QUANTIZED MATRIX COMPLETION FOR LOW RANK MATRICES

Sonia A. Bhaskar

Department of Electrical Engineering
Stanford University, Stanford, CA 94304, USA
sbhaskar@stanford.edu

ABSTRACT

In this paper, we consider the recovery of a low rank matrix M given a subset of noisy quantized (or ordinal) measurements. We consider a constrained maximum likelihood estimation of M, under a constraint on the entry-wise infinity-norm of M and an exact rank constraint. We provide an upper bound on the Frobenius norm of the matrix estimation error under this model. Past theoretical investigations have been restricted to binary quantizers, and based on convex relaxation of the rank. We propose a globally convergent optimization algorithm exploiting existing work on low rank matrix factorization, and validate the method on synthetic data, with improved performance over past methods.

Index Terms— Quantization, matrix completion, convex optimization, constrained maximum likelihood.

1. INTRODUCTION

The problem of recovering a low rank matrix from an incomplete or noisy sampling of its entries arises in a variety of applications, including collaborative filtering [1, 2], sensor network localization [3, 4, 5], manifold learning [6, 7] and rank aggregation [8]. In many applications, the observations are not only missing, but are also highly discretized, or quantized, e.g. binary-valued (1-bit) [9, 10, 11], or multiple-valued [12]. For example, in the Netflix problem where a subset of the users' ratings is observed, the ratings take integer values between 1 and 5. Although one can apply existing matrix completion techniques to discrete-valued observations by treating them as continuous-valued, performance improvement can be achieved by treating the values as discrete [10, 12].

Relation to Prior Work: Recent work in the quantized matrix completion literature has followed a probabilistic model for relating the matrix Y of quantized observations to the underlying true matrix M (further described in (1)-(2) in the next section), and has performed estimation of low rank M via solving a constrained maximum likelihood (ML) optimization problem. In [10, 11], 1-bit (binary) observations are considered, and performance analysis is given based on a convex program and assumptions on the sampling scheme. In these works, directly using the binary observations is shown to result in better performance than treating the observations as real-valued. In [12], this work is extended to multi-level quantization and it is similarly shown that performance is improved when the number of levels is less than ten. However, no performance analysis exists for the multi-level case.

Contributions: In this paper, we follow the 1-bit formulation of [10, 11] in seeking an ML estimate of M, but consider multi-level observations. Instead of using a convex relaxation for the rank on M as in previous work, we use an exact rank constraint. We provide an upperbound on the Frobenius norm of the matrix estimation error under this model. We follow the sampling model of [13], which relates Ω, the set of indices of the observed data, to the edges of a bipartite graph. This allows our results to be more general in that it includes the commonly used sampling schemes in previous work such as uniform sampling of [10, 12] and non-uniform sampling. Finally, we present an algorithm based on matrix factorization for solving our optimization problem, which is globally convergent, and evaluate it on synthetic data.

Notation: We use capital letters, such as M, to denote a matrix, and M_{ij} as its (i,j)th entry. We let $\|M\|_2$, $\|M\|_F$ and $\|M\|_\infty$ denote the operator, Frobenius 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 $\{1,\dots,n\}$, $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all ones, $\dot{f}(x) := (\mathrm{d}f(x)/\mathrm{d}x)$, and $\mathbf{1}_{[A]}$ denotes the indicator function, i.e. $\mathbf{1}_{[A]} = 1$ when A is true, and $\mathbf{1}_{[A]} = 0$ otherwise.

2. SYSTEM MODEL

Given $M \in \mathbb{R}^{d_1 \times d_2}$, a subset of indices $\Omega \subseteq [d_1] \times [d_2]$, a twice differentiable known function $f_\ell : \mathbb{R} \to [0,1]$, and $\ell \in [K]$, $(K \geq 2)$, we observe

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

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 [12]:

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

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 [12]

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

$\omega_0 < \omega_1 < \cdots < \omega_K$, and we will take $\omega_0 = -\infty$ and $\omega_K = \infty$. In this paper, the quantization bin boundaries are assumed to be known in contrast to [12], where they are estimated in the algorithm. It then follows that

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

For $K = 2$ and $\omega_1 = 0$, one obtains the model of [10, 11]. We observe a subset $\Omega \subseteq [d_1] \times [d_2]$ of the entries in Y, which are related to M via the probabilistic model given in (1)-(2).

We wish to estimate unknown M using a constrained ML approach. We use $X \in \mathbb{R}^{d_1 \times d_2}$ to denote the optimization variable. The negative log-likelihood function is given by

$$F_{\Omega,Y}(X) = -\sum_{(i,j)\in\Omega} \left[\sum_{\ell=1}^{K} \left(\mathbf{1}_{[Y_{ij}=\ell]} \log(f_\ell(X_{ij})) \right) \right] \tag{5}$$

which is a convex function when the function f_ℓ is log-concave. Two natural choices for the function f 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 standard normal distribution $\mathcal{N}(0, 1)$. One motivation for these choices is that logistic and normal CDFs and pdfs are log-concave (e.g. [14, pp. 104-5]). Then, by [14, p. 105-7], f_ℓ is log-concave for logistic and normal $\Phi(x)$. These models have also been used in [9, 10, 11].

We assume that M is a low-rank matrix with rank bounded by r. 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. We refer the reader to [10, 11, 15] for further details. Coupling these constraints, based on the assumptions on M, with the negative log-likelihood function given in (5), the constrained ML estimate we wish to obtain is given by the solution to the optimization problem (s.t.: subject to):

$$\widehat{M} = \arg\min_X F_{\Omega,Y}(X) \text{ s.t. } \|X\|_\infty \leq \alpha,\ \text{rank}(X) \leq r. \tag{6}$$

We now discuss our assumptions on the set Ω, on which we follow [13]. Consider a bipartite graph $G = ([d_1], [d_2], E)$, where the edge set $E \subseteq [d_1] \times [d_2]$ 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 (BAM) where BAM $G_{ij} = 1$ if $(i, j) \in E$, $G_{ij} = 0$ if $(i, j) \notin E$. In [13], square matrices M are considered, and Ω is related to the edges of a bipartite graph with an equal number of left and right nodes. In this paper, M is rectangular. Thus, we will consider a larger bipartite graph $\tilde{G}$ with equal number of left and right nodes, with the original bipartite graph G being a "truncation" of $\tilde{G}$, and base our assumptions on this larger graph $\tilde{G}$, as follows. Set $d_{\max} = \max(d_1, d_2)$ and without loss of generality, take $d_1 \geq d_2$. Define the bipartite graph $\tilde{G} = ([d_{\max}], [d_{\max}], \tilde{E})$ such that the first d_2 columns of BAM $\tilde{G}$ equal BAM G and consider $\tilde{\Omega} \subseteq [d_{\max}] \times [d_{\max}]$ with $\tilde{\Omega}$ related to Ω just as $\tilde{G}$ is related to G. Then $(i, j) \in \tilde{E}$ iff $(i, j) \in \tilde{\Omega}$ and $\Omega \subseteq \tilde{\Omega}$. We denote the association of $\tilde{G}$ to $\tilde{\Omega}$ by $\tilde{G}\backslash\tilde{\Omega}$, and similarly for $G\backslash\Omega$. Following [13], we assume that the bipartite graph $\tilde{G}\backslash\tilde{\Omega}$ is d_g-regular with the following properties on its SVD:

(A1) All top singular vectors of $\tilde{G}$ are $\mathbf{1}_{d_{\max}}/\sqrt{d_{max}}$.

(A2) We have $\sigma_1(\tilde{G}) = d_g$ and $\sigma_2(\tilde{G}) \leq C_3\sqrt{d_g}$, where $C_3 > 0$ is some universal constant.

Thus we require $\tilde{G}$ to have a large enough spectral gap. Note that these assumptions imply that $|\Omega| \equiv m = d_g \max\{d_1, d_2\}$.

A family of graphs that satisfy (A1) and (A2) are Ramanujan graphs, a class of regular expander graphs [16]. As shown in [17] and discussed in [13], a Erdos-Renyi random graph with average degree $\geq c \log(d_{max})$ satisfies this spectral gap property with high probability (w.h.p.), and as shown in [18] and discussed in [13], so do stochastic block models for certain choices of inter- and intra-cluster edge connection probabilities. Thus, the sampling scheme of [13] is more general than a uniform sampling assumption, used in [10, 9], and it also includes the stochastic block model [13] resulting in non-uniform sampling.

3. PERFORMANCE BOUND

We now present a performance bound for $\widehat{M}$ in (6). Define

$$\gamma_\alpha \leq \min_{\ell\in[K]} \inf_{|x|\leq\alpha} \left\{ \frac{\dot{f}_\ell^2(x)}{f_\ell^2(x)} - \frac{\ddot{f}_\ell(x)}{f_\ell(x)} \right\}, \tag{7}$$

$$L_\alpha \geq \max_{\ell\in[K]} \sup_{|x|\leq\alpha} \left\{ \left| \dot{f}_\ell(x) \right| / f_\ell(x) \right\}, \tag{8}$$

where α is the bound on the entry-wise infinity-norm of $\widehat{M}$ (see (6)). For further reference, define the constraint set

$$\mathcal{C} := \left\{ X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_\infty \leq \alpha,\ \text{rank}(X) \leq r \right\}. \tag{9}$$

Theorem 3.1. *Assume that $M \in \mathcal{C}$, $|\Omega| = m$ and $\tilde{G}\backslash\tilde{\Omega}$ associated with $G\backslash\Omega$ satisfies assumptions (A1) and (A2). Further, suppose Y is generated as in (1) where $f_\ell(x)$ is log-concave in x $\forall \ell \in [K]$. Then with probability at least $1 - C_1 \exp(-C_2 m)$, any global minimizer $\widehat{M}$ of (6) satisfies*

$$\frac{\|\widehat{M} - M\|_F}{\sqrt{d_1 d_2}} \leq \max\left(C_{1\alpha}\sqrt{\frac{r^2 d_{\max}^3}{m d_1 d_2}}, C_{2\alpha}\sqrt{\frac{r^3 d_{\max}^9}{m^4 d_1 d_2}} \right) \tag{10}$$

if $\gamma_\alpha > 0$, where $C_{1\alpha} = 8\alpha C_3$, $C_{2\alpha} = 32.16\sqrt{2}L_\alpha/\gamma_\alpha$, $d_{\max} = \max(d_1, d_2)$. Here, $C_1, C_2, C_3 > 0$ are universal constants, and γ_α and L_α are given by (7), (8).

The proof is omitted for lack of space. It is based on a second-order Taylor expansion of the negative log-likelihood function around the true matrix, exploitation of [19, Theorem 8.4] concerning spectral norms of random matrices, and a modification of [13, Theorem 4.1] regarding closeness of a fixed matrix to its sampled version. By [14, p. 105-107] $f_\ell(x)$ is log-concave iff $\ddot{f}_\ell(x) f_\ell(x) \leq (\dot{f}_\ell(x))^2$. Thus $\gamma_\alpha \geq 0$, in general, and it can be shown to be > 0 for the logistic and Gaussian models of Sec. 2.

Of particular interest is the case where $p = \frac{m}{d_1 d_2}$ is fixed and we let d_1 and d_2 become large, with d_1/d_2 fixed. In this case with high probability (w.h.p.), we have

$$\frac{\|\widehat{M} - M\|_F}{\sqrt{d_1 d_2}} \leq \mathcal{O}\left(\frac{1}{p^2}\sqrt{\frac{r^3 d_{\max}^9}{(d_1 d_2)^5}}\right). \tag{11}$$

Comparison with previous work for the binary case ($K = 2$): We now provide a comparison of our results with those of [10, 11] which are restricted to the binary case (K=2). Consider $M \in \mathbb{R}^{n\times n}$, with p fraction of its entries sampled. Then, $d_1 = d_2 = n$, $m = p d_1 d_2 = p n^2$. The bounds proposed in [10, 11] yield w.h.p. $\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq C_\alpha \sqrt{\frac{r}{pn}} = \mathcal{O}\left(\sqrt{\frac{r}{pn}}\right)$ where C_α depends upon α and $f(x)$. In our bound we have w.h.p. $\frac{1}{n^2}\|\widehat{M} - M\|_F^2 \leq \mathcal{O}\left(\frac{r^3}{p^4 n}\right)$. Comparing the two, we see our method has faster convergence rate in n for fixed rank r and fraction of revealed entries p. One may notice that if the missing entries scale with n according to $p \sim \mathcal{O}(1/n)$ then [10] yields bounded error while our bound grows with n; in our case we need $p \geq \mathcal{O}(1/n^{1/4})$. We believe this to be an artifact of our proof, as our numerical results (Fig. 1a) show our method significantly outperforms [10] for low values of p and higher values of rank r, and is comparable for higher values of p and lower values of r.

4. OPTIMIZATION

We use the factorization technique of [20, 21, 22] where instead of optimizing with respect to X, it is factorized into two matrices $U \in \mathbb{R}^{d_1\times k}$ and $V \in \mathbb{R}^{d_2\times k}$ such that $X = UV^\top$. One then chooses $k = r + 1$, and optimizes with respect to the factors U, V. Following [11] for the case of $K = 2$, we have the following approximate projected gradient method.

Algorithm 1: Approximate Projected Gradient Method

Given estimates U^t, V^t at iteration t, one updates

$$\begin{bmatrix} U^{t+1} \\ V^{t+1} \end{bmatrix} = \mathcal{P}_\alpha\left(\begin{bmatrix} U^t - \frac{\tau}{\sqrt{t}}\nabla_X F_{\Omega,Y}(U^t V^{t\top})V^t \\ V^t - \frac{\tau}{\sqrt{t}}\nabla_X F_{\Omega,Y}(U^t V^{t\top})^\top U^t \end{bmatrix}\right) \tag{12}$$

where $\mathcal{P}_\alpha\left([U^\top\ V^\top]^\top\right) = \sqrt{\alpha/\|UV^\top\|_\infty}[U^\top\ \ V^\top]^\top$ if $\|UV^\top\|_\infty > \alpha$, and $= [U^\top\ V^\top]^\top$, otherwise. In (12) 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 (12) and approximate projection $\mathcal{P}_\alpha$, [11] ($K = 2$) also uses another projection to enforce a max-norm constraint. If a matrix A has rank r and $\|A\|_\infty \leq \alpha$, then $\|A\|_{\max} \leq \sqrt{r}\alpha$ [11]. Therefore, in our case the max-norm constraint is unnecessary; in this sense, our Algorithm 1 is the same as the approach of [11] for $K = 2$.

Remark 4.1. *The hard rank constraint results in a nonconvex constraint set; thus* (6) *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 [20, Proposition 4], following [21] 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* (6)*, so long as U^* and V^* are rank-deficient. (Rank deficiency of (U^*, V^*) is a sufficient condition, not necessary.) This result is utilized in [23], [22] and [11] for problems of this form. Thus one would expect to achieve global convergence for the problem of* (6)*. However, the "projection" $\mathcal{P}_\alpha$ in (12) is not an orthogonal projection and the set $\{\|UV^\top\|_\infty \leq \alpha\}$ is not convex in U, V (although $\{\|X\|_\infty \leq \alpha\}$ 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 [11]).*

The convergence deficiency discussed in Remark 4.1 motivates the following log-barrier penalty function approach [14, Sec. 11.2].

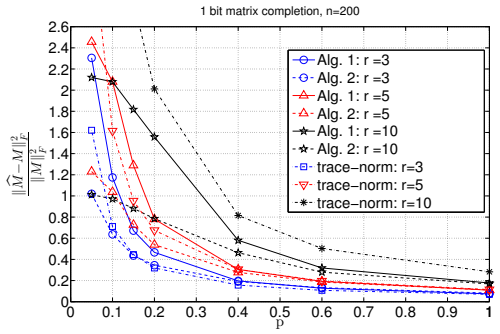
Algorithm 2: Logarithmic Barrier Gradient Method The constraint $\max_{i,j}|X_{ij}| \leq \alpha$ translates to $X_{ij} - \alpha \leq 0$ and $-X_{ij} - \alpha \leq 0\ \forall (i,j)$, which motivates the log-barrier penalty function $-\log\left(1 - (X_{ij}/\alpha)^2\right)$ which is finite for $|X_{ij}|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

$$\tilde{F}_{\Omega,Y}(X) = F_{\Omega,Y}(X) - \lambda \sum_{(i,j)} \log\left(1 - (X_{ij}/\alpha)^2\right) \tag{13}$$

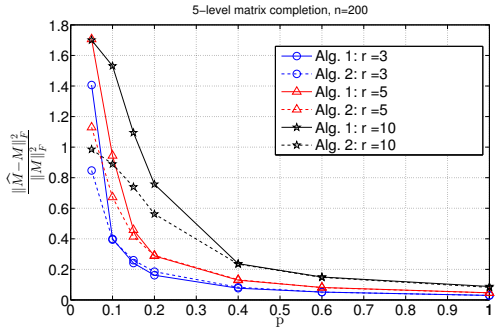
and the optimization problem

$$\widehat{M} = \arg\min_X \tilde{F}_{\Omega,Y}(X) \text{ s.t. } \operatorname{rank}(X) \leq r. \tag{14}$$

The parameter $\lambda > 0$ in (13) 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, convergence is guaranteed, and the factorization approach $X = UV^\top$ is well-justified, per Remark 4.1. The above problem is typically solved via a sequence of central path following solutions [14, Sec. 11.2] where one gradually reduces λ toward 0. In our approach we initialize it with the solution to Algorithm 1 and then either use a single "small" value of λ, or select λ via 5-fold cross-validation. One may therefore view augmentation with log-barrier cost as regularization of $F_{\Omega,Y}(X)$. We solve the factored version $X = UV^\top$ of problem (14) for a fixed λ using a Nesterov/accelerated gradient method with adaptive restart [24], or a gradient descent method.



(a) K=2: binary case, $w_1 = 0$, "trace-norm" refers to [10], Alg. 1 is the method of [11] for known r.



(b) K=5: $w_1 = -0.4$, $w_2 = -0.15$, $w_3 = 0.15$, $w_4 = 0.4$

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

5. NUMERICAL EXPERIMENTS

In this section, we test our method on synthetic data. We set $d_1 = d_2 = n$ 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 i.i.d. entries drawn from a uniform distribution on $[-0.5, 0.5]$ (as in [11, 10] for $K = 2$). Then we scale M to achieve $\|M\|_\infty = 1 = \alpha$. We pick $r = 3$, 5 or 10, vary matrix sizes $n = 100, 200,$ or 400. We generate the set Ω of revealed indices via a stochastic block model as in [13]. 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 when the spectral gap is the largest. This setting also corresponds to the Bernoulli sampling model of [10]. 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. We used the model (2) with Z_{ij} as zero-mean Gaussian with standard deviation $\sigma = 0.18$ (as in [11, 10] for $K = 2$). 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. The resulting relative mean-square error (MSE) $\|\widehat{M} - M\|_F^2/\|M\|_F^2$ is shown in Figures 1a and 1b for $n = 200$, and $K = 2$, which we compare with [10], and $K = 5$, respectively, where we average over 20 Monte Carlo runs. As expected, the performance improves with increasing n and increasing p. For $K = 2$, Alg. 2 significantly outperforms [10, 11] for low values of p and high values of r.

In Fig. 2 we show the relative MSE for $n = 100, 200, 400$, $p = 0.2, 0.4, 0.6$ $(= q)$. We also plot the line $1/n$ in Fig. 2 to show the scale of the upper bound $\mathcal{O}\left(r^3/(p^4 n)\right)$ established in Section 3. As we can see, the empirical estimation errors follow approximately the same scaling, suggesting that our analysis is tight, up to some constant. In Fig. 3 we plot the relative MSE for $n = 200$, 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 [13] and MSE is the smallest, and as p gets larger, the spectral gap decreases, leading to larger MSE.

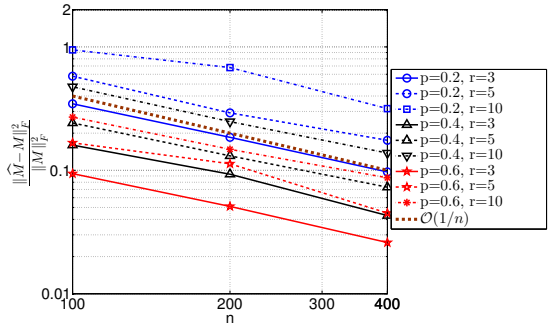


Fig. 2: Relative MSE: $K = 5$, $p = q$, $r = 3$, 5 or 10, $n = 100$, 200, 400, $\alpha = 1$, Gaussian noise $\sigma = 0.18$

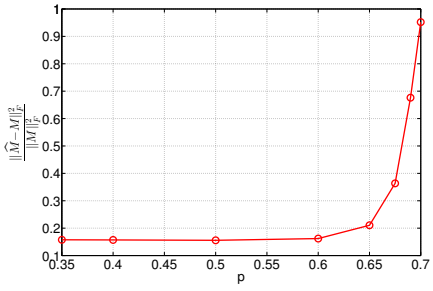


Fig. 3: Relative MSE versus p, fixed $p + q = 0.7$, $K = 5$

6. CONCLUSIONS

We investigated constrained ML estimation of an unknown low-rank matrix M given noisy quantized (or ordinal) measurements, under a constraint on the entry-wise infinity-norm of M and an exact rank constraint (upper bound). We provided an upper bound on the Frobenius norm of matrix estimation error, proposed a globally convergent optimization algorithm, and validated the method on synthetic data.

7. REFERENCES

[1] 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*, ser. ICML '05, 2005, pp. 713–719.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[3] Y. Shang, W. Ruml, Y. Zhang, and M. Fromherz, "Localization from connectivity in sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 11, pp. 961–974, 2004.

[4] A. Javanmard and A. Montanari, "Localization from incomplete noisy distance measurements," *Found. Comput. Math.*, vol. 13, no. 3, pp. 297–345, 2013.

[5] A. Karbasi and S. Oh, "Robust localization from incomplete local information," *IEEE Transactions on Networking*, vol. 21, pp. 1131 – 1144, August 2013.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2003.

[8] D. F. Gleich and L.-H. Lim, "Rank aggregation via nuclear norm minimization," in *Proc. SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '11. ACM, 2011, pp. 60–68.

[9] 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*, vol. 15, pp. 1959–2008, 2014.

[10] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, pp. 189–223, 2014.

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

[12] A. S. Lan, C. Studer, and R. G. Baraniuk, "Matrix recovery from quantized and corrupted measurements," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[13] S. Bhojanapalli and P. Jain, "Universal matrix completion," in *Proceedings of the 31st International Conference on Machine Learning*, ser. ICML '14, 2014.

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

[15] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[16] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications," *Bulletin of American Mathematical Society*, vol. 43, no. 4, pp. 439–561, 2006.

[17] U. Feige and E. Ofek, "Spectral techniques applied to sparse random graphs," *Random Structures & Algorithms*, vol. 27, pp. 251–275, 2005, DOI: 10.1002/rsa.20089.

[18] R. R. Nadakuditi and M. E. Newman, "Graph spectra and the detectability of community structure in networks," *Physical Review Letters*, vol. 108, no. 18, pp. 188 701–5, 2012.

[19] S. Chatterjee, "Matrix estimation by universal singular value thresholding," 2013, arXiv:1212.1247v5 .

[20] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," 2008, arXiv:0812.1869v1 .

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

[22] J. D. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. A. Tropp, "Practical large-scale optimization for max-norm regularization," in *NIPS*, 2010, pp. 1297–1305.

[23] B. Recht and C. Re, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Program. Comput.*, vol. 5, no. 2, pp. 201–226, 2013.

[24] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Foundations of Computational Mathematics*, 2013.