, an additional algorithm, Algorithm 3, and more extensive numerical experiments which include validation on the MovieLens 1M dataset.

The paper is organized as follows. In Section 2 we discuss the assumptions on the target matrix to make it identifiable and also discuss our sampling model on which we follow Bhojanapalli and Jain (2014). In Section 3, we state our main results regarding theoretical guarantees on matrix recovery. We first describe the proposed constrained ML estimation of the target matrix M from multi-level quantized observations. We then establish upper bounds on the estimation error norm for this problem, which yield a faster rate of convergence than previously established upper bounds for the binary case. In Section 4 we present several algorithms based on matrix factorization for solving our optimization problem, one of which is globally convergent. We corroborate our results with numerical examples in Section 5 where we test our methods on synthetic and real data, and also compare our methods with that of Keshavan et al. (2009, 2010) (OptSpace), Cai et al. (2010) (SVT), Cai and Zhou (2013) and Davenport et al. (2014). Proofs of technical claims are given in the two appendices.

Notation: We use capital letters, such as M, to denote a matrix, and M_{ij} for its (i, j)th entry. We let $\|M\|_2$, $\|M\|_F$, $\|M\|_*$ and $\|M\|_\infty$ denote the operator, Frobenius, nuclear (or trace) and entry-wise infinity norm, respectively, of M. The notation $M^\top$ denotes the transpose of M, $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$, $[n]$ denotes the set of integers $\{1,\ldots,n\}$, $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all ones, $\tilde{\mathbf{1}}_n = \mathbf{1}_n/\sqrt{n}$, and $\mathbf{1}_{[A]}$ denotes the indicator function, i.e. $\mathbf{1}_{[A]} = 1$ when A is true, and $\mathbf{1}_{[A]} = 0$ otherwise. We use $\langle A, B\rangle$ to denote $\mathrm{tr}(A^\top B) = \sum_{ij} A_{ij}B_{ij}$. The abbreviations w.h.p. and w.p.1 stand for with high probability and with probability one, respectively.

2. Preliminaries and Model Assumptions

In this section, we discuss the assumptions on the target matrix M to make it identifiable. We also discuss our sampling model on which we follow Bhojanapalli and Jain (2014).

2.1 Low-rank Matrices

The problem of completing a matrix from a subset of its entries is ill-posed without imposing structural assumptions on the matrix. Hence, some relationship between the entries must be assumed to reconstruct M from a subset of its entries. The majority of the literature on matrix completion assumes that the matrix M to be recovered is low rank, i.e., that it spans a low-dimensional subspace. This assumption is reasonable in many applications.

We assume that M is a low-rank matrix with rank bounded by r. We note that in many applications, such as sensor network localization, where M is known to exist in 2 or 3-dimensional grid, or DNA haplotype assembly, the rank r is known. In examples such as collaborative filtering, where M is a matrix in which rows may represent users and columns may represent their preferences for an item, M is low rank since the users' preferences are believed to be a function of just a few factors. In applications where the rank r is not explicitly known, as in the former example, it can be reliably estimated (Keshavan et al., 2010).

We furthermore assume that the true matrix M satisfies $\|M\|_\infty \leq \alpha$, which helps make the recovery of M well-posed by preventing excessive "spikiness" of the matrix. In classical matrix completion (Cai et al., 2010), the incoherence assumption is used to ensure that the left and right singular vectors are not aligned with the standard basis vectors, and to facilitate establishment of recovery guarantees. This assumption was made less stringent in Negahban and Wainright (2012) by instead restricting the "spikiness" ratio of the matrix. The assumption $\|M\|_\infty \leq \alpha$ follows from this condition (Gunasekar et al., 2014). We refer the reader to Davenport et al. (2014), Cai and Zhou (2013), Klopp (2014) and Negahban and Wainright (2012) for further details.

2.2 Sampling Model

We now discuss our assumptions on the set Ω, on which we follow Bhojanapalli and Jain (2014), where the classical matrix completion problem is considered. Consider a bipartite graph $G = ([m],[n],E)$, where the edge set $E \subseteq [m]\times[n]$ is related to the index set of revealed entries Ω as $(i,j) \in E$ iff $(i,j) \in \Omega$. Abusing the notation, we use G for both the graph and its bi-adjacency matrix where $G_{ij} = 1$ if $(i,j) \in E$, $G_{ij} = 0$ if $(i,j) \notin E$.

We denote the association of G to Ω by $G\backslash\Omega$. Without loss of generality we take $m \geq n$. We assume that each row of G has d nonzero entries (thus $|\Omega| = md$) with the following properties on its singular value decomposition (SVD):

(A1) The left and right top singular vectors of G are $\mathbf{1}_m/\sqrt{m}$ and $\mathbf{1}_n/\sqrt{n}$, respectively.

(A2) We have $\sigma_1(G) \geq d$ and $\sigma_2(G) \leq C\sqrt{d}$, where $C > 0$ is some universal constant. Here $\sigma_1(G)$ and $\sigma_2(G)$ respectively denote the largest and the second largest singular values of G.

Thus we require G to have a large enough spectral gap.

Examples. We now discuss a few examples of graphs families which satisfy assumptions (A1) and (A2).

(1) Ramanujan graphs, a class of regular expander graphs (Hoory et al., 2006).

(2) Erdös-Renyi random graphs with average degree $d \geq c\,\log(m)$. These graphs satisfy this spectral gap property with high probability (Feige and Ofek, 2005). More explicitly, if G is an Erdös-Renyi bipartite random graph with probability p of an edge being placed, then the ensemble average of the bi-adjacency matrix $\mathbb{E}[G] = p\mathbf{1}_m\mathbf{1}_n^\top = \sqrt{nmp^2}\tilde{\mathbf{1}}_m\tilde{\mathbf{1}}_n^\top$ and $\tilde{G} = G - \mathbb{E}[G]$ is a random matrix with zero-mean i.i.d. entries of variance $p(1-p)$ with the largest singular value having $\mathcal{O}(\sqrt{m+n})$ with high probability (and also with probability 1) (Bolla et al., 2010). Thus, $\sigma_1(G) = \sqrt{nmp^2}$ is the dominant singular value of G, and (A1) and (A2) hold with high probability (and also with probability 1) (Bolla et al., 2010). Note that the uniform sampling assumption used in Davenport et al. (2014), Gunasekar et al. (2014), and Lan et al. (2014c), generates an Erdös-Renyi random graph.

(3) Stochastic block models for certain choices of inter- and intra-cluster edge connection probabilities. Consider the case of two clusters of the left and right vertices, with $m/2$ left vertices and $n/2$ right vertices of graph G belonging to the first cluster and the remaining left and right vertices to second cluster. Suppose that each intra-cluster edge is placed with probability p and an inter-cluster edge is placed with probability q. Then the ensemble average of the bi-adjacency matrix $\mathbb{E}[G]$ consists of elements equal to p for edges with both vertices in the same cluster and q for edges with vertices in different clusters. This can be expressed as (see also Nadakuditi and Newman, 2012)

$$\mathbb{E}[G] = \frac{\sqrt{mn}(p+q)}{2}\tilde{\mathbf{1}}_m\tilde{\mathbf{1}}_n^\top + \frac{\sqrt{mn}|p-q|}{2}\tilde{\mathbf{u}}_m\tilde{\mathbf{u}}_n^\top \tag{5}$$

where $\tilde{\mathbf{u}}_m = \mathbf{u}_m/\sqrt{m}$, the elements of $\mathbf{u}_m \in \mathbb{R}^m$ are ± 1, $+1$ if the left vertex is in the first cluster, -1 otherwise; similarly for $\mathbf{u}_n$ and $\tilde{\mathbf{u}}_n$. For even m and n (clusters of equal size), (5) is an SVD of $\mathbb{E}[G]$ since $\{\tilde{\mathbf{1}}_m, \tilde{\mathbf{u}}_m\}$ and $\{\tilde{\mathbf{1}}_n, \tilde{\mathbf{u}}_n\}$ are sets of orthonormal vectors representing the left and right singular vectors of non-zero singular values of $\mathbb{E}[G]$, which is of rank 2. The matrix $\tilde{G} = G - \mathbb{E}[G]$ is a random matrix with bounded and independent entries, with the largest singular value having $\mathcal{O}(\sqrt{m+n})$ (Bolla et al., 2010). Thus, the two largest singular values of G are $\frac{\sqrt{mn}(p+q)}{2}$ and $\frac{\sqrt{mn}|p-q|}{2}$ (perturbed by random $\tilde{G}$). When $p = q$, we have an Erdös-Renyi random graph with the largest spectral gap. As q becomes smaller, the spectral gap decreases. By (5), (A1) is true, but for (A2) to be true, one should have $|p-q| = \mathcal{O}(1/\sqrt{m})$. For fixed p and q,

$\sigma_2(G) = \frac{\sqrt{mn}|p-q|}{2}$ does not satisfy (A2) although the spectral gap can be made smaller by making $|p-q|$ smaller. As shown in Bhojanapalli and Jain (2014), the stochastic block model is a useful device to study the effect of the spectral gap on the performance of matrix recovery approaches. Such stochastic block models also apply to practical settings such as modeling connectivity subnetworks in the brain (Ghanbari et al., 2013).

3. Main Results

In this section, we describe the rank-constrained ML estimation of the target matrix M from multi-level quantized observations. We then establish upper bounds on the estimation error norm for this problem, which yield a faster rate of convergence than previously established upper bounds for the binary case.

3.1 Rank-Constrained Maximum Likelihood Estimation

We wish to estimate unknown M from the observed entries of Y using a constrained ML approach. We assume Y is related to M via the probabilistic model given in (1)-(4). We use $X \in \mathbb{R}^{m\times n}$ to denote the optimization variable. The negative log-likelihood function is given by

$$F_{\Omega,Y}(X) = -\sum_{(i,j)\in\Omega}\left[\sum_{\ell=1}^{K}\log(f_\ell(X_{ij}))\mathbf{1}_{[Y_{ij}=\ell]}\right] \tag{6}$$

which is a convex function in X when the function f_ℓ is log-concave in X_{ij}, and can be an implicit function of $\boldsymbol{\omega}$, where

$$\boldsymbol{\omega} = [\omega_1\ \omega_2\ \cdots\ \omega_{K-1}]^\top \in \mathbb{R}^{K-1} \tag{7}$$

is the vector of bin boundaries.

Two common choices for which the function f_ℓ, and its associated CDFs and pdfs, are log-concave, are:

(i) Logistic model with logistic CDF $\Phi(x) = \Phi_{log}(x/\sigma) = \frac{1}{1+e^{-x/\sigma}}$, $\sigma > 0$.

(ii) Probit model with $\Phi(x) = \Phi_{norm}(x/\sigma)$ where $\sigma > 0$ and $\Phi_{norm}(x)$ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$.

We consider two classes of problems. In the first, $\boldsymbol{\omega}$ is known, and thus the distribution in (4) is completely specified. We will assume that the bin boundaries ω_ℓ, $\ell \in [K]$, are known for our theoretical results. In the other, $\boldsymbol{\omega}$ is unknown and will be determined as part of the optimization problem. By allowing the bin boundaries to be determined by the optimization, we allow the distribution in (4) to be tuned to real data. For our numerical results, we allow the ω_ℓs to be unknown and estimate them (along with M), as in Lan et al. (2014a).

When $\boldsymbol{\omega}$ is known, we seek the solution to the optimization problem (P1): (s.t. stands for subject to)

$$\begin{aligned}(\text{P1}): \quad & \widehat{M} = \arg\min_X F_{\Omega,Y}(X)\\ & \text{s.t.}\quad \|X\|_\infty \le \alpha,\ \text{rank}(X) \le r.\end{aligned} \tag{8}$$

As a result of the rank constraint, (P1) is a nonconvex optimization problem. In Section 4, we will discuss factorization methods for solving this problem which come with guarantees of global convergence. In Section 3.2, we present performance upper bounds for problem (P1).

When $\boldsymbol{\omega}$ is unknown, the constrained ML estimate we wish to obtain is given by the solution to the optimization problem (P2):

$$\text{(P2)}: \quad \left(\widehat{M}, \widehat{\boldsymbol{\omega}}\right) = \arg\min_{X,\boldsymbol{\omega}} F_{\Omega,Y}(X) \text{ s.t. } \|X\|_\infty \leq \alpha,$$

$$\operatorname{rank}(X) \leq r \text{ and } \omega_1 < \omega_2 < \cdots < \omega_{K-1}. \tag{9}$$

The negative log-likelihood $F_{\Omega,Y}(X)$ is not jointly convex in X and $\boldsymbol{\omega}$. However, we show in Lemma 3 in Appendix A that f_ℓ is log-concave in ω_k for fixed X and ω_is $(i \neq k)$, and in Section 3.3, that it is strictly log-concave in X for fixed $\boldsymbol{\omega}$ whenever $\Phi(x)$ is log-concave. Thus, $F_{\Omega,Y}(X)$ is convex in ω_k for fixed X and ω_is $(i \neq k)$, and convex in X for fixed $\boldsymbol{\omega}$. Consequently, as seen later in Section 4, it will require an alternating minimization procedure (block-coordinate descent).

3.2 Performance Upper Bounds

We now present performance upper bounds for $\widehat{M}$ in (8), i.e., problem (P1) where $\boldsymbol{\omega}$, the vector of true bin boundaries, is assumed to be known. We first define some constants which appear in the bound, involving functions of $f_\ell(x)$ and its first two derivatives. With $\dot{f}(x) := (\mathrm{d}f(x)/\mathrm{d}x)$, define

$$\gamma_\alpha \leq \min_{\ell\in[K]} \inf_{|x|\leq\alpha} \left\{ \frac{\dot{f}_\ell^2(x)}{f_\ell^2(x)} - \frac{\ddot{f}_\ell(x)}{f_\ell(x)} \right\}, \tag{10}$$

$$L_\alpha \geq \max_{\ell\in[K]} \sup_{|x|\leq\alpha} \left\{ \left|\dot{f}_\ell(x)\right| / f_\ell(x) \right\}, \tag{11}$$

where α is the bound on the entry-wise infinity-norm of $\widehat{M}$ (see Equation 8). For further reference, define the constraint set

$$\mathcal{C} := \left\{ X \in \mathbb{R}^{m\times n} \,:\, \|X\|_\infty \leq \alpha,\ \operatorname{rank}(X) \leq r \right\}. \tag{12}$$

Theorem 1 *Suppose that $M \in \mathcal{C}$, and $G\backslash\Omega$ satisfies assumptions (A1) and (A2). Without loss of generality, assume $m \geq n$. Further, suppose Y is generated according to (1) where $f_\ell(x)$ is log-concave in x $\forall \ell \in [K]$. Then, provided $\gamma_\alpha > 0$, with probability at least $1 - 2(9\alpha\sqrt{mn})^{-r(m+n+1)} - C_1\exp(-C_2 m)$, any global minimizer $\widehat{M}$ of (8) satisfies*

$$\frac{1}{\sqrt{mn}}\|\widehat{M} - M\|_F \leq \min\left(2\alpha,\ U_1,\ U_2\right) \tag{13}$$

where

$$U_1 = \max\left(\frac{C_{1\alpha} r \sigma_2(G)}{\sigma_1(G)}, \frac{C_{2\alpha} m \sqrt{r^3 n}}{\sigma_1^2(G)} \right) \leq \max\left(\frac{C_{1\alpha} C r\sqrt{m}}{\sqrt{|\Omega|}}, \frac{C_{2\alpha} m^3 \sqrt{r^3 n}}{|\Omega|^2} \right), \tag{14}$$

$$U_2 = \max\left(\frac{C_{1\alpha} r \sigma_2(G)}{\sigma_1(G)},\ \frac{C_{3\alpha} r^{0.75}\left(|\Omega|(m+n+1)\log(9\alpha\sqrt{mn})\right)^{0.25}}{\sigma_1(G)}\right) \tag{15}$$

$$\leq \max\left(\frac{C_{1\alpha} C r\sqrt{m}}{\sqrt{|\Omega|}},\ C_{3\alpha}\left(\frac{r}{|\Omega|}\right)^{0.75} m\left((m+n+1)\log(9\alpha\sqrt{mn})\right)^{0.25}\right). \tag{16}$$

Here, $C_{1\alpha} = 4\sqrt{2}\alpha$*,* $C_{2\alpha} = 32.16\sqrt{2}L_\alpha/\gamma_\alpha$*,* $C_{3\alpha} = 8\sqrt{\frac{(1+\alpha)L_\alpha}{\gamma_\alpha}}$*,* $C_1, C_2, C > 0$ *are universal constants, and* γ_α *and* L_α *are given by (10), (11).*

A proof of Theorem 1 is given in Appendix B. In the binary case ($K = 2$) the link function f in (4) belongs to the exponential family for a large class of CDFs $\Phi(x)$ (e.g., logistic or Gaussian), but not for $K > 3$. The bounding approaches in Gunasekar et al. (2014), Lafond (2015) and Cao and Xie (2015) for $K > 3$ require f to come from the exponential family whereas our approach based on a Taylor series approximation and some concentration inequalities applies to the quantization model (4). One of the novelties in our proof compared to existing works is how we bound the gradient of the cost function in two different ways (see Lemmas 5 and 6 in Appendix B).

Of some interest is the case where the fraction of revealed entries $p = \frac{|\Omega|}{mn}$ is fixed and we let m and n become large, with $m/n \equiv \delta \geq 1$ fixed. In this case we have the following Corollary.

Corollary 2 *Assume the conditions of Theorem 1. Let* $p = \frac{|\Omega|}{mn}$ *be fixed independent of* m *and* n*. Then*

$$U_1 \leq \mathcal{O}\left(\frac{\delta}{p^2}\sqrt{\frac{r^3}{n}}\right), \quad U_2 \leq \mathcal{O}\left(\left(\frac{r^3\delta^2\log(n)}{p^3 n}\right)^{1/4}\right).$$

Then with probability at least $1 - C_1\exp(-C_2 m) - 2/(9\alpha n\sqrt{\delta})^{2rn}$*, any global minimum* $\widehat{M}$ *to (8) satisfies*

$$\frac{1}{\sqrt{mn}}\|\widehat{M} - M\|_F \leq \min\left(\mathcal{O}\left(\frac{\delta}{p^2}\sqrt{\frac{r^3}{n}}\right), \mathcal{O}\left(\left(\frac{r^3\delta^2\log(n)}{p^3 n}\right)^{1/4}\right)\right). \tag{17}$$

Corollary 2 suggests that for "larger" fixed values of p, U_1 dominates, signifying a convergence rate of at least $1/\sqrt{n}$ for $\frac{1}{\sqrt{mn}}\|\widehat{M} - M\|_F$ whereas for "small" values of p, U_2 is likely to dominate signifying convergence rate of at least $(\log(n)/n)^{1/4}$. In our simulation results shown later in Figure 5 for $m = n = 100, 200$, or 400, and $p = 0.2, 0.4$, or 0.6, we find that $\frac{1}{mn}\|\widehat{M} - M\|_F^2$ decays approximately as $\mathcal{O}(1/n)$.

3.3 Constants γ_α and L_α for the logistic and probit models

It is known that $f_\ell(x)$ is log-concave iff $\ddot{f}_\ell(x) f_\ell(x) \leq (\dot{f}_\ell(x))^2$ (Boyd and Vandenberghe, 2004). Thus $\gamma_\alpha \geq 0$ for log-concave $f_\ell(x)$ and $\gamma_\alpha > 0$ for strictly log-concave $f_\ell(x)$. For the logistic model, i.e., when $\Phi(x) = \Phi_{log}(x/\sigma)$, some tedious calculations show that

$$\begin{aligned}\frac{\dot{f}_\ell^2(x)}{f_\ell^2(x)} - \frac{\ddot{f}_\ell(x)}{f_\ell(x)} = \frac{1}{\sigma^2}\Big[&\Phi_{log}\left(\frac{\omega_\ell - x}{\sigma}\right)\left(1 - \Phi_{log}\left(\frac{\omega_\ell - x}{\sigma}\right)\right) \\ &+ \Phi_{log}\left(\frac{\omega_{\ell-1} - x}{\sigma}\right)\left(1 - \Phi_{log}\left(\frac{\omega_{\ell-1} - x}{\sigma}\right)\right)\Big] \\ > 0 \;\; \forall x \in \mathbb{R},\ \forall \ell \in [K]. \qquad (18)\end{aligned}$$

Therefore, by (10), $\gamma_\alpha > 0$ for the logistic model. Similarly one can verify numerically that $\gamma_\alpha > 0$ for the probit model when $\Phi(x) = \Phi_{norm}(x/\sigma)$. For the logistic model, it turns out that $L_\alpha = 1/(2\sigma\beta_{\alpha\sigma})$ where

$$\beta_{\alpha\sigma} := \min_{\ell \in [K]} \inf_{|x| \leq \alpha} \left\{ \Phi_{log}(\frac{\omega_\ell - x}{\sigma}) - \Phi_{log}(\frac{\omega_{\ell-1} - x}{\sigma}) \right\} > 0. \qquad (19)$$

For the probit model, we have $L_\alpha = \sqrt{2}/(\sqrt{\pi}\sigma\beta_{n\alpha\sigma})$ where

$$\beta_{n\alpha\sigma} := \min_{\ell \in [K]} \inf_{|x| \leq \alpha} \left\{ \Phi_{norm}(\frac{\omega_\ell - x}{\sigma}) - \Phi_{norm}(\frac{\omega_{\ell-1} - x}{\sigma}) \right\} > 0. \qquad (20)$$

3.4 Comparison of Convergence Rates

We first provide a comparison of our bounds with those of Davenport et al. (2014) and Cai and Zhou (2013), who have established bounds for only the binary ($K = 2$) level case. Consider $M \in \mathbb{R}^{n \times n}$, with p fraction of its entries sampled. Then $m = n$, and $|\Omega| = pn^2$. The bounds proposed in Davenport et al. (2014) and Cai and Zhou (2013) yield (for $|\Omega| \geq 4n\log(n)$)

$$\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq \mathcal{O}\left(\sqrt{\frac{r}{pn}}\right). \qquad (21)$$

Our bound (Corollary 2) yields

$$\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq \min\left(\mathcal{O}\left(\frac{r^3}{p^4 n}\right), \mathcal{O}\left(\sqrt{\frac{r^3\log(n)}{p^3 n}}\right)\right). \qquad (22)$$

For $K = 2$, the results of Lafond et al. (2014), Gunasekar et al. (2014) and Lafond (2015) apply to our model but not for $K > 2$. The bound of Lafond et al. (2014) and Lafond (2015) yields

$$\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq \mathcal{O}\left(\frac{r\log(n)}{pn}\right) \qquad (23)$$

and the bound of (Gunasekar et al., 2014, Corollary 1) yields

$$\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq \mathcal{O}\left(\alpha^{*2}\frac{r\log(n)}{pn}\right) = \mathcal{O}\left(\frac{rm\log(n)}{p}\right) \qquad (24)$$

since in Gunasekar et al. (2014), $\alpha^* \geq \sqrt{mn}\|M\|_\infty$. The bound in Soni et al. (2014) as applied to non-sparse models and $K = 2$, yields

$$\frac{1}{n^2}\mathbb{E}\left[\|\widehat{M} - M\|_F^2\right] \leq \mathcal{O}\left(\frac{r\log(n)}{pn}\right) \qquad (25)$$

where the expectation is over the noise (Z_{ij} in Equation 2) and sampling distributions of the revealed matrix entries. Thus, (25) is similar to (23) but is averaged over the noise and sampling distributions. The multi-level observation model of Cao and Xie (2015) does not include the quantized observation model given in (2) but applies to a multinomial logistic model. The bound in Cao and Xie (2015) yields

$$\frac{1}{n^2}\|\widehat{M}-M\|_F^2 \le \mathcal{O}\left(\sqrt{\frac{r}{pn}}\right). \tag{26}$$

In (21)-(26) the omitted constants do not depend on r, p or n.

Comparing (21)-(26) for the case $K = 2$, we see our method has faster convergence rate in n for fixed rank r and fraction of revealed entries p, compared to Davenport et al. (2014), Cai and Zhou (2013), Lafond et al. (2014), Gunasekar et al. (2014), Lafond (2015) and Soni et al. (2014); the same comment applies for $K > 2$ when comparing with Cao and Xie (2015). On the other hand, for fixed n, our bound is inferior to these other bounds in p or r. One may notice if the revealed entries scale with n according to $p \sim \mathcal{O}(\log(n)/n)$ then Davenport et al. (2014) and Cao and Xie (2015) yield bounded error while our bound grows with n; in our case we need $p \ge \mathcal{O}(1/n^{1/3})$. We believe this to be an artifact of our proof, as our numerical results in Figure 1 show our method outperforms Cai and Zhou (2013) and Davenport et al. (2014), especially for low values of p and higher values of rank r.

4. Optimization

In this section, we describe the algorithms used to solve problems (P1) and (P2). We use the matrix factorization technique (Bach et al., 2008; Burer and Monteiro, 2003; Lee et al., 2010) where instead of optimizing with respect to X, it is factorized into two matrices $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ such that $X = UV^\top$. One then optimizes with respect to the factors U, V. This method is non-convex, however, it is known (Bach et al., 2008; Burer and Monteiro, 2003; Lee et al., 2010) that if k is chosen to be large enough, it is guaranteed that the local minimum of the problem is also the global minimum of the non-factorized problem.

4.1 Known Bin Boundaries

We have the following approximate projected gradient method for solving problem (P1) following the algorithm of Cai and Zhou (2013), where the case of $K = 2$ was considered.

4.1.1 Algorithm 1: Approximate Projected Gradient Method

Given initial estimates U^0, V^0, one updates

$$\begin{bmatrix} U^{t+1} \\ V^{t+1} \end{bmatrix} = \mathcal{P}_\alpha\left(\begin{bmatrix} U^t - \frac{\tau}{\sqrt{t}}\nabla_X F_{\Omega,Y}(U^t {V^t}^\top)V^t \\ V^t - \frac{\tau}{\sqrt{t}}\nabla_X F_{\Omega,Y}(U^t {V^t}^\top)^\top U^t \end{bmatrix}\right) \tag{27}$$

where U^t, V^t are the estimates at iteration t, and

$$\mathcal{P}_\alpha\left([U^\top \; V^\top]^\top\right) = \begin{cases} \sqrt{\alpha/\|UV^\top\|_\infty}[U^\top \; V^\top]^\top & \text{if} \quad \|UV^\top\|_\infty > \alpha \\ \left[U^\top \; V^\top\right]^\top & \text{if} \quad \|UV^\top\|_\infty \le \alpha. \end{cases} \tag{28}$$

In (27) the stepsize τ is selected via a backtracking line search using Armijo's rule, to minimize the cost $F_{\Omega,Y}(U^{t+1}V^{t+1\top})$.

In addition to the approximate projection $\mathcal{P}_\alpha$ in (28), Cai and Zhou (2013) (for $K=2$) also uses another projection to enforce a max-norm constraint. In Cai and Zhou (2013), for $K=2$, the negative log-likelihood cost is minimized w.r.t. X subject to the constraints $\|X\|_\infty \le \alpha$ and $\|X\|_{\max} \le R$ for some $\alpha > 0$ and $R > 0$. The operator $\mathcal{P}_\alpha$ enforces the constraint $\|X\|_\infty \le \alpha$. The factored form definition of the max norm (Lee et al., 2010) is given by $\|X\|_{\max} = \inf\left\{\max(\|\bar{U}\|_{2,\infty}^2, \|\bar{V}\|_{2,\infty}^2) : X = \bar{U}\bar{V}^\top\right\}$ where $\|\bar{U}\|_{2,\infty} = \max_i \sqrt{\sum_j \bar{U}_{ij}^2}$, $\bar{U} \in \mathbb{R}^{m\times k}$, $\bar{V} \in \mathbb{R}^{n\times k}$, $k = 1,2,\ldots,\min(m,n) = n$. For fixed k and $X = UV^\top$, Cai and Zhou (2013) enforce the constraint set $\|X\|_{\max} \le R$ by requiring $\max(\|U\|_{2,\infty}^2, \|V\|_{2,\infty}^2) \le R$. As stated in Cai and Zhou (2013), the global minimum of the cost over the constraints $\|X\|_\infty \le \alpha$ and $\|X\|_{\max} \le R$ is the same as that over the constraints $\|X\|_\infty \le \alpha$, $\|U\|_{2,\infty}^2 \le R$ and $\|V\|_{2,\infty}^2 \le R$ if $\operatorname{rank}(X) \le k$. If a matrix A has rank r and $\|A\|_\infty \le \alpha$, then $\|A\|_{\max} \le \sqrt{r}\alpha$ (Cai and Zhou, 2013). Therefore, in our case the max-norm constraint is unnecessary as it is automatically satisfied for any $R \ge \sqrt{r}\alpha$. In this sense, our Algorithm 1 is the same as the approach of Cai and Zhou (2013) when $K=2$ and one picks $R \ge \sqrt{r}\alpha$.

Remark 1 *The hard rank constraint results in a nonconvex constraint set. Thus,* (8) *is a nonconvex optimization problem; similarly for Algorithm 1 for which the rank constraint is implicit in the factorization of X. However, the following result is shown in (Bach et al., 2008, Proposition 4), based on Burer and Monteiro (2003), for nonconvex problems of this form. If (U^*, V^*) is a local minimum of the reformulated (i.e., factored) problem, then $X^* = U^*V^{*\top}$ is the global minimum of problem* (8)*, so long as U^* and V^* are rank-deficient. (Rank deficiency of (U^*, V^*) is a sufficient condition, not necessary.) This result is invoked in Recht and Re (2013), Lee et al. (2010) and Cai and Zhou (2013) for problems of this form. Thus one would expect to achieve global convergence for the problem of* (8) *provided iterations (27)-(28) converge to a local minimum. These iterations represent a projected gradient method which converges to a stationary point if one has orthogonal projection onto a convex constraint set (Bertsekas, 1999, Prop. 2.3.1). However, the "projection" $\mathcal{P}_\alpha$ in (28) is not an orthogonal projection and the set $\{\|UV^\top\|_\infty \le \alpha\}$ is not convex in U, V (although $\{\|X\|_\infty \le \alpha\}$ is convex in X), therefore, convergence to even a local minimum is not ensured. However, numerically, this method has still provided good results (similarly reported in Cai and Zhou (2013)). In Cai and Zhou (2013), for the $K=2$ case, there are two additional constraints $\|U\|_{2,\infty}^2 \le R$ and $\|V\|_{2,\infty}^2 \le R$ which are convex sets in U and V, and the corresponding projections are orthogonal projections. However, the projection corresponding to our $\mathcal{P}_\alpha$ in Cai and Zhou (2013) is not orthogonal.*

Thus, for problems of this form, one can choose $k = r+1$ to achieve global convergence if an upper bound r on the rank of M is known.

The convergence deficiency discussed in Remark 1 motivates the following log-barrier penalty function approach.

4.1.2 Algorithm 2: Logarithmic Barrier Gradient Method

The constraint $\max_{i,j} |X_{ij}| \leq \alpha$ translates to $X_{ij} - \alpha \leq 0$ and $-X_{ij} - \alpha \leq 0 \; \forall (i,j)$, which motivates the log-barrier penalty function $-\log\left(1 - (X_{ij}/\alpha)^2\right)$, which is finite for $|X_{ij}| < \alpha$, and is infinite otherwise. This leads to the objective function

$$\tilde{F}_{\Omega,Y}(X) = F_{\Omega,Y}(X) - \lambda \sum_{(i,j)} \log\left(1 - (X_{ij}/\alpha)^2\right) \tag{29}$$

and the optimization problem

$$\widehat{M} = \arg\min_X \tilde{F}_{\Omega,Y}(X) \, \text{s.t.} \, \text{rank}(X) \leq r. \tag{30}$$

The parameter $\lambda > 0$ in (29) sets the accuracy of approximation of $\max_{i,j} |X_{ij}| \leq \alpha$ via the log-barrier function (which is twice-differentiable and convex in X, hence so is $\tilde{F}_{\Omega,Y}(X)$). Now, however, the factorization approach $X = UV^\top$ is well-justified, per Remark 1, and convergence is guaranteed.

The log-barrier method is ill-conditioned and solving the problem for a fixed value of λ generally only works well for small problems or good choices of initialization (Boyd and Vandenberghe, 2004). The above problem is typically solved via a sequence of central path following solutions (Boyd and Vandenberghe, 2004) where one gradually reduces λ toward 0. In our approach we initialize it with the solution to Algorithm 1, providing a good initialization, and then either use a single "small" value of λ, or select λ via 5-fold cross-validation. One may therefore view augmentation with the log-barrier cost as regularization of $F_{\Omega,Y}(X)$.

We solve the factored version $X = UV^\top$ of problem (30) for a fixed λ using a gradient method as follows. Given initial estimates U^0, V^0, one updates

$$\begin{bmatrix} U^{t+1} \\ V^{t+1} \end{bmatrix} = \begin{bmatrix} U^t - \frac{\tau}{\sqrt{t}} \nabla_X \tilde{F}_{\Omega,Y}(U^t {V^t}^\top) V^t \\ V^t - \frac{\tau}{\sqrt{t}} \nabla_X \tilde{F}_{\Omega,Y}(U^t {V^t}^\top)^\top U^t \end{bmatrix} \tag{31}$$

where U^t, V^t are the estimates at iteration t and $\nabla_X \tilde{F}_{\Omega,Y}(U^t {V^t}^\top) = \nabla_X \tilde{F}_{\Omega,Y}(X)\big|_{X=U^t {V^t}^\top}$. In (31) the stepsize τ is selected via a backtracking line search using Armijo's rule, to minimize the cost $\tilde{F}_{\Omega,Y}(U^{t+1} {V^{t+1}}^\top)$.

Define $Z = [U^\top \; V^\top]^\top \in \mathbb{R}^{(m+n)\times k}$ and let $\bar{F}_{\Omega,Y}(Z^t) = \tilde{F}_{\Omega,Y}(U^t {V^t}^\top)$. Then (31) can be rewritten as

$$Z^{t+1} = Z^t - \frac{\tau}{\sqrt{t}} \nabla_Z \bar{F}_{\Omega,Y}(Z^t), \quad \nabla_Z \bar{F}_{\Omega,Y}(Z^t) = \begin{bmatrix} \nabla_X \tilde{F}_{\Omega,Y}(U^t {V^t}^\top) V^t \\ \nabla_X \tilde{F}_{\Omega,Y}(U^t {V^t}^\top)^\top U^t \end{bmatrix}. \tag{32}$$

Thus, (31) is a gradient descent method using Armijo's rule for stepsize selection, for unconstrained minimization of the continuously differentiable cost $\tilde{F}_{\Omega,Y}(UV^\top) = \bar{F}_{\Omega,Y}(Z)$ w.r.t. Z. By (Bertsekas, 1999, Prop. 1.2.1), the iterations given in (31) converge to a stationary point of $\tilde{F}_{\Omega,Y}(UV^\top)$. Since the cost is non-increasing at every iteration, the stationary point is either a local minimum or a saddle point. In theory, convergence to saddle points can not be excluded for gradient descent algorithms for continuously differentiable functions. However, we assume that we escape saddle points in practice as saddle points are generally unstable from a numerical point of view, i.e., a perturbation always exists at the saddle point which will decrease the cost function.

4.2 Unknown Bin Boundaries

As noted earlier $F_{\Omega,Y}(X)$ is convex in ω_k for fixed X and ω_is $(i \neq k)$, and convex in X for fixed $\boldsymbol{\omega}$, however, $F_{\Omega,Y}(X)$ is not jointly convex in X and $\boldsymbol{\omega}$. Thus the problem (P2) specified in (9) is multi-convex in X and $\omega_1, \omega_2, \cdots, \omega_{K-1}$, and one approach to solve it is via block-coordinate descent (Xu and Yin, 2013), for which there are no convergence guarantees, in general. In order to detail this approach, with an abuse of notation, we now explicitly denote the dependence of $F_{\Omega,Y}(X)$ on the ω_is as $F_{\Omega,Y}(X, \omega_1, \cdots, \omega_{K-1})$, and that of $\tilde{F}_{\Omega,Y}(X)$ as $\tilde{F}_{\Omega,Y}(X, \omega_1, \cdots, \omega_{K-1})$. Following Xu and Yin (2013) and our Algorithms 1 and 2, our optimization algorithm for the case of unknown bin boundaries is given in Algorithm 3 where, in Step 5, δ_0 makes the constraint set for ω_i convex.

Algorithm 3 Block-Coordinate Descent Method for Solving (9)

Input: Set of observed entries Y_{ij} for $(i,j) \in \Omega$, initialization $U^0 \in \mathbb{R}^{m\times k}$, $V^0 \in \mathbb{R}^{m\times k}$, $\omega_1^0, \omega_2^0, \cdots, \omega_{K-1}^0 \in \mathbb{R}$, $\omega_1^0 < \omega_2^0 < \cdots < \omega_{K-1}^0$, parameters α, λ
Output: Solution $X^* = U^* {V^*}^\top$, $\boldsymbol{\omega}^*$

```
1: for $\ell = 1, 2, \cdots$, until convergence, do
2:     $(L^\ell, R^\ell) \leftarrow \arg\min_{U,V} F_{\Omega,Y}(UV^\top, \omega_1^{\ell-1}, \cdots, \omega_{K-1}^{\ell-1})$ subject to $\|UV^\top\|_\infty \leq \alpha$. Solve using approximate projected gradient method (27) initialized with $U^{\ell-1}, V^{\ell-1}$.
3:     $(U^\ell, V^\ell) \leftarrow \arg\min_{U,V} \tilde{F}_{\Omega,Y}(UV^\top, \omega_1^{\ell-1}, \cdots, \omega_{K-1}^{\ell-1})$. Solve using log-barrier gradient method (31) initialized with $L^\ell, R^\ell$.
4:     for $i = 1, 2, \cdots, K-1$, do
5:         $\omega_i^\ell \leftarrow \arg\min_{\omega_i} F_{\Omega,Y}(U^\ell {V^\ell}^\top, \omega_1^\ell, \cdots, \omega_{i-1}^\ell, \omega_i, \omega_{i+1}^{\ell-1}, \cdots, \omega_{K-1}^{\ell-1})$ subject to $\omega_{i-1}^\ell + \delta_0 \leq \omega_i \leq \omega_{i+1}^{\ell-1} - \delta_0$ for some "small" $\delta_0 > 0$ . Solve using a gradient descent method initialized with $\omega_i = \omega_i^{\ell-1}$.
6:     end for
7: end for
8: return $X^* = U^* {V^*}^\top$, $\boldsymbol{\omega}^*$
```

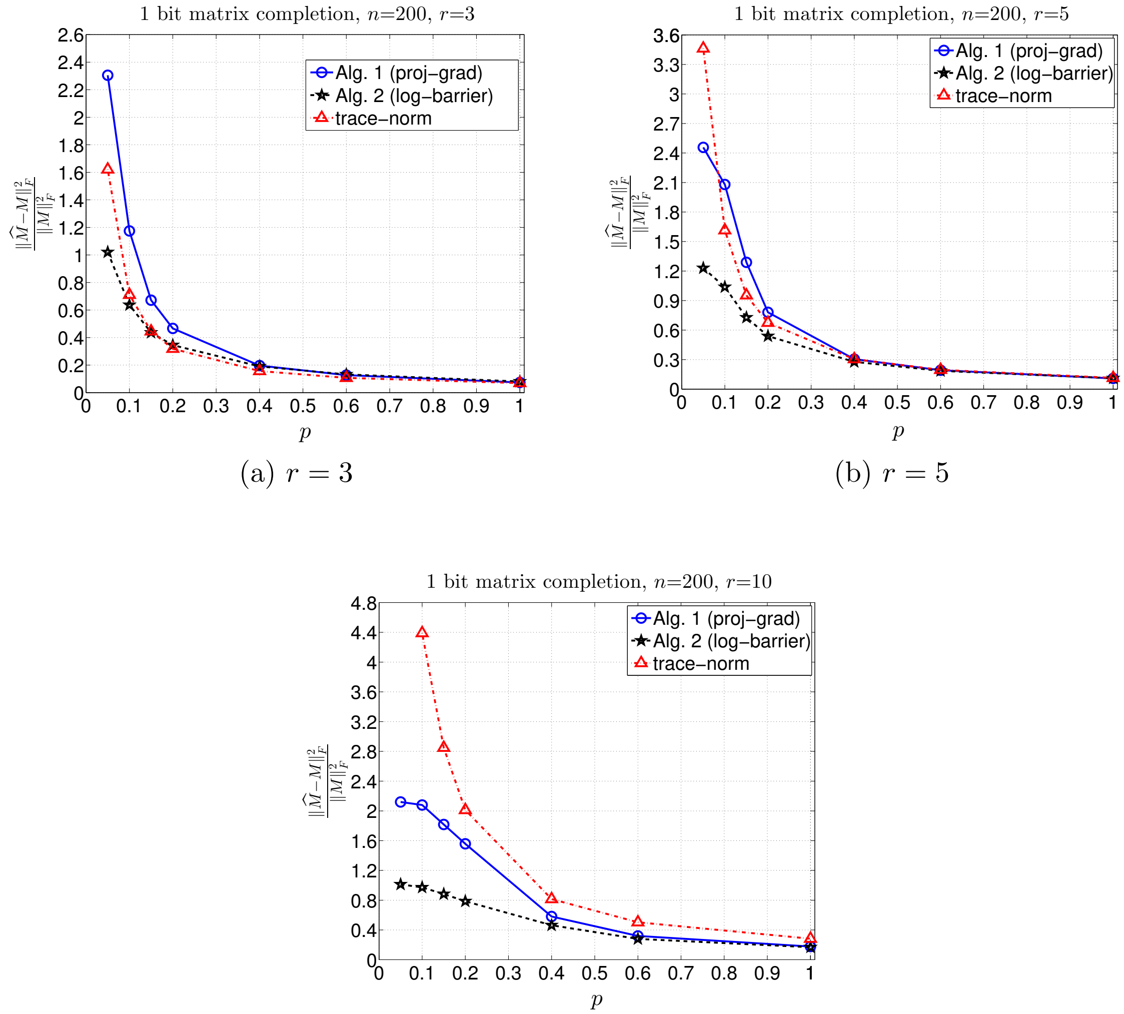
In Step 2 of Algorithm 3, we have used the solution of the approximate projected gradient method (27), to provide a good initialization to the log-barrier algorithm in Step 3, since we are not using central path following, for the reasons aforementioned in Section 4.1.2.

5. Numerical Experiments

In this section, we test our methods on synthetic and real data, and also compare our methods with that of Keshavan et al. (2009, 2010) (OptSpace), Cai et al. (2010) (SVT), Cai and Zhou (2013) and Davenport et al. (2014).

5.1 Synthetic Data

In this section, we report the results of evaluating our method on synthetic data. We set $m = n$ and construct $M \in \mathbb{R}^{n\times n}$ as $M = M_1 M_2^\top$ where M_1 and M_2 are $n \times r$ matrices with



(a) $r = 3$

(b) $r = 5$

(c) $r = 10$

Figure 1: Relative MSE $\|\widehat{M} - M\|_F^2/\|M\|_F^2$ for varied values of $p = q$, $n = 200$, $\alpha = 1$, Gaussian noise with $\sigma = 0.18$, K=2: binary case, $w_1 = 0$, known bin boundaries, "trace-norm" refers to Davenport et al. (2014), the proposed Alg. 1 (proj-grad) coincides with the algorithm of Cai and Zhou (2013) when $K = 2$ and one picks $R \geq \sqrt{r}\alpha$ in Cai and Zhou (2013).

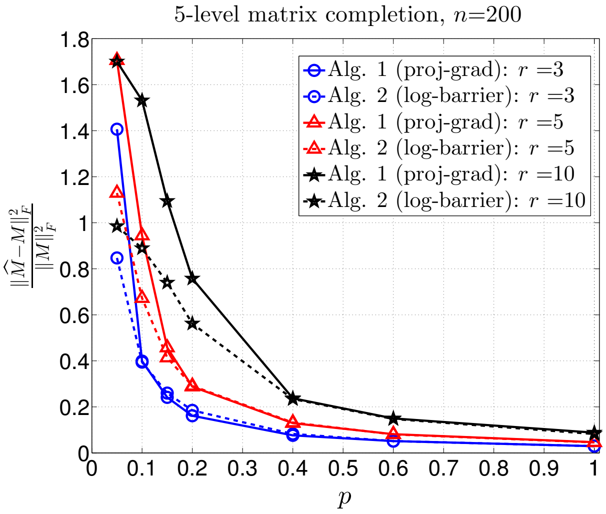


Figure 2: Relative MSE $\|\widehat{M} - M\|_F^2 / \|M\|_F^2$ for varied values of $p = q$, $n = 200$, $\alpha = 1$, Gaussian noise with $\sigma = 0.18$, K=5: $w_1 = -0.4$, $w_2 = -0.15$, $w_3 = 0.15$, $w_4 = 0.4$, known bin boundaries.

i.i.d. entries drawn from a uniform distribution on $[-0.5, 0.5]$. Then we scaled M to achieve $\|M\|_\infty = 1 = \alpha$. We pick $r = 3$, 5 or 10, and vary matrix sizes $n = 100, 200$, or 400. We used the model (2) with Z_{ij} as a zero-mean Gaussian with standard deviation $\sigma = 0.18$. These choices follow the numerical experiments of Cai and Zhou (2013) and Davenport et al. (2014), which dealt with the case of binary observations (i.e., in which $K = 2$). We generate the set Ω of revealed indices via a stochastic block model as in Bhojanapalli and Jain (2014). In the basic stochastic block model, we divide the set of nodes $[n]$ into two clusters, where each intra-cluster edge is sampled uniformly with probability p and an inter-cluster edge is sampled with probability q. For our simulations, initially we chose $p = q$ which corresponds to the Bernoulli sampling model of Davenport et al. (2014). Then we change the fraction of revealed 1-bit entries as $p = 0.05, 0.1, 0.15, 0.2, 0.4, 0.6$ or 1. Algorithm 1 was implemented with random initialization and its result was used to initialize Algorithm 2 where we either picked λ via 5-fold cross-validation (how well the label values of revealed Y_{ij} in the test set are matched), or simply used a small fixed λ. We assumed the bin boundaries to be known. The resulting relative mean-square error (MSE) $\|\widehat{M} - M\|_F^2 / \|M\|_F^2$, averaged over 20 Monte Carlo runs, is shown for $n = 200$ in Figure 1 for $K = 2$, and Figure 2 for $K = 5$. As expected, the performance improves with increasing n and increasing p. For comparison, Figure 1 also shows the MSE for Davenport et al. (2014) and Cai and Zhou (2013), and it is seen that Algorithm 2 (log-barrier) significantly outperforms Davenport et al. (2014) and Cai and Zhou (2013) for low values of p and high values of r (e.g., $r = 10$ and $p < 0.4$), and the performances are comparable for higher values of p and lower values of r (e.g., $r = 3$ and $p > 0.1$). Note that as aforementioned, Algorithm 1 (proj-grad) coincides with the algorithm of Cai and Zhou (2013) when $K = 2$ and one picks $R \geq \sqrt{r}\alpha$ in Cai and Zhou (2013).

In Figure 3 we show the results for the case of $r = 5$ and $m = n = 200$ for both known and estimated bin boundaries. As before, the results were averaged over 20 Monte Carlo runs, and the missing entries were set via the stochastic block model with $p = q$. For the case of unknown bin boundaries, we used Algorithm 3 with initialization $\omega_1^0 = -0.3$, $\omega_2^0 = -0.1$, $\omega_3^0 = 0.1$ and $\omega_4^0 = 0.3$, and $\alpha = 1$, $\lambda = 0.5$. Also shown are the results of the algorithm OptSpace of Keshavan et al. (2010) which is a matrix completion algorithm that assumes a real-valued low-rank matrix. The results using OptSpace were obtained for two cases: the case labeled "quantized noisy M" refers to the case where OptSpace is provided with revealed Y_{ij}'s, and the case labeled "unquantized noisy M" refers to the case where OptSpace works on revealed noisy M_{ij}s (i.e., $M_{ij} + Z_{ij}$s). For OptSpace, we scaled the estimate $\widehat{M}$ to have the same Frobenius norm as the true M before computing the MSE. It is seen that for $p \geq 0.2$, there is no loss in performance when bin boundaries are estimated, using the proposed approach. The algorithm OptSpace performs poorly for quantized data.

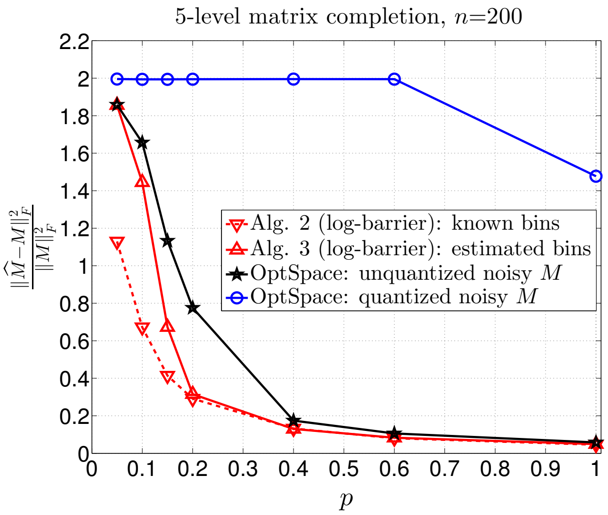


Figure 3: Relative MSE $\|\widehat{M} - M\|_F^2 / \|M\|_F^2$ for varied values of $p = q$, $n = 200$, $\alpha = 1$, Gaussian noise with $\sigma = 0.18$, K=5: true bin boundaries $w_1 = -0.4$, $w_2 = -0.15$, $w_3 = 0.15$, $w_4 = 0.4$, known and estimated bin boundaries. OptSpace is the method of Keshavan et al. (2010).

In Figure 4, we show the results for varying number of quantization levels K, with $r = 3, 5, 10$, $p = 0.2$, $m = n = 200$ and known bin boundaries. The matrix M is constructed as for Figure 1. The results were over 20 Monte Carlo runs, and the missing entries were set via the stochastic block model with $p = q$. With $\alpha = 1$, the bin boundaries were picked as $w_1 = 0$ for $K = 2$, $w_1 = -0.2$, $w_2 = 0.2$ for $K = 3$, $w_1 = -0.25$, $w_2 = 0$, $w_3 = 0.25$ for $K = 4$, $w_1 = -0.4$, $w_2 = -0.15$, $w_3 = 0.15$, $w_4 = 0.4$ for $K = 5$ and $w_1 = -0.4$, $w_2 = -0.2$, $w_3 = -0.05$, $w_4 = 0.05$, $w_5 = 0.2$, $w_6 = 0.4$ for $K = 7$. These choices yield a comparable number of entries in each bin. It is seen from Figure 4 that performance improves with increasing K. This is not surprising since with increasing K, bin intervals shrink and the quantization error becomes smaller. For the considered model there are two sources of error/noise: additive noise and quantization error.

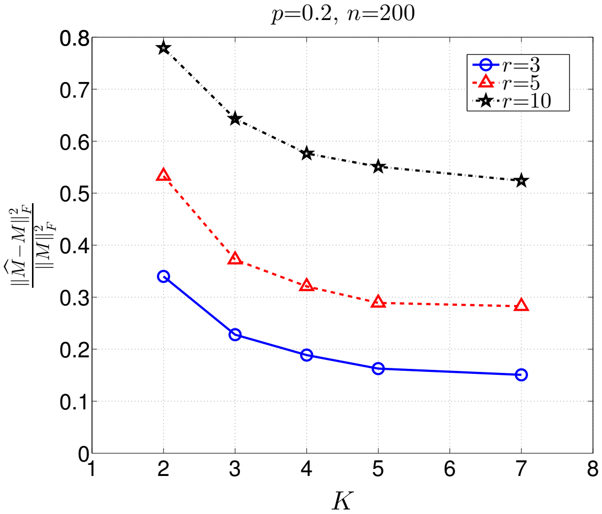
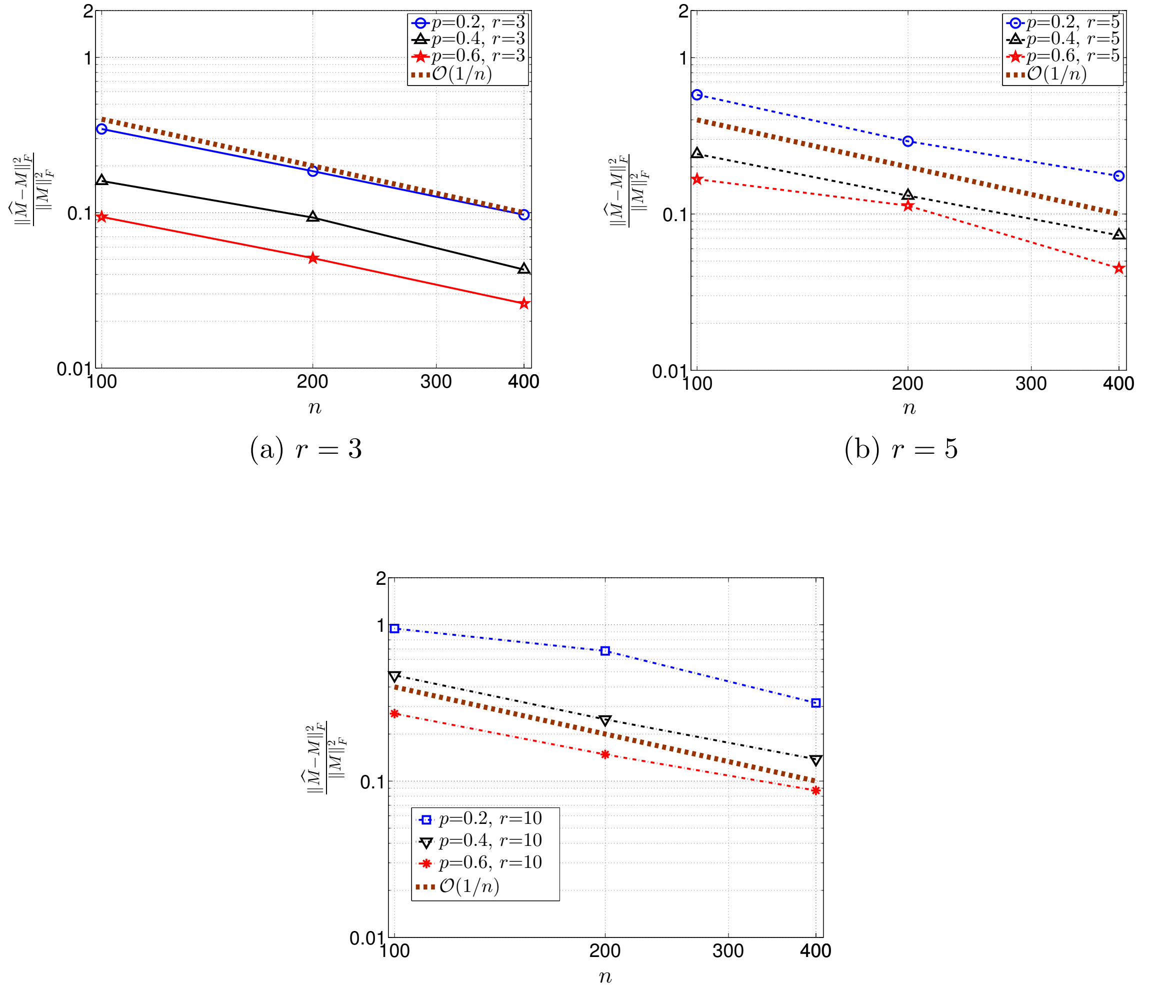


Figure 4: Relative MSE $\|\widehat{M}-M\|_F^2/\|M\|_F^2$ for varied values of K, $p=0.2$, $n=200$, $\alpha=1$, Gaussian noise with $\sigma=0.18$.

In Figure 5 we show the relative MSE for $r=3$, 5 and 10, respectively, and $n=100,200,400$, $p=q=0.2,0.4,0.6$. In Section 3.2 (see Equation 22), the upper bound on MSE was established as $\min\left(\mathcal{O}\left(\frac{r^3}{p^4 n}\right), \mathcal{O}\left(\sqrt{\frac{r^3\log(n)}{p^3 n}}\right)\right)$. Therefore, for fixed r and p, the bound is $\mathcal{O}\left(\frac{1}{n}\right)$, whereas for fixed n and p, the bound is $\mathcal{O}\left(r^{1.5}\right)$, and for fixed n and r, the bound is $\mathcal{O}\left(p^{-1.5}\right)$. We also plot the lines $1/n$ in Figure 5 to show the scale of the upper bound $\mathcal{O}(1/n)$ for fixed r and p. As we can see, the empirical estimation errors follow approximately the same scaling, suggesting that our analysis is tight with respect to n, up to some constant. In Figures 6 and 7 we show the relative MSE as a function of r and p, respectively, for $p=q=0.2$ and $n=100,200,400$. We also plot the lines $r^{1.5}$ and $1/p^{1.5}$ in Figures 6 and 7, respectively, to show the scale of the upper bound $\mathcal{O}\left(r^{1.5}\right)$ for fixed n and p, and the upper bound $\mathcal{O}\left(1/p^{1.5}\right)$ for fixed n and r. Now we see that these bounds are not tight. The empirical MSE results are approximately $\mathcal{O}(r)$ for fixed n and p and $\mathcal{O}(1/p)$ for fixed n and r.

In Figure 8 we additionally plot the relative MSE for $n=200$ and rank $r=5$, via the same method described above, but with varying p and keeping $p+q=0.7$, under the probit model. This enables us to study the performance of the model under nonuniform sampling. Note that when $p=q=0.35$, then the spectral gap is largest (Bhojanapalli and Jain, 2014) and the MSE is the smallest, and as p gets larger, the spectral gap decreases, leading to larger MSE.



(a) $r = 3$

(b) $r = 5$

(c) $r = 10$

Figure 5: Relative MSE: $K = 5$, $p = q$, $r = 3$, 5 or 10, $n = 100$, 200, 400, known bin boundaries, $\alpha = 1$, Gaussian noise $\sigma = 0.18$.

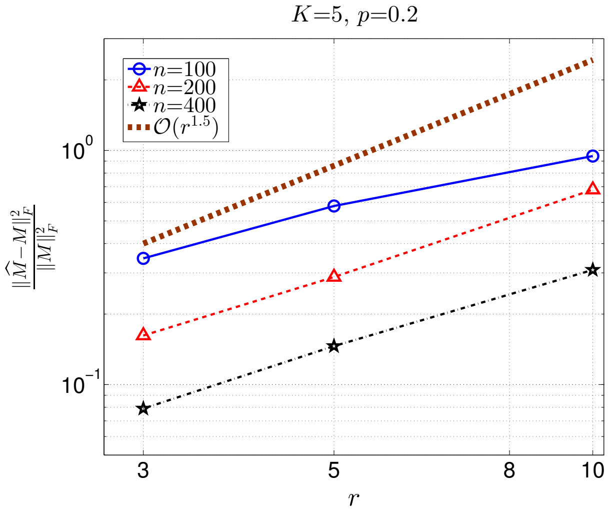


Figure 6: Relative MSE $\|\widehat{M} - M\|_F^2/\|M\|_F^2$ for varied values of r, $p = 0.2$, $K = 5$, $\alpha = 1$, Gaussian noise with $\sigma = 0.18$.

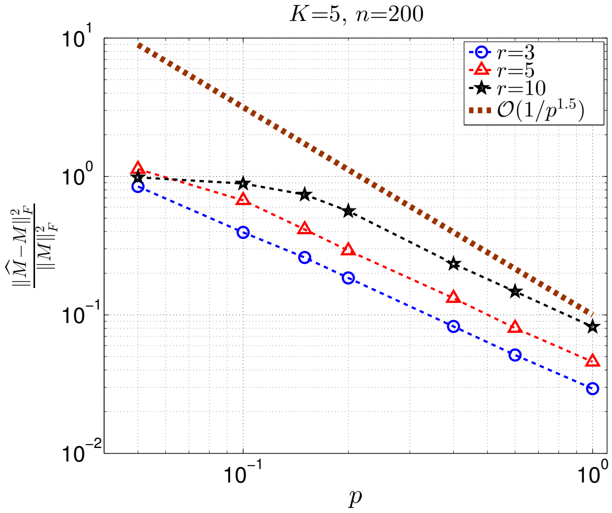


Figure 7: Relative MSE $\|\widehat{M} - M\|_F^2/\|M\|_F^2$ for varied values of p, $n = 200$, $K = 5$, $\alpha = 1$, Gaussian noise with $\sigma = 0.18$.

5.2 MovieLens Dataset

Now we consider the MovieLens 1M dataset (available from http://www.grouplens.org) consisting of 1,000,000 movie ratings on a scale from 1 to 5, from 6040 users on 3952 movies (95.8% missing entries). A given set of ratings has a matrix representation with rows representing the users and columns representing the movies, and the (i, j)th entry of the matrix is non-zero if user i has given a rating for movie j. Thus estimating the remaining ratings in the matrix corresponds to a matrix completion problem. We consider

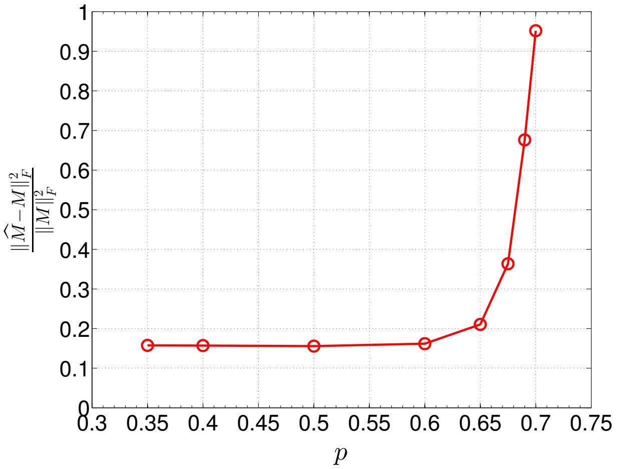


Figure 8: Relative MSE versus p, known bin boundaries, fixed $p+q = 0.7$, $K = 5$, $n = 200$, Gaussian noise $\sigma = 0.18$.

20 independent realizations of 80%/20% training/test splits of the 1 million revealed entries. For each data split, we train the proposed Algorithms 1 and 3, and the approaches OptSpace (Keshavan et al., 2009, 2010) and SVT (Cai et al., 2010), on the training set and compare their performance on the corresponding test set in predicting the revealed matrix entries in the test set. For implementation of OptSpace and SVT, we used the code made available by the authors online. Let Ω_T denote the test set, Y_{ij} the original rating in the data set and $\widehat{Y}_{ij}$ the predicted rating of user i for movie j. For performance assessment, we use the normalized mean absolute error (NMAE) and the root mean-square error (RMSE) in prediction on test set, defined as

$$RMSE = \sqrt{\frac{1}{|\Omega_T|} \sum_{(i,j)\in\Omega_T} \left(Y_{ij} - \widehat{Y}_{ij}\right)^2},$$

$$NMAE = \frac{1}{|\Omega_T|(Y_{max} - Y_{min})} \sum_{(i,j)\in\Omega_T} |Y_{ij} - \widehat{Y}_{ij}|,$$

where Y_{max} and Y_{min} are the upper and lower bounds, respectively, on the ratings (5 and 1, respectively). Both metrics are widely used for evaluation of prediction accuracy (Gunawardana and Shani, 2009); for instance, RMSE has been used in Salakhutdinov and Mnih (2008) and NMAE in Keshavan et al. (2009).

Remark 2 *Estimation of missing entries of Y. In many applications such as this one, one may wish to estimate missing discrete Y_{ij}s instead of (or in addition to) the continuous-valued M. We will use the model (1). If one wishes to pick the estimate $\widehat{Y}_{ij}$ of Y_{ij} given M_{ij}, to minimize the MSE $\mathbb{E}\{[\widehat{Y}_{ij} - Y_{ij}]^2\}$, then $\widehat{Y}_{ij} = \mathbb{E}\{Y_{ij} \mid M_{ij}\} = \sum_{\ell=1}^{K} \ell\, f_\ell(M_{ij})$. If, on the other hand, the optimality criterion is the MAE $\mathbb{E}\{|\widehat{Y}_{ij} - Y_{ij}|\}$, then given M_{ij}, $\widehat{Y}_{ij}$ is a*

MovieLens 1M		
Approach	RMSE	NMAE
Alg. 1 (proj-grad): logistic	0.8698 ± 0.0029	0.1590 ± 0.0004
Alg. 3 (log-barrier + unknown bins): logistic	**0.8568** ± 0.0014	**0.1559** ± 0.0004
Alg. 3 (log-barrier + unknown bins): probit	0.8580 ± 0.0027	0.1561 ± 0.0005
OptSpace	0.8947 ± 0.0033	0.1767 ± 0.0006
SVT	0.9023 ± 0.0014	0.1754 ± 0.0003

Table 1: Test data RMSE and NMAE (± one standard deviation) averaged over 20 realizations of 80%/20% training/test splits drawn from MovieLens 1M data

median of conditional distribution $f_\ell(M_{ij})$. For the model (1)-(4), if $\Phi(0) = 0.5$ *(true for the logistic and probit models considered in this paper), then a median of $f_\ell(M_{ij})$ is given by $\mathcal{Q}(M_{ij})$. These estimators are used with M_{ij} replaced with the estimated $\widehat{M_{ij}}$.*

For Algorithm 1 we used $\alpha = 1$, $\text{rank}(M) = 7$, the logistic model with $\sigma = 1/16$, and considered fixed bin boundaries $\omega_1^0 = -0.6$, $\omega_2^0 = -0.2$, $\omega_3^0 = 0.2$ and $\omega_4^0 = 0.6$ spaced (arbitrarily) uniformly over $[-\alpha, \alpha]$ to get equal width bins. An alternative initialization would be to pick the bin boundaries to match the distribution of the revealed entries. Let p_ℓ denote the fraction of revealed entries with $Y_{ij} = \ell$. Then a reasonable choice is to satisfy $\Phi(\omega_\ell) - \Phi(\omega_{\ell-1}) = p_\ell$, leading to $\omega_0^0 = -\infty$, $\omega_L^0 = \infty$, and $\omega_\ell^0 = \Phi^{-1}(\sum_{i=1}^{\ell} p_i)$ for $\ell = 1, \cdots, L-1$. For a given choice of α, this requires a proper choice of σ to ensure that ω_ℓ, $\ell = 1, \cdots, L-1$, are within $[-\alpha, \alpha]$. Note that scaling the variables, M, α, σ, and the bin boundaries by the same factor will lead to the same likelihood for observed data. Since we fixed (arbitrarily) $\alpha = 1$, different values of σ, the bin boundaries, and X will lead to a different likelihood for the observed data.

For Algorithm 3, based on additional optimization w.r.t. $\boldsymbol{\omega}$, we used $\alpha = 1$, $\text{rank}(M) = 7$, initialization $\omega_1^0 = -0.6$, $\omega_2^0 = -0.2$, $\omega_3^0 = 0.2$ and $\omega_4^0 = 0.6$, and either the logistic model with $\sigma = 1/16$, or the probit model with $\sigma = 1/13$. The results (RMSE and NMAE) averaged over 20 runs are shown in Table 1. It is seen that our methods outperform OptSpace and SVT under both performance measures, and fixed bin boundaries also yield useful and improved predictions. Tran et al. (2012) reported the results shown in Table 2 for their Matrix Cumulative RBM (restricted Boltzmann machines) based method and the OrdRec method of Koren and Sill (2011, 2013) when tested on the MovieLens 1M dataset; both these approaches are based on a quantization observation model. Comparing Tables 1 and 2 we see that our methods outperform OrdRec and Matrix Cumulative RBM under both performance measures.

Acknowledgments

I would like to thank Andrea Montanari and Adel Javanmard for helpful discussions.

MovieLens 1M: Results from Tran et al. (2012)						
	r=50		r=100		r=200	
Approach	RMSE	NMAE	RMSE	NMAE	RMSE	NMAE
OrdRec	0.904	0.1705	0.902	0.1705	0.902	0.1700
Matrix Cumulative RBM	0.904	0.1665	0.904	0.1655	0.906	0.1660

Table 2: RMSE and NMAE results from Tran et al. (2012) for MovieLens 1M data for various values of r=rank(M). OrdRec is the method of Koren and Sill (2011, 2013) and Matrix Cumulative RBM is the restricted Boltzman machines based method of Tran et al. (2012)

Appendix A. Proof of a Technical Lemma

Here we provide a proof of the assertion made in Section 3.1 (after Equation 9) that $f_\ell(X_{ij})$ is log-concave in ω_k for fixed X and ω_is ($i \neq k$) for log-concave $\Phi(x)$.

Lemma 3 *The probability $f_\ell(X_{ij})$ defined in (4) is log-concave in ω_k for fixed X and $\omega_i s$ ($i \neq k$) for log-concave $\Phi(x)$.*

Proof We have $f_\ell(x) = \Phi(\omega_\ell - x) - \Phi(\omega_{\ell-1} - x)$. Therefore, it is obviously log-concave in ω_i, $i \in [K-1]$, $i \neq \ell$ or $\ell - 1$. By p. 121, Problem 3.48, of Boyd and Vandenberghe (2004) $\Phi(\omega_\ell - x) - \Phi(\omega_{\ell-1} - x)$ is log-concave in ω_ℓ for fixed x and $\omega_{\ell-1}$. To show that $\Phi(\omega_\ell - x) - \Phi(\omega_{\ell-1} - x)$ is log-concave in $\omega_{\ell-1}$ for fixed x and ω_ℓ, we will modify a proof given in Prop. 1 of An (1995) to prove log-concavity of $1 - \Phi(x)$ in x. Let $\phi(x) = \frac{d\Phi(x)}{dx} \geq 0$. Then with $y_0 = \omega_\ell - x$ and $y = \omega_{\ell-1} - x$, we have $y_0 > y$ for every x,

$$s(y) := \Phi(\omega_\ell - x) - \Phi(\omega_{\ell-1} - x) = \int_y^{y_0} \phi(u)\, du \geq 0$$

and

$$h(y) := -\frac{d \log s(y)}{dy} = \frac{\phi(y)}{s(y)}.$$

By Prop. 1 of An (1995), $s(y)$ is log-concave iff $h(y)$ is non-decreasing is y. For $y_1 < y_2$, we have

$$\begin{aligned}
& h(y_2) - h(y_1) \geq 0 \\
\iff & \phi(y_2)s(y_1) - \phi(y_1)s(y_2) \\
& = \phi(y_2)\int_{y_1}^{y_0} \phi(u)\, du - \phi(y_1)\int_{y_2}^{y_0} \phi(u)\, du \\
& = \int_0^{y_0 - y_2} \left[\phi(y_2)\phi(y_1 + v) - \phi(y_1)\phi(y_2 + v)\right] dv + \int_{y_0 - y_2}^{y_0 - y_1} \phi(y_2)\phi(y_1 + v) dv \\
& \geq 0
\end{aligned}$$

where we have used the fact that $\int_{y_0-y_2}^{y_0-y_1} \phi(y_2)\phi(y_1+v)dv \geq 0$ since the integrand is non-negative and $y_1 < y_2$, and since $\phi(x)$ is log-concave, by Lemma 1 of An (1995)

$$\int_0^{y_0-y_2} \left[\phi(y_2)\phi(y_1+v) - \phi(y_1)\phi(y_2+v)\right] dv \geq 0.$$

Thus, for fixed x and ω_ℓ, $s(y)$ is log-concave in y, hence, in $\omega_{\ell-1}$. ■

Appendix B. Proof of Theorem 1

Our proof is based on a second-order Taylor series expansion and a matrix concentration inequality. Let $\theta = \mathrm{vec}(X) \in \mathbb{R}^{mn}$ and $\tilde{F}_{\Omega,Y}(\theta) = F_{\Omega,Y}(X)$. The objective function $F_{\Omega,Y}(X)$ is continuous in X and the set $\mathcal{C}$ is compact, therefore, $F_{\Omega,Y}(X)$ achieves a minimum in $\mathcal{C}$. If $\hat{\theta} = \mathrm{vec}(\widehat{M})$ minimizes $\tilde{F}_{\Omega,Y}(\theta)$ subject to the constraints, then $\tilde{F}_{\Omega,Y}(\hat{\theta}) \leq \tilde{F}_{\Omega,Y}(\theta^*)$ where $\theta^* = \mathrm{vec}(M)$. By the second-order Taylor's theorem, expanding around θ^* we have

$$\tilde{F}_{\Omega,Y}(\theta) = \tilde{F}_{\Omega,Y}(\theta^*) + \langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), \theta - \theta^* \rangle + \frac{1}{2}\langle \theta - \theta^*, \left(\nabla^2_{\theta\theta}\tilde{F}_{\Omega,Y}(\tilde{\theta})\right)(\theta - \theta^*)\rangle \tag{33}$$

where $\tilde{\theta} = \theta^* + \gamma(\theta - \theta^*)$ for some $\gamma \in [0,1]$, with corresponding matrices $\tilde{X} = M + \gamma(X - M)$. We need several auxiliary results before we can prove Theorem 1.

We need the following result from Chatterjee (2013) concerning spectral norms of random matrices for Lemma 5.

Lemma 4 *(Theorem 8.4 of Chatterjee (2013)) Take any two numbers m and n such that $1 \leq n \leq m$. Suppose that $A = [a_{ij}]_{1\leq i\leq m, 1\leq j\leq n}$ is a matrix whose entries are independent random variables that satisfy, for some $\sigma^2 \in [0,1]$,*

$$\mathbb{E}[a_{ij}] = 0,\ \mathbb{E}[a_{ij}^2] \leq \sigma^2,\ \text{and } |a_{ij}| \leq 1 \text{ a.s.}$$

Suppose that $\sigma^2 \geq m^{-1+\varepsilon}$ for some $\varepsilon > 0$. Then

$$P\left(\|A\|_2 \geq 2.01\sigma\sqrt{m}\right) \leq C_1(\varepsilon)e^{-C_2\sigma^2 m},$$

where $C_1(\varepsilon)$ is a constant that depends only on ε and C_2 is a positive universal constant. The same result is true when $m = n$ and A is symmetric or skew-symmetric, with independent entries on and above the diagonal, all other assumptions remaining the same. Lastly, all results remain true if the assumption $\sigma^2 \geq m^{-1+\varepsilon}$ is changed to $\sigma^2 \geq m^{-1}(\log(m))^{6+\varepsilon}$.

Using (6), it follows that

$$\frac{\partial F_{\Omega,Y}(X)}{\partial X_{ij}} = -\left(\sum_{\ell=1}^{K} \frac{\dot{f}_\ell(X_{ij})}{f_\ell(X_{ij})}\mathbf{1}_{[Y_{ij}=\ell]}\right)\mathbf{1}_{[(i,j)\in\Omega]}, \tag{34}$$

$$\frac{\partial^2 F_{\Omega,Y}(X)}{\partial X_{ij}^2} = \sum_{\ell=1}^{K}\left(\frac{\dot{f}_\ell^2(X_{ij})}{f_\ell^2(X_{ij})} - \frac{\ddot{f}_\ell(X_{ij})}{f_\ell(X_{ij})}\right)\mathbf{1}_{[Y_{ij}=\ell]}\mathbf{1}_{[(i,j)\in\Omega]} \tag{35}$$

and

$$\frac{\partial^2 F_{\Omega,Y}(X)}{\partial X_{i_1 j_1} \partial X_{i_2 j_2}} = 0 \text{ if } (i_1, j_1) \neq (i_2, j_2). \tag{36}$$

Let $w = \operatorname{vec}(X^{(1)} - M) = \theta^{(1)} - \theta^*$; for later use, we would like $\theta^{(1)}$ in w to be not necessarily equal to θ in the gradient or the Hessian of the objective function. Using (34) and the notation

$$\nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*) = \operatorname{vec}\left(\left[\frac{\partial F_{\Omega,Y}(X)}{\partial X_{ij}} \right] \Big|_{X=M} \right) = \operatorname{vec}\left(\left[\frac{\partial F_{\Omega,Y}(M)}{\partial X_{ij}} \right] \right),$$

we have

$$\langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), w \rangle = \langle \nabla_X F_{\Omega,Y}(M), X^{(1)} - M \rangle \tag{37}$$

where $\langle A, B \rangle := \operatorname{tr}(A^\top B)$, $|\langle A, B \rangle| \leq \|A\|_2 \|B\|_*$, $\|B\|_*$ is the nuclear (or Schatten) norm of B, and

$$[\nabla_X F_{\Omega,Y}(M)]_{ij} =: z_{ij} = -\left(\sum_{\ell=1}^{K} \frac{\dot{f}_\ell(M_{ij})}{f_\ell(M_{ij})} \mathbf{1}_{[Y_{ij}=\ell]} \right) \mathbf{1}_{[(i,j)\in\Omega]} . \tag{38}$$

Using (1), (11), and the fact that $\sum_{\ell=1}^{K} f_\ell(X_{ij}) = 1$, we have

$$\mathbb{E}[z_{ij}] = -\left(\sum_{\ell=1}^{K} \dot{f}_\ell(M_{ij}) \right) \mathbf{1}_{[(i,j)\in\Omega]} = 0, \tag{39}$$

and

$$|z_{ij}| \leq L_\alpha \implies \mathbb{E}[z_{ij}^2] \leq L_\alpha^2 . \tag{40}$$

Lemma 5 *Let $w = \operatorname{vec}(X^{(1)} - M) = \theta^{(1)} - \theta^*$ and $X^{(1)}, M \in \mathcal{C}$. Then with probability at least $1 - C_1(\varepsilon)\exp(-C_2 m)$, we have*

$$\left| \langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), w \rangle \right| \leq 2.01 L_\alpha \sqrt{2rm} \|X^{(1)} - M\|_F ,$$

where $\varepsilon \in (0,1)$, $C_1(\varepsilon)$ is a constant that depends only on ε and C_2 is a positive universal constant.

Proof Using (37), we have

$$\begin{aligned} |\langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), w \rangle| &= |\langle \nabla_X F_{\Omega,Y}(M), X^{(1)} - M \rangle| \\ &\leq \|\nabla_X F_{\Omega,Y}(M)\|_2 \|X^{(1)} - M\|_*. \end{aligned} \tag{41}$$

Consider $\tilde{z}_{ij} := \left[L_\alpha^{-1} \nabla_X F_{\Omega,Y}(M) \right]_{ij}$. Then we have $\mathbb{E}[\tilde{z}_{ij}] = 0$, $|\tilde{z}_{ij}| \leq 1$ and $\mathbb{E}[\tilde{z}_{ij}^2] \leq 1$. We will apply Lemma 4 to $L_\alpha^{-1} \nabla_X F_{\Omega,Y}(M)$, for which we have to ensure that $\mathbb{E}[\tilde{z}_{ij}^2] \leq \sigma^2$ and $m^{-1+\varepsilon} \leq \sigma^2$ for some $\varepsilon > 0$. Therefore, we pick $\sigma^2 = 1 = \max\left(1, \frac{1}{m^{1-\varepsilon}}\right)$. Hence, by Lemma 4, $\|L_\alpha^{-1} \nabla_X F_{\Omega,Y}(M)\|_2 \leq 2.01\sqrt{m}$ with probability at least $1 - C_1(\varepsilon)\exp(-C_2 m)$ for some positive constants $C_1(\varepsilon)$ and C_2. Since for any rank r matrix A, $\|A\|_* \leq \sqrt{r}\|A\|_F$, we have $\|X^{(1)} - M\|_* \leq \sqrt{2r}\|X^{(1)} - M\|_F$, yielding the desired result. ■

Lemma 6 *Let $w = \text{vec}(X - M) = \theta - \theta^*$ and $X, M \in \mathcal{C}$. Then with probability at least $1 - 2(9\alpha\sqrt{mn})^{-r(m+n+1)} - C_1 \exp(-C_2 m)$, we have*

$$\left|\langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), w\rangle\right| \leq 4(1+\alpha)L_\alpha \sqrt{|\Omega| r(m+n+1)\log(9\alpha\sqrt{mn})}\,.$$

Proof Define the set $S_{rK} = \{X \in \mathbb{R}^{m\times n} : \text{rank}(X) \leq r,\ \|X\|_F \leq K\}$ for some $K > 0$. By Lemma A.2 in the supplementary material of Wang and Xu (2012) (which is based on Lemma 3.1 in Candes and Plan (2011)), there exists a 1-net for the Frobenius norm, $\overline{S}_{rK}(1) = \{Q \in \mathbb{R}^{m\times n} : \text{rank}(Q) \leq r,\ \|X - Q\|_F \leq 1\} \subset S_{rK}$ with its cardinality $|\overline{S}_{rK}(1)| \leq (9K)^{(m+n+1)r}$. That is, given any $X \in S_{rK}$, there exists $Q \in \overline{S}_{rK}(1) \subset S_{rK}$ such that $\|X - Q\|_F \leq 1$. Suppose that $X \in S_{rK}$ is such that $\|X\|_\infty \leq \alpha$. Then for $Q \in \overline{S}_{rK}(1)$, we have

$$\|Q\|_\infty \leq \|Q - X\|_\infty + \|X\|_\infty \leq \|Q - X\|_F + \|X\|_\infty \leq 1 + \alpha. \tag{42}$$

Also, $\|X\|_F \leq \sqrt{mn}\|X\|_\infty \leq \alpha\sqrt{mn}$. For $X, M \in S_{rK}$ and $Q \in \overline{S}_{rK}(1)$, consider

$$\begin{aligned} u_X &:= \langle \nabla_\theta \tilde{F}_{\Omega,Y}(\theta^*), w\rangle = \langle \nabla_X F_{\Omega,Y}(M), X - M\rangle \\ &= \langle \nabla_X F_{\Omega,Y}(M), Q - M\rangle + \langle \nabla_X F_{\Omega,Y}(M), X - Q\rangle = u_Q + \langle \nabla_X F_{\Omega,Y}(M), X - Q\rangle. \end{aligned} \tag{43}$$

Therefore, for any $X \in S_{rK}$ and corresponding $Q \in \overline{S}_{rK}(1)$, we have

$$\begin{aligned} |u_X| &\leq |u_Q| + |\langle \nabla_X F_{\Omega,Y}(M), X - Q\rangle| \\ &\leq |u_Q| + \|\nabla_X F_{\Omega,Y}(M)\|_2 \|X - Q\|_* \\ &\leq |u_Q| + 2.01 L_\alpha \sqrt{2rm} \ \text{ with probability } \geq 1 - C_1\exp(-C_2 m) \end{aligned} \tag{44}$$

where in the last inequality we have used Lemma 5 with $C_1 = C_1(1/2)$, and the fact that $\|X - Q\|_F \leq 1$ and both X and Q are of rank r. Now consider u_Q and rewrite it as

$$u_Q = \sum_{(i,j)\in\Omega} h_{ij} \quad \text{where} \quad h_{ij} = \frac{\partial F_{\Omega,Y}(M)}{\partial X_{ij}}\,(Q_{ij} - M_{ij})\,.$$

We have $\mathbb{E}[h_{ij}] = 0$ (see Equations 38 and 39), and

$$\begin{aligned} |h_{ij}| &\leq |\frac{\partial F_{\Omega,Y}(M)}{\partial X_{ij}}| \times |Q_{ij} - M_{ij}| \\ &\leq L_\alpha(|Q_{ij}| + |M_{ij}|) \leq L_\alpha(1 + 2\alpha) =: \beta_\alpha. \end{aligned} \tag{45}$$

Apply the Hoeffding inequality to u_Q to obtain

$$\mathbb{P}\left(|u_Q| > t\right) \leq 2\exp\left(-2\frac{t^2}{|\Omega|\beta_\alpha^2}\right).$$

Set $K = \alpha\sqrt{mn}$ (since $\|X\|_F \leq \alpha\sqrt{mn}$) and apply the union bound over all $Q \in \overline{S}_{rK}(1)$ to obtain

$$\begin{aligned} \mathbb{P}\left(\cup_{Q\in\overline{S}_{rK}(1)} \{|u_Q| > t\}\right) &\leq 2\left(9\alpha\sqrt{mn}\right)^{(m+n+1)r} \exp\left(-2\frac{t^2}{|\Omega|\beta_\alpha^2}\right) \\ &\leq 2\exp\left(-\frac{2t^2}{|\Omega|\beta_\alpha^2} + (m+n+1)r\log\left(9\alpha\sqrt{mn}\right)\right). \end{aligned}$$

We pick

$$t = \beta_\alpha \sqrt{|\Omega|(m+n+1)r \log\left(9\alpha\sqrt{mn}\right)} \tag{46}$$

to achieve

$$\begin{aligned}\mathbb{P}\left(\cup_{Q\in\overline{S}_{rK}(1)} \{|u_Q| > t\}\right) &\leq 2\exp\left(-(m+n+1)r\log\left(9\alpha\sqrt{mn}\right)\right)\\ &= \frac{2}{(9\alpha\sqrt{mn})^{r(m+n+1)}}. \end{aligned}\tag{47}$$

Now using (44)-(47), the union bound and the fact that the chosen $t > 2.01 L_\alpha\sqrt{2rm}$, we have the desired result. ■

Lemma 7 *Let* $w = \mathrm{vec}(X^{(1)} - M) = \theta^{(1)} - \theta^*$ *and* $X^{(1)}, X, M \in \mathcal{C}$. *Then for any* $\tilde{\theta} = \theta^* + \gamma(\theta - \theta^*)$ *and any* $\gamma \in [0,1]$, *we have*

$$\langle w, \left[\nabla^2_{\theta\theta}\tilde{F}_{\Omega,Y}(\tilde{\theta})\right] w\rangle \geq \gamma_\alpha \left\|\left(X^{(1)} - M\right)_\Omega\right\|_F^2$$

where $X_\Omega = X_{ij}$ *if* $(i,j)\in\Omega$, *and* $= 0$ *otherwise.*

Proof Using (10), (35) and (36), we have

$$\begin{aligned}\langle w, \left[\nabla^2_{\theta\theta}\tilde{F}_{\Omega,Y}(\tilde{\theta})\right] w\rangle &= \sum_{(i,j)\in\Omega}\left(\frac{\partial^2 F_{\Omega,Y}(\tilde{X})}{\partial X_{ij}^2}\right)(X_{ij}^{(1)} - M_{ij})^2\\ &\geq \gamma_\alpha \sum_{(i,j)\in\Omega}(X_{ij}^{(1)} - M_{ij})^2 = \gamma_\alpha\left\|\left(X^{(1)} - M\right)_\Omega\right\|_F^2. \end{aligned}\tag{48}$$

■

We need a result similar to Theorem 4.1 of Bhojanapalli and Jain (2014) regarding closeness of a fixed matrix to its sampled version, which is proved therein for square matrices M under an incoherence assumption on M. In Lemma 8 we prove a similar result for rectangular Z with bounded $\|Z\|_\infty$. Take $Z\in\mathbb{R}^{m\times n}$, $m \geq n$, and as in Lemma 7, define the operator $\mathcal{R}_\Omega$ as $Z_\Omega := \mathcal{R}_\Omega(Z) = Z_{ij}$ if $(i,j)\in\Omega$, and $= 0$ otherwise.

Lemma 8 *Let* $G\backslash\Omega$ *satisfy assumptions (A1) and (A2) in Section 2. Let* $Z\in\mathbb{R}^{m\times n}$ *have rank* $\leq r$. *Then we have*

$$\left\|\left(\frac{\sqrt{mn}}{\sigma_1(G)}R_\Omega - I\right)(Z)\right\|_2 \leq \frac{\sqrt{mn}\sigma_2(G)}{\sigma_1(G)}\|Z\|_{\max} \tag{49}$$

$$\leq \frac{\sqrt{rmn}\sigma_2(G)}{\sigma_1(G)}\|Z\|_\infty \leq Cm\sqrt{\frac{nr}{|\Omega|}}\|Z\|_\infty. \tag{50}$$

Proof Normalize $\mathbf{1}_m$ to unit norm as $\tilde{\mathbf{1}}_m = \mathbf{1}_m/\sqrt{m}$, and similarly for $\tilde{\mathbf{1}}_n$. It then follows from the properties (A1)-(A2) that

$$G = \sigma_1(G)\tilde{\mathbf{1}}_m\tilde{\mathbf{1}}_n^\top + \sum_{i=2}^{n} \sigma_i(G) U_i V_i^\top \tag{51}$$

where the SVD of G is $G = U\Sigma_G V^\top$ and U_i is the i-th column of U. First some preliminaries. If $\mathrm{rank}(Z) \le r$, then we have $\|Z\|_{\max} \le \sqrt{r}\|Z\|_\infty$. By the factored form definition of the max norm (Lee et al., 2010), we have $\|Z\|_{\max} = \inf\Big\{ \max(\|\bar{U}\|_{2,\infty}^2, \|\bar{V}\|_{2,\infty}^2) : Z = \bar{U}\bar{V}^\intercal \Big\}$ where $\|\bar{U}\|_{2,\infty} = \max_i \sqrt{\sum_j \bar{U}_{ij}^2}$, $\bar{U} \in \mathbb{R}^{m\times k}$, $\bar{V} \in \mathbb{R}^{n\times k}$, $k = 1, 2, \cdots, \min(m,n) = n$. Hence, there exist $U_Z \in \mathbb{R}^{m\times k}$ and $V_Z \in \mathbb{R}^{n\times k}$ for some $1 \le k \le \min(m,n)$ such that $Z = U_Z V_Z^\top$, $\|U_Z\|_{2,\infty}^2 \le \|Z\|_{\max}$ and $\|V_Z\|_{2,\infty}^2 \le \|Z\|_{\max}$. For Z of rank $\le r$, one should have $k \le r$, but this fact is not needed in our proof. We now follow the proof of Theorem 4.1 of Bhojanapalli and Jain (2014) with $Z = \sum_{i=1}^k U_{Zi}V_{Zi}^\top$. Note that

$$\|\frac{\sqrt{mn}}{\sigma_1(G)} R_\Omega(Z) - Z\|_2 = \max_{x,y\colon \|x\|_2 = 1 = \|y\|_2} y^\top \left(\frac{\sqrt{mn}}{\sigma_1(G)} R_\Omega(Z) - Z \right) x.$$

As in the proof of Theorem 4.1 of Bhojanapalli and Jain (2014), noting that $R_\Omega(Z) = Z \circ G$ where $\circ$ denotes the Hadamard (elementwise) product, we have

$$y^\top \left(\frac{\sqrt{mn}}{\sigma_1(G)} \mathcal{R}_\Omega(Z) - Z \right) x = \sum_{i=1}^{k} \left(\frac{\sqrt{mn}}{\sigma_1(G)} (y \circ U_{Zi})^\top G (x \circ V_{Zi}) - (y^\top U_{Zi})(x^\top V_{Zi}) \right). \tag{52}$$

Let $y \circ U_{Zi} = \alpha_i \tilde{\mathbf{1}}_m + \beta_i \tilde{\mathbf{1}}_{m\perp}^i$ where $\tilde{\mathbf{1}}_{m\perp}^i$ is a unit norm vector orthogonal to $\tilde{\mathbf{1}}_m$. Then $\alpha_i = \tilde{\mathbf{1}}_m^\top (y \circ U_{Zi}) = y^\top U_{Zi}/\sqrt{m}$. Using the fact that $\tilde{\mathbf{1}}_m^\top G = \sigma_1(G)\tilde{\mathbf{1}}_n^\top$, we have

$$\begin{aligned}
& y^\top \left(\frac{\sqrt{mn}}{\sigma_1(G)} \mathcal{R}_\Omega(Z) - Z \right) x \\
&\quad = \sum_{i=1}^{k} \Big(\frac{\sqrt{mn}}{\sigma_1(G)} \Big[(1/\sqrt{m}) y^\top U_{Zi} \tilde{\mathbf{1}}_m^\top G (x \circ V_{Zi}) + \beta_i \tilde{\mathbf{1}}_{m\perp}^{i\top} G (x \circ V_{Zi}) \Big] - (y^\top U_{Zi})(x^\top V_{Zi}) \Big) \\
&\quad = \sum_{i=1}^{k} \left(\frac{\sqrt{mn}}{\sigma_1(G)} \beta_i \tilde{\mathbf{1}}_{m\perp}^{i\top} G (x \circ V_{Zi}) \right) \qquad (53)
\end{aligned}$$

where we have also used $\tilde{\mathbf{1}}_n^\top (x \circ V_{Zi}) = x^\top V_{Zi}/\sqrt{n}$. Using the SVD (51) of G, we have

$$\tilde{\mathbf{1}}_{m\perp}^{i\top} G = \sum_{\ell=2}^{n} \sigma_\ell(G) (\tilde{\mathbf{1}}_{m\perp}^{i\top} U_\ell) V_\ell^\top$$

$$\implies |\tilde{\mathbf{1}}_{m\perp}^{i\top} G z| \le \sigma_2(G) \|z\|_2 \text{ for any } z \in \mathbb{R}^n.$$

Using the above inequality in (53) we obtain

$$\begin{aligned} y^\top \left(\frac{\sqrt{mn}}{\sigma_1(G)} R_\Omega(Z) - Z \right) x &\leq \frac{\sqrt{mn}}{\sigma_1(G)} \sigma_2(G) \sum_{i=1}^k |\beta_i| \|x \circ V_{Zi}\|_2 \\ &\leq \frac{\sqrt{mn}}{\sigma_1(G)} \sigma_2(G) \sqrt{\sum_{i=1}^k \beta_i^2} \sqrt{\sum_{i=1}^k \|x \circ V_{Zi}\|_2^2} \,. \end{aligned} \tag{54}$$

We have $\beta_i = \tilde{\mathbf{1}}_{m\perp}^{i\top}(y \circ U_{Zi})$, hence, $|\beta_i| \leq \|(y \circ U_{Zi})\|_2$. Therefore,

$$\begin{aligned} \sum_{i=1}^k \beta_i^2 \leq \sum_{i=1}^k \|(y \circ U_{Zi})\|_2^2 &= \sum_{j=1}^m \sum_{i=1}^k y_j^2 U_{Zji}^2 \\ &\leq \sum_{j=1}^m y_j^2 \|U_Z^j\|_2^2 \leq \|U_Z\|_{2,\infty}^2 \sum_{j=1}^m y_j^2 \leq \|Z\|_{\max} \end{aligned} \tag{55}$$

where U_Z^j denotes the jth row of U_Z and $\sum_{j=1}^m y_j^2 = 1$. Similarly, we have

$$\begin{aligned} \sum_{i=1}^k \|x \circ V_{Zi}\|_2^2 = \sum_{j=1}^n \sum_{i=1}^k x_j^2 V_{Zji}^2 &\leq \sum_{j=1}^n x_j^2 \|\tilde{V}_Z^j\|_2^2 \\ &\leq \|\tilde{V}_Z\|_{2,\infty}^2 \sum_{j=1}^n x_j^2 \leq \|Z\|_{\max} \,. \end{aligned} \tag{56}$$

It then follows from (54)-(56) that

$$\begin{aligned} &y^\top \left(\frac{\sqrt{mn}}{\sigma_1(G)} R_\Omega(Z) - Z \right) x \leq \frac{\sqrt{mn}\sigma_2(G)}{\sigma_1(G)} \|Z\|_{\max} \\ \implies &\| \frac{\sqrt{mn}}{\sigma_1(G)} R_\Omega(Z) - Z \|_2 \leq \frac{\sqrt{mn}\sigma_2(G)}{\sigma_1(G)} \|Z\|_{\max}. \end{aligned} \tag{57}$$

This establishes (49). Now use the facts $\|Z\|_{\max} \leq \sqrt{r}\|Z\|_\infty$ and $|\Omega| = md$ to establish (50). ■

Lemma 9 *Let $X, M \in \mathcal{C}$. Then we have*

$$\|(X-M)_\Omega\|_F \geq \frac{\sigma_1(G)}{\sqrt{2rmn}} \|X - M\|_F - 2\alpha\sqrt{r}\sigma_2(G).$$

Proof Let $Z = X - M$, $a = \sqrt{mn}/\sigma_1(G)$, and $b = (\sigma_2(G)/\sigma_1(G))\sqrt{rmn}$. Then by Lemma 8 and the fact that $\text{rank}(Z) \leq \text{rank}(X) + \text{rank}(M) \leq 2r$, we have

$$|a\|Z_\Omega\|_2 - \|Z\|_2| \leq \|aZ_\Omega - Z\|_2 \leq b\|Z\|_\infty. \tag{58}$$

Using $\|Z\|_\infty = \|X - M\|_\infty \leq \|X\|_\infty + \|M\|_\infty \leq 2\alpha$, (58) can be expressed as $\|Z\|_2 \leq a\|Z_\Omega\|_2 + 2\alpha b$. Since $\|A\|_2 \leq \|A\|_F \; \forall A$, we then have $\|Z\|_2 \leq a\|Z_\Omega\|_F + 2\alpha b$. Since $\|A\|_F \leq \sqrt{\text{rank}(A)}\|A\|_2 \; \forall A$, we have $\|Z\|_F \leq \sqrt{2r}\|Z\|_2 \leq (\sqrt{2r}a)\|Z_\Omega\|_F + \sqrt{2r}2\alpha b$, leading to the desired result. ■

We now turn to the proof of Theorem 1.

Proof of Theorem 1 The bound 2α follows from the fact that $\widehat{M}, M \in \mathcal{C}$. To establish bound U_1, we will use Lemma 5 and to establish U_2, we will use Lemma 6. We first prove U_1. Consider $\tilde{F}_{\Omega,Y}(\theta) = F_{\Omega,Y}(X)$. The objective function $F_{\Omega,Y}(X)$ is continuous in X and the set $\mathcal{C}$ is compact, therefore, $F_{\Omega,Y}(X)$ achieves a minimum in $\mathcal{C}$. Now suppose that $\widehat{M} \in \mathcal{C}$ minimizes $F_{\Omega,Y}(X)$. Then $F_{\Omega,Y}(\widehat{M}) \leq F_{\Omega,Y}(X) \; \forall X \in \mathcal{C}$, including $X = M$. Define

$$c_g = 2.01 L_\alpha \sqrt{2rm}, \quad c_h = \frac{\sigma_1^2(G)\gamma_\alpha}{4rmn}, \quad \bar{c}_h = \frac{\gamma_\alpha}{2}. \tag{59}$$

Using (33) and Lemmas 5 and 7, we have w.h.p. (specified in Lemma 5)

$$F_{\Omega,Y}(\widehat{M}) \geq F_{\Omega,Y}(M) - c_g\|\widehat{M} - M\|_F + \bar{c}_h \left\| \left(\widehat{M} - M\right)_\Omega \right\|_F^2 . \tag{60}$$

Since $\widehat{M}$ minimizes $F_{\Omega,Y}(X)$, we have

$$\begin{aligned} 0 &\geq F_{\Omega,Y}(\widehat{M}) - F_{\Omega,Y}(M) \\ &\geq -c_g\|\widehat{M} - M\|_F + \bar{c}_h \left\| \left(\widehat{M} - M\right)_\Omega \right\|_F^2 . \end{aligned} \tag{61}$$

Set

$$\eta = 2\alpha r(\sigma_2(G)/\sigma_1(G))\sqrt{2mn} \quad \text{and} \quad a_0 = \sigma_1(G)/\sqrt{2rmn} \,.$$

Then Lemma 9 implies $\|(X - M)_\Omega\|_F \geq a_0 \left[\|X - M\|_F - \eta\right]$. Now consider two cases: (i) $\|\widehat{M} - M\|_F < 2\eta$, (ii) $\|\widehat{M} - M\|_F \geq 2\eta$. In case (i), we clearly have an obvious upper bound on $\|\widehat{M} - M\|_F$. Turning to case (ii), we have

$$\begin{aligned} \|\widehat{M} - M\|_F - \eta &\geq \|\widehat{M} - M\|_F - \frac{1}{2}\|\widehat{M} - M\|_F \\ &= \frac{1}{2}\|\widehat{M} - M\|_F. \end{aligned} \tag{62}$$

Using (61), (62) and Lemma 9 with $X = \widehat{M}$, we have

$$\begin{aligned} 0 &\geq F_{\Omega,Y}(\widehat{M}) - F_{\Omega,Y}(M) \\ &\geq -c_g\|\widehat{M} - M\|_F + \frac{c_h}{4}\|\widehat{M} - M\|_F^2 \\ &= \|\widehat{M} - M\|_F \left[-c_g + \frac{c_h}{4}\|\widehat{M} - M\|_F\right]. \end{aligned} \tag{63}$$

In order for (63) to be true, we must have $\|\widehat{M}-M\|_F \leq 4c_g/c_h$ otherwise the right-side of (63) is positive violating (63). Combining the two cases, we obtain

$$\begin{aligned}\|\widehat{M}-M\|_F &\leq \max\left(2\eta, \frac{4c_g}{c_h}\right)\\ &= \max\left(4\alpha r\sqrt{2mn}\frac{\sigma_2(G)}{\sigma_1(G)}, \frac{32.16\sqrt{2}L_\alpha(rm)^{1.5}n}{\gamma_\alpha\sigma_1^2(G)}\right)\ . \qquad (64)\end{aligned}$$

This is the bound U_1 stated in (13)-14) of the theorem after division by $\sqrt{mn}$. The high probability stated in the theorem follows from Lemma 5 after setting $\varepsilon = 0.5$. Finally, we use $(\sigma_2(G)/\sigma_1(G)) \leq (C/\sqrt{d}) = C\sqrt{m}/\sqrt{|\Omega|}$ and $(1/\sigma_1^2(G)) \leq (1/d^2) = m^2/|\Omega|^2$ to derive (14).

Finally we turn to proving U_2. Define

$$\bar{c}_g = 4(1+\alpha)L_\alpha\sqrt{|\Omega|r(m+n+1)\log(9\alpha\sqrt{mn})}. \qquad (65)$$

Using (33) and Lemmas 6 and 7, we have w.h.p. (specified in Lemma 6)

$$F_{\Omega,Y}(\widehat{M}) \geq F_{\Omega,Y}(M) - \bar{c}_g + \bar{c}_h\left\|\left(\widehat{M}-M\right)_\Omega\right\|_F^2\ . \qquad (66)$$

Arguing as earlier, we then have

$$\left\|\left(\widehat{M}-M\right)_\Omega\right\|_F \leq \sqrt{\frac{2\bar{c}_g}{\gamma_\alpha}}. \qquad (67)$$

As before, we have either $\|\widehat{M}-M\|_F < 2\eta$ or $\|\widehat{M}-M\|_F \geq 2\eta$; the former yields an obvious upper bound while the latter case yields

$$\left\|\widehat{M}-M\right\|_F \leq \frac{2}{a_0}\left\|\left(\widehat{M}-M\right)_\Omega\right\|_F \leq \frac{2}{a_0}\sqrt{\frac{2\bar{c}_g}{\gamma_\alpha}}. \qquad (68)$$

The stated bound U_2 in (15)-16) then follows just as U_1. This completes the proof. ■

References

M.Y. An. Log-concave probability distributions: Theory and statistical testing. Working Paper No. 95-03, Department of Economics, Duke University, Durham, North Carolina, 1995.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869v1*, 2008.

D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

S.A. Bhaskar. Quantized matrix completion for low rank matrices. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3741–3745, Brisbane, Queensland, Australia, 2015.

S. Bhojanapalli and P. Jain. Universal matrix completion. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

M. Bolla, K. Friedl, and A. Kramli. Singular value decomposition of large random matrices (for two-way classification of microarrays). *Journal Multivariate Analysis*, 101:434–446, 2010.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ Press, 2004.

S. Burer and R.D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming (series B)*, 95:329–357, 2003.

J.F. Cai, E.J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optimization*, 20(4):1956–1982, 2010.

T. Cai and W.-X. Zhou. A Max-Norm Constrained Minimization Approach to 1-Bit Matrix Completion. *Journal of Machine Learning Research*, 14:3619–3647, 2013.

E.J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57:2342–2359, 2011.

Y. Cao and Y. Xie. Categorical matrix completion. *arXiv preprint arXiv:1507.00421v1*, 2015.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv preprint arXiv:1212.1247v5*, 2013.

M.A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3:189–223, 2014.

U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27:251–275, 2005.

Y. Ghanbari, A.R. Smith, R.T. Schultz, and R. Verma. Connectivity subnetwork learning for pathology and developmental variations. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 90–97. Springer, 2013.

D.F. Gleich and L.-H. Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 60–68, 2011.

P. Gopalan, F.J.R. Ruiz, R. Ranganath, and D.M. Blei. Bayesian nonparametric poisson factorization for recommendation systems. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014.

S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1917–1925, 2014.

A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of American Mathematical Society*, 43(4):439–561, 2006.

A. Karbasi and S. Oh. Robust localization from incomplete local information. *IEEE/ACM Transactions on Networking*, 21:1131–1144, August 2013.

R.H. Keshavan, A. Montanari, and S. Oh. Low-rank matrix completion with noisy observations: a quantitative comparison. In *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, Urbana, Illinois, 2009.

R.H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Y. Koren and J. Sill. OrdRec: An ordinal method for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pages 117–124, Chicago, Illinois, 2011.

Y. Koren and J. Sill. Collaborative filtering on ordinal user feedback. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 3022–3026, Beijing, China, 2013.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

J. Lafond. Low rank matrix completion with exponential family noise. *arXiv preprint arXiv:1502.06919v2*, 2015.

J. Lafond, O. Klopp, E. Moulines, and J. Salmon. Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2014.

A.S. Lan, C. Studer, and R.G. Baraniuk. Matrix recovery from quantized and corrupted measurements. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014a.

A.S. Lan, C. Studer, and R.G. Baraniuk. Quantized matrix completion for personalized learning. In *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK, 2014b.

A.S. Lan, A.E. Waters, C. Studer, and R.G. Baraniuk. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15:1959–2008, 2014c.

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

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 42(2):109–142, 1980.

R.R. Nadakuditi and M.E.J. Newman. Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18):188701–5, 2012.

S. Negahban and M.J. Wainright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

B. Recht and C. Re. Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.*, 5(2):201–226, 2013.

J.D.M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 713–719, 2005.

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, 2008.

L.K. Saul and S.T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz. Localization from connectivity in sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 15(11):961–974, 2004.

A. Soni, S. Jain, J. Haupt, and S. Gonella. Noisy matrix completion under sparse factor models. *arXiv preprint arXiv:1411.0282v1*, 2014.

J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

T. Tran, D. Phung, and S. Venkatesh. Cumulative restricted Boltzmann machines for ordinal matrix data analysis. In *Proceedings of the 4th Asian Conference on Machine Learning*, volume 25 of *JMLR Workshop and Conference Proceedings*, pages 411–426, 2012.

Y.-X. Wang and H. Xu. Stability of matrix factorization in collaborative filtering. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sciences*, 6(3):1758–1789, 2013.