# Noisy Inductive Matrix Completion Under Sparse Factor Models

Akshay Soni, Troy Chevalier
Yahoo! Research
Sunnyvale, CA, USA
Email: {akshaysoni,troyc}@yahoo-inc.com

Swayambhoo Jain
Department of Electrical and Computer Engineering
University of Minnesota – Twin Cities
Email: jainx174@umn.edu

*Abstract*—Inductive Matrix Completion (IMC) is an important class of matrix completion problems that allows direct inclusion of available features to enhance estimation capabilities. These models have found applications in personalized recommendation systems, multilabel learning, dictionary learning, etc. This paper examines a general class of noisy matrix completion tasks where the underlying matrix is following an IMC model i.e., it is formed by a mixing matrix (a priori unknown) sandwiched between two known feature matrices. The mixing matrix here is assumed to be well approximated by the product of two sparse matrices—referred here to as "sparse factor models." We leverage the main theorem of [1] and extend it to provide theoretical error bounds for the sparsity-regularized maximum likelihood estimators for the class of problems discussed in this paper. The main result is general in the sense that it can be used to derive error bounds for various noise models. In this paper, we instantiate our main result for the case of Gaussian noise and provide corresponding error bounds in terms of squared loss.

## I. Introduction

In recent years, matrix completion has been a venue for constant research—both from theoretical as well as algorithmic front. The applications include collaborative filtering, multiclass learning, dimensionality reduction etc. In most general terms, the aim of these problems is to estimate the elements of a matrix $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ from noisy observations collected on only a subset of its locations. Of course, such estimation problems are ill-posed without further assumptions, since the value of $\mathbf{X}^*$ at the unobserved locations can be arbitrary. A common approach is to augment the inference method with the assumption that the underlying matrix to be estimated exhibits some form of intrinsic low-dimensional structure, making, in some cases, the problem feasible. For example, the assumption of low-rank on the target matrix has been the main focus of research for these problems. Even though the rank constrained problem is non-convex and difficult to analyze, convex relaxations to these problems have been well studied in noiseless [2], additive uncertainty [3], and even in settings where observations are nonlinear (e.g., one-bit quantized [4]). However, due to the computational complexity reasons, the most popular algorithms to solve these problems are variants of alternating minimization [1]; the low-rank structure is imposed explicitly via bilinear models i.e., $\mathbf{X}^*$ is assumed to be a product of two factor-matrices as

$$\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*, \tag{1}$$

where $\mathbf{D}^* \in \mathbb{R}^{n_1 \times r}$, $\mathbf{A}^* \in \mathbb{R}^{r \times n_2}$ and $r \ll \min(n_1, n_2)$. Slight modifications of this formulation also allows for explicit control and facilitates imposing additional constraints like positivity [5], sparsity [6], or even tree-sparsity [7] on the factors.

In terms of the well-known user-item recommendation problem [8]: $\mathbf{X}^*$ pertains to the rating matrix, $\mathbf{D}^*$ and $\mathbf{A}^*$ denote the matrices of user and item latent-factors, respectively. Apart from the rating information, generally, we also have other user and item side-features (like age, gender, etc) that can be used to further improve the quality of estimation. These features can be included via a Bayesian framework [9], [10] or by directly changing the model as in *Inductive*

*Matrix Completion* (IMC) [11], [12] where we have the following model assumption on $\mathbf{X}^*$:

$$\mathbf{X}^* = \mathbf{A}\mathbf{W}^*\mathbf{B} = \mathbf{A}\mathbf{P}^*\mathbf{Q}^*\mathbf{B}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{B} \in \mathbb{R}^{r_2 \times n_2}$ are known matrices, for example, user and item feature matrices, $\mathbf{W}^* \in \mathbb{R}^{r_1 \times r_2}$ is a mixing matrix whose low-rank structure is explicitly imposed via a bilinear model $\mathbf{W}^* = \mathbf{P}^*\mathbf{Q}^*$, $\mathbf{P}^* \in \mathbb{R}^{r_1 \times r}$ and $\mathbf{Q}^* \in \mathbb{R}^{r \times r_2}$.

The goal of this paper is to provide estimation error bounds for matrix completion under inductive setting, as in (2), by using the framework of regularized maximum likelihood estimation [1]. For generality, we derive these bounds under the assumption that both $\mathbf{P}^*$ and $\mathbf{Q}^*$ can be sparse, and noise can follow any generic distribution. This kind of sparse-structure arises in recommendation problems where the desire is to have sparse latent vectors due to requirement of low computational and storage complexity [13], and variant of dictionary learning problems where apart from the coefficient matrix being sparse, dictionary itself is modeled to be sparse [14].

Before delving into the specific contributions of this paper, we first put our work in the context of exiting literature in the next subsection.

### A. Related Works

Our inference tasks here essentially entails learning two structured factors in a bilinear model. With a few notable exceptions (e.g., low-rank matrices and certain non-negative matrices [15], [16]), the non-convex bilinear models studied in this paper are difficult to analyze. Several recent efforts in the dictionary learning literature have established theoretical guarantees on identifiability, as well as algorithmic analyses of a number of optimization approaches [17]–[20]. The most closely related to our work are [1], [6], [17], where authors provide a framework for deriving estimation error bounds for the models of form (1) under various noise distributions and sparsity structure on $\mathbf{A}^*$ by using a regularized maximum-likelihood approach based on the ideas from [21]. The Overall theme of the estimation bounds obtained using this framework is that the error of their estimate, $\hat{\mathbf{X}}$ of $\mathbf{X}^*$, decays as the ratio of number of degrees of freedom to estimate over the number of observations.

The idea of including side features in the factorization models to improve recommendations is considered using Bayesian formulations in [22] or via directly incorporating these features in a generative model in IMC [12], Factorization Machines [23]. The statistical performance of IMC has been theoretically analyzed via convex relaxations based on nuclear-norm in noiseless [12] and noisy settings [24]. From the algorithmic perspective, provable alternating-minimization based approach with bilinear models has been recently proposed and analyzed [11]. Our results in this paper provide additional insights into these problems under a general setting of noise and sparsity of the bilinear factors. It is to be noted here that the aim of all these related works is to accurately estimate the mixing matrix $\mathbf{W}^*$; we provide error bounds for the estimation of $\mathbf{X}^*$.

Many recent works have used IMC framework for large scale recommendations [25] or extended IMC to automatically construct better features [26].

### B. Specific Contributions

In this paper we provide estimation error guarantees for matrix completion using regularized maximum likelihood estimator under the model described in (2) for the cases where the factor matrices are dense (either one or both), or when $\mathbf{W}^*$ is itself sparse. In order to arrive at these bounds, we extend the fundamental theorem proposed in [1] to the inductive matrix completion setting studied here. This extension is detailed in Theorem II.1. The main utility of this theorem is that it can provide estimation error guarantees for a variety of noise distributions; for example, Gaussian, Poisson, Laplace, and even under the extreme quantized setting of 1-bit observations. Here we instantiate it for the Gaussian noise and obtain corresponding error guarantees.

### C. Preliminaries

To set the stage for the statement of our main result, we recall several information-theoretic preliminaries. When $p(Y)$ and $q(Y)$ denote the pdf (or pmf) of a real-valued random variable $Y$, the Kullback-Leibler divergence (or KL divergence) of $q$ from $p$ is denoted $\mathrm{D}(p\|q)$ and given by

$$\mathrm{D}(p\|q) = \mathbb{E}_p\left[\log\frac{p(Y)}{q(Y)}\right]$$

where the logarithm is taken to be the natural log. By definition, $\mathrm{D}(p\|q)$ is finite only if the support of $p$ is contained in the support of $q$. Further, the KL divergence satisfies $\mathrm{D}(p\|q) \geq 0$ and $\mathrm{D}(p\|q) = 0$ when $p(Y) = q(Y)$. We also use the Hellinger affinity denoted by $\mathrm{A}(p,q)$ and given by

$$\mathrm{A}(p,q) = \mathbb{E}_p\left[\sqrt{\frac{q(Y)}{p(Y)}}\right] = \mathbb{E}_q\left[\sqrt{\frac{p(Y)}{q(Y)}}\right]$$

It is known that $0 \leq \mathrm{A}(p,q) \leq 1$. When $p$ and $q$ are parameterized by elements $X_{i,j}$ and $\widetilde{X}_{i,j}$ of matrices $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, respectively, so that $p(Y_{i,j}) = p_{X_{i,j}}(Y_{i,j})$ and $q(Y_{i,j}) = q_{\widetilde{X}_{i,j}}(Y_{i,j})$, we use the shorthand notation $\mathrm{D}(p_{\mathbf{X}}\|q_{\widetilde{\mathbf{X}}}) \triangleq \sum_{i,j} \mathrm{D}(p_{X_{i,j}}\|q_{\widetilde{X}_{i,j}})$ and $\mathrm{A}(p_{\mathbf{X}}, q_{\widetilde{\mathbf{X}}}) \triangleq \prod_{i,j} \mathrm{A}(p_{X_{i,j}}, q_{\widetilde{X}_{i,j}})$. Finally, for a matrix $\mathbf{M}$ we denote by $\|\mathbf{M}\|_0$ its number of nonzero elements, and $\|\mathbf{M}\|_{\max}$ the magnitude of its largest element (in absolute value). Also, $(a_1 \vee a_2) = \max(a_1, a_2)$.

## II. PROBLEM STATEMENT AND RECOVERY RESULT

Our focus here is to estimate a matrix $\mathbf{X}^*$ that admits a model as in (2) from noisy observations $\mathbf{Y}_{\mathcal{S}} = \{Y_{i,j}\}_{(i,j)\in\mathcal{S}}$ collected at a subset $\mathcal{S} \subset [n_1] \times [n_2]$ of its locations. Sampling locations are randomly chosen in the sense that for an integer $m$ satisfying $4 \leq m \leq n_1 n_2$ and $\gamma = m(n_1 n_2)^{-1}$, each location is observed independently via a Bernoulli($\gamma$) model. Going forward, we assume the following bounds on the participating matrices:

$$\|\mathbf{X}^*\|_{\max} \leq \mathrm{X}_{\max}/2, \ \|\mathbf{P}^*\|_{\max} \leq 1, \ \|\mathbf{Q}^*\|_{\max} \leq \mathrm{Q}_{\max},$$
$$\|\mathbf{A}\|_{\max} \leq \mathrm{A}_{\max}, \text{ and } \|\mathbf{B}\|_{\max} \leq \mathrm{B}_{\max},$$

for some positive constants $\mathrm{Q}_{\max}$, $\mathrm{A}_{\max}$, and $\mathrm{B}_{\max}$.

Given $\mathcal{S}$ and $\mathbf{Y}_{\mathcal{S}}$, we write the joint pdf (or pmf) of the observations as

$$p_{\mathbf{X}_{\mathcal{S}}^*}(\mathbf{Y}_{\mathcal{S}}) \triangleq \prod_{(i,j)\in\mathcal{S}} p_{X_{i,j}^*}(Y_{i,j}), \tag{3}$$

where $p_{X_{i,j}^*}(Y_{i,j})$ denotes the corresponding scalar pdf (or pmf), and we use the shorthand $\mathbf{X}_{\mathcal{S}}^*$ to denote the collection of elements of $\mathbf{X}^*$ indexed by $(i,j) \in \mathcal{S}$. Then our task may be described succinctly as follows: given $\mathcal{S}$ and corresponding noisy observations $\mathbf{Y}_{\mathcal{S}}$ of $\mathbf{X}^*$ distributed according to (3), our goal is to estimate $\mathbf{X}^*$ under the assumption that it admits a model as in (2).

Our approach will be to estimate $\mathbf{X}^*$ via sparsity-penalized maximum likelihood; we consider estimates of the form

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}=\mathbf{APQB}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) + \lambda_{\mathrm{P}}\|\mathbf{P}\|_0 + \lambda_{\mathrm{Q}}\|\mathbf{Q}\|_0\right\}, \tag{4}$$

where $\{\lambda_{\mathrm{P}}, \lambda_{\mathrm{Q}}\} > 0$ are user-specified regularization parameter, $\mathbf{X}_{\mathcal{S}}$ is shorthand for the collection $\{X_{i,j}\}_{(i,j)\in\mathcal{S}}$ of entries of $\mathbf{X}$ indexed by $\mathcal{S}$, and $\mathcal{X}$ is an appropriately constructed class of candidate estimates. To facilitate our analysis here, we take $\mathcal{X}$ to be a countable class of estimates constructed as follows: for a specified $\beta \geq 1$, we set $L_{\mathrm{lev}} = 2^{\lceil \log_2(n_1 \vee n_2)^\beta \rceil}$ and construct $\mathcal{P}$ to be the set of all matrices $\mathbf{P} \in \mathbb{R}^{r_1 \times r}$ whose elements are either 0, or are discretized to one of $L_{\mathrm{lev}}$ uniformly spaced levels in the range $[-1, 1]$ and $\mathcal{Q}$ to be the set of all matrices $\mathbf{Q} \in \mathbb{R}^{r \times r_2}$ whose elements either take the value zero, or are discretized to one of $L_{\mathrm{lev}}$ uniformly-spaced levels in the range $[-\mathrm{Q}_{\max}, \mathrm{Q}_{\max}]$. Then, we let

$$\mathcal{X} \triangleq \{\mathbf{X} = \mathbf{APQB} : \mathbf{P} \in \mathcal{P}, \ \mathbf{Q} \in \mathcal{Q}, \ \|\mathbf{X}\|_{\max} \leq \mathrm{X}_{\max}\}. \tag{5}$$

Our main result establishes error bounds under general noise or corruption models. We state the result here as a theorem; its proof appears in Section IV.

**Theorem II.1.** *Let $\mathcal{S}$, $\gamma$, $m$, $\beta$, $\mathcal{X}$, and $\mathbf{Y}_{\mathcal{S}}$ be as defined above and $C_{\mathrm{D}}$ is any constant satisfying*

$$C_{\mathrm{D}} \geq \max_{\mathbf{X}\in\mathcal{X}} \max_{i,j} \ D(p_{X_{i,j}^*}\|p_{X_{i,j}}).$$

*Then for any*

$$\lambda_{\mathrm{P}} \geq 2\left(1 + \frac{2C_{\mathrm{D}}}{3}\right)(4\log(r_1) + \beta\log(n_1 \vee n_2)) \tag{6}$$

$$\lambda_{\mathrm{Q}} \geq 2\left(1 + \frac{2C_{\mathrm{D}}}{3}\right)(4\log(r_2) + \beta\log(n_1 \vee n_2)) \tag{7}$$

*the complexity penalized maximum likelihood estimator* (4) *satisfies the (normalized, per-element) error bound*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log\mathrm{A}(p_{\widehat{\mathbf{X}}\lambda}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_{\mathrm{D}}\log m}{m} +$$
$$3 \cdot \min_{\mathbf{X}\in\mathcal{X}}\left\{\frac{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left((\lambda_{\mathrm{P}} + \lambda_{\mathrm{Q}}) + \frac{4C_{\mathrm{D}}(\beta+4)\log(n_1 \vee n_2)}{3}\right)\left(\frac{\|\mathbf{P}\|_0 + \|\mathbf{Q}\|_0}{m}\right)\right\}.$$

Before proceeding to specific case of Gaussian noise, we note a few salient points about this result. First, as alluded above, our results is general and can be used to analyze the error performance under a variety of noise models. Specialization to a given noise model requires us to compute the upper bounds for KL divergences in terms of problem parameters, and the lower bound of negative log Hellinger affinities in terms of the required error function. Second, the above result is kind of an *oracle* bound, in that it is specified in terms of a minimum over $\mathbf{X} \in \mathcal{X}$. We can construct a valid upper bound by evaluating this for any $\mathbf{X} \in \mathcal{X}$. We construct a specific $\mathbf{X} \in \mathcal{X}$ that is close to $\mathbf{X}^*$, and evaluate the oracle bound for the constructed point to get a non-trivial upper bound.

Finally, we note that the optimization problem (4) is non-convex, as often is the case, not only because of $\ell_0$ penalty, or discretized

space ($\mathcal{X}$) used for proofs in this paper, but due the fundamental challenge posed by bilinear models.

### A. Implications for Gaussian Noise

Below we provide scaling behavior of error bound with problem parameters for Gaussian noise with known variance by instantiating Theorem II.1. The joint likelihood of observations can be written as

$$p_{\mathbf{X}_S^*}(\mathbf{Y}_S) = \frac{1}{(2\pi\sigma^2)^{|S|/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y}_S - \mathbf{X}_S^*\|_F^2\right), \quad (8)$$

where we have used the shorthand notation $\|\mathbf{Y}_S - \mathbf{X}_S^*\|_F^2 \triangleq \sum_{(i,j)\in S}(Y_{i,j} - X_{i,j}^*)^2$. To that end, we fix $\beta$ as

$$\beta = \max\left\{1, 1 + \frac{\log(6\sqrt{m}\cdot(r\cdot r_1\cdot r_2)\cdot \mathrm{A_{max}B_{max}Q_{max}/X_{max}})}{\log(n_1\vee n_2)}\right\} \quad (9)$$

for describing the number of discretization levels in the elements of each factor matrix and regularization parameters $(\lambda_P, \lambda_Q)$ equal to right hand side of (6). In this setting we have the following result; its proof appears in Section IV-B.

**Corollary II.1.** *Let $\beta$, $\lambda_P$, and $\lambda_Q$ be as defined above with $C_D = 2\mathrm{X_{max}^2}/\sigma^2$. The estimate $\widehat{\mathbf{X}}$ obtained via (4) satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_S}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} = \quad (10)$$
$$\mathcal{O}\left((\sigma^2 + \mathrm{X_{max}^2})\left(\frac{\|\mathbf{P}^*\|_0 + \|\mathbf{Q}^*\|_0}{m}\right)\log(n_1\vee n_2)\right).$$

A few comments are in order regarding these error guarantees. We may interpret the quantity $\|\mathbf{P}^*\|_0 + \|\mathbf{Q}^*\|_0$ as the number of degrees of freedom to be estimated, and in this sense we see that the error rate of the penalized maximum likelihood estimator exhibits characteristics of the well-known parametric rate (modulo the logarithmic factor). If $\mathbf{P}^*$ and $\mathbf{Q}^*$ are instead dense, then we get the familiar error bound that decays as $(r_1 + r_2)r/m$ with $m$. Compared this with the case where matrix $\mathbf{X}^*$ is just low-rank (as in formal matrix completion tasks), the error rates are of the form $(n_1 + n_2)r/m$ [1]. Clearly, the IMC model has a better error performance if $r_2 \ll n_2$ or even better if the factor matrices are sparse.

The presence of the factor $\mathrm{X_{max}^2}$ is akin to the incoherence conditions prevalent in the matrix-completion literature. In essence, it pertains to the spikiness of the underlying matrix we are trying to estimate [27].

### III. Conclusion and future directions

In this paper, we considered the matrix completion problem under the IMC model assumption. We provided a generic theorem which can be extended to many settings including different noise distributions and underlying matrix structural assumptions. We considered the specific case of Gaussian noise and demonstrated the estimation error guarantees. The extensions to other noise distributions and matrix structures are left for future research direction. Obtaining lower bounds for the estimation problem discussed here and understanding the min-max optimality of the bounds obtained in this paper is yet another interesting direction for future exploration.

### IV. Proofs

### A. Proof of Theorem II.1

We utilize the following Lemma from [1] and apply it to our problem.

**Lemma IV.1.** *Let $\mathbf{X}^*$ be an $n_1 \times n_2$ matrix whose elements we aim to estimate, and let $\mathcal{X}$ be a countable collection of candidate reconstructions $\mathbf{X}$ of $\mathbf{X}^*$, each with corresponding penalty $\mathrm{pen}(\mathbf{X}) \geq 1$, so that the collection of penalties satisfies the summability condition $\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$. Given $m$, $\gamma$, $\mathcal{S}$, $\mathbf{Y}_S$, and $C_D$ as explained earlier, we have that for any $\xi \geq \left(1 + \frac{2C_D}{3}\right)\cdot 2\log 2$, the complexity penalized maximum likelihood estimator*

$$\widehat{\mathbf{X}}^\xi = \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{-\log p_{\mathbf{X}_S}(\mathbf{Y}_S) + \xi\cdot\mathrm{pen}(\mathbf{X})\right\},$$

*satisfies the (normalized, per-element) error bound*

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_S}\left[-2\log\mathrm{A}(p_{\widehat{\mathbf{X}}\xi}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_D\log m}{m} +$$
$$3\cdot\min_{\mathbf{X}\in\mathcal{X}}\left\{\frac{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left(\xi + \frac{4C_D\log 2}{3}\right)\frac{\mathrm{pen}(\mathbf{X})}{m}\right\},$$

*where, as denoted, the expectation is with respect to the joint distribution of $\mathcal{S}$ and $\mathbf{Y}_S$.*

Now, consider any discretized matrix factors $\mathbf{P}\in\mathcal{P}$ and $\mathbf{Q}\in\mathcal{Q}$, as described in Section II. For $L_{\mathrm{loc}}^{\mathcal{P}} \triangleq 2^{\lceil\log_2(r_1\cdot r)\rceil}$, we encode each nonzero element of $\mathbf{P}$ using $\log_2 L_{\mathrm{loc}}^{\mathcal{P}}$ bits to denote its location and $\log_2 L_{\mathrm{lev}}$ bits for its amplitude. We can then encode $\mathbf{P}$ with $\|\mathbf{P}\|_0$ nonzero entries by using $\|\mathbf{P}\|_0(\log_2 L_{\mathrm{loc}}^{\mathcal{P}} + \log_2 L_{\mathrm{lev}})$ bits. Similarly, we can encode $\mathbf{Q}$ using $\|\mathbf{Q}\|_0(\log_2 L_{\mathrm{loc}}^{\mathcal{Q}} + \log_2 L_{\mathrm{lev}})$ where $L_{\mathrm{loc}}^{\mathcal{Q}} \triangleq 2^{\lceil\log_2(r_2\cdot r)\rceil}$. Now, we let $\mathcal{X}'$ be set of all such $\mathbf{X} = \mathbf{APQB}$, and let the binary code for each $\mathbf{X}$ be the concatenation of the binary code for $\mathbf{P}$ followed by the binary code for $\mathbf{Q}$. It follows that we may assign penalties $\mathrm{pen}(\mathbf{X})$ to all $\mathbf{X}\in\mathcal{X}'$ whose length satisfy

$$\mathrm{pen}(\mathbf{X}) = \|\mathbf{P}\|_0(\log_2(L_{\mathrm{loc}}^{\mathcal{P}}\cdot L_{\mathrm{lev}})) + \|\mathbf{Q}\|_0(\log_2(L_{\mathrm{loc}}^{\mathcal{Q}}\cdot L_{\mathrm{lev}})), \quad (11)$$

given that $\mathbf{A}$ and $\mathbf{B}$ are known matrices.

It is easy to see that such codes are (by construction) uniquely decodable, so we have that $\sum_{\mathbf{X}\in\mathcal{X}'} 2^{-\mathrm{pen}(\mathbf{X})} \leq 1$ by the well-known Kraft-McMillan Inequality [28]. Now let $\mathcal{X}$ be any subset of $\mathcal{X}'$. For randomly subsampled and noisy observations $\mathbf{Y}_S$ our estimates take the form

$$\widehat{\mathbf{X}}^\xi = \arg\min_{\mathbf{X}=\mathbf{APQB}\in\mathcal{X}}\left\{-\log p_{\mathbf{X}_S}(\mathbf{Y}_S) + \xi\cdot\mathrm{pen}(\mathbf{X})\right\}.$$

Further, using Lemma IV.1 we have

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_S}\left[-2\log\mathrm{A}(p_{\widehat{\mathbf{X}}\xi}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_D\log m}{m} +$$
$$3\cdot\min_{\mathbf{X}\in\mathcal{X}}\left\{\frac{\mathrm{D}(p_{\mathbf{X}^*}\|p_{\mathbf{X}})}{n_1 n_2} + \left(\xi + \frac{4C_D\log 2}{3}\right)(\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}})\left(\frac{\|\mathbf{P}\|_0 + \|\mathbf{Q}\|_0}{m}\right)\right\},$$

where $\log_2 L_{\mathrm{loc}} = (\log_2 L_{\mathrm{loc}}^{\mathcal{P}} \vee \log_2 L_{\mathrm{loc}}^{\mathcal{Q}})$. Finally, letting

$$\lambda_P = \xi\cdot(\log_2 L_{\mathrm{loc}}^{\mathcal{P}} + \log_2 L_{\mathrm{lev}})$$
$$\lambda_Q = \xi\cdot(\log_2 L_{\mathrm{loc}}^{\mathcal{Q}} + \log_2 L_{\mathrm{lev}})$$

and using the fact that

$$\log_2 L_{\mathrm{loc}}^{\mathcal{P}} + \log_2 L_{\mathrm{lev}} \leq 8\log(r_1) + 2\beta\log(n_1\vee n_2)$$
$$\log_2 L_{\mathrm{loc}}^{\mathcal{Q}} + \log_2 L_{\mathrm{lev}} \leq 8\log(r_2) + 2\beta\log(n_1\vee n_2)$$
$$\log_2 L_{\mathrm{loc}} + \log_2 L_{\mathrm{lev}} \leq (\beta + 4)\cdot\log(n_1\vee n_2)\cdot 2\log_2(e)$$

which follows by our selection of $L_{\mathrm{lev}}$ and $L_{\mathrm{loc}}$ and the fact that $r_1 < n_1$, $r_2 < n_2$; it follows (after some straightforward simplification)

that $(\lambda_P, \lambda_Q)$ follow (6), and the estimate (4) satisfies

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log A(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} \leq \frac{8C_D \log m}{m} +$$

$$3 \cdot \min_{\mathbf{X} \in \mathcal{X}} \left\{ \frac{D(p_{\mathbf{X}^*} \| p_{\mathbf{X}})}{n_1 n_2} + \right.$$

$$\left. \left( (\lambda_P + \lambda_Q) + \frac{4C_D(\beta+4)\log(n_1 \vee n_2)}{3} \right) \left( \frac{\|\mathbf{P}\|_0 + \|\mathbf{Q}\|_0}{m} \right) \right\},$$

as claimed.

### B. Corollary II.1

*Proof:* For $\mathbf{X}^*$ as specified and any $\mathbf{X} \in \mathcal{X}$, using the model (8) we have

$$D(p_{X_{i,j}^*} \| p_{X_{i,j}}) = \frac{(X_{i,j}^* - X_{i,j})^2}{2\sigma^2}$$

for any fixed $(i,j) \in S$. It follows that $D(p_{\mathbf{X}^*} \| p_{\mathbf{X}}) = \|\mathbf{X}^* - \mathbf{X}\|_F^2 / 2\sigma^2$, and using the fact that the amplitudes of entries of $\mathbf{X}^*$ and all $\mathbf{X} \in \mathcal{X}$ are no larger than $X_{\max}$, it is clear that we may choose $C_D = 2X_{\max}^2 / \sigma^2$. Further, for any $\mathbf{X} \in \mathcal{X}$ and any fixed $(i,j) \in \mathcal{S}$ it is easy to show that in this case

$$-2\log A(p_{X_{i,j}}, p_{X_{i,j}^*}) = \frac{(X_{i,j}^* - X_{i,j})^2}{4\sigma^2},$$

so that $-2\log A(p_{\mathbf{X}}, p_{\mathbf{X}^*}) = \|\mathbf{X}^* - \mathbf{X}\|_F^2 / 4\sigma^2$. It follows that

$$\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[-2\log A(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right] = \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{4\sigma^2}.$$

Incorporating this into Theorem II.1, we obtain that for regularization parameters as in (6) with $C_D$ substituted, the sparsity penalized maximum-likelihood estimate satisfies the per-element mean-square error bound

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \frac{64 X_{\max}^2 \log m}{m}$$

$$+ 6 \cdot \min_{\mathbf{X} \in \mathcal{X}} \left\{ \frac{\|\mathbf{X}^* - \mathbf{X}\|_F^2}{n_1 n_2} + \right.$$

$$\left( 2\sigma^2(\lambda_P + \lambda_Q) + \frac{16 X_{\max}^2(\beta+4)\log(n_1 \vee n_2)}{3} \right)$$

$$\left. \left( \frac{\|\mathbf{P}\|_0 + \|\mathbf{Q}\|_0}{m} \right) \right\}.$$

We now establish the error bound for the case where $\mathbf{P}^*$ and $\mathbf{Q}^*$ are sparse with corresponding cardinality $\|\mathbf{P}^*\|_0$ and $\|\mathbf{Q}^*\|_0$. Consider a candidate reconstruction of the form $\mathbf{X}_q^* = \mathbf{A}\mathbf{P}_q^*\mathbf{Q}_q^*\mathbf{B}$, where the elements of $\mathbf{P}_q^*$ are the closest discretized surrogates of the *nonzero* entries of $\mathbf{P}^*$, and the entries of and $\mathbf{Q}_q^*$ are the closest discretized surrogates of the *nonzero* entries of $\mathbf{Q}^*$. Denote $\mathbf{P}_q^* = \mathbf{P}^* + \triangle_{\mathbf{P}^*}$ and $\mathbf{Q}_q^* = \mathbf{Q}^* + \triangle_{\mathbf{Q}^*}$. Then it is easy to see that

$$\mathbf{X}_q^* - \mathbf{X}^* = \mathbf{A}(\mathbf{P}^*\triangle_{\mathbf{Q}^*} + \triangle_{\mathbf{P}^*}\mathbf{Q}^* + \triangle_{\mathbf{P}^*}\triangle_{\mathbf{Q}^*})\mathbf{B}.$$

Given the range limits on allowable $\mathbf{P}$ and $\mathbf{Q}$ and that each range is quantized to $L_{\text{lev}}$ levels, we have that $\|\triangle_{\mathbf{P}^*}\|_{\max} \leq 1/(L_{\text{lev}} - 1)$ and $\|\triangle_{\mathbf{Q}^*}\|_{\max} \leq Q_{\max}/(L_{\text{lev}} - 1)$. Now, we can obtain a bound on the magnitudes of the elements of $\mathbf{X}_q^* - \mathbf{X}^*$ that hold uniformly over all $i, j$, as follows

$$\|\mathbf{X}_q^* - \mathbf{X}^*\|_{\max}$$
$$= \max_{i,j} |(\mathbf{A}(\mathbf{P}^*\triangle_{\mathbf{Q}^*} + \triangle_{\mathbf{P}^*}\mathbf{Q}^* + \triangle_{\mathbf{P}^*}\triangle_{\mathbf{Q}^*})\mathbf{B})_{i,j}|$$
$$\leq \frac{6 \cdot r \cdot r_1 \cdot r_2 \cdot A_{\max} B_{\max} Q_{\max}}{L_{\text{lev}}}, \tag{12}$$

where the inequality follows from a application of triangle inequality followed by the bounds on $\|\triangle_{\mathbf{P}^*}\|_{\max}$ and $\|\triangle_{\mathbf{Q}^*}\|_{\max}$ and the entry-wise bounds on elements of allowable $\mathbf{P}$ and $\mathbf{Q}$, and because $L_{\text{lev}} \geq 2$. Now, it is straight-forward to show that our choice of $\beta$ in (9) implies

$$L_{\text{lev}} \geq \frac{12 \cdot r \cdot r_1 \cdot r_2 \cdot A_{\max} B_{\max} Q_{\max}}{X_{\max}}, \tag{13}$$

so each entry of $\mathbf{X}_q^* - \mathbf{X}^*$ is bounded in magnitude by $X_{\max}/2$. It follows that each element of the candidate $\mathbf{X}_q^*$ constructed above is bounded in magnitude by $X_{\max}$, so $\mathbf{X}_q^*$ is indeed a valid element of the set $\mathcal{X}$.

Further, the approximation error analysis above also implies directly that

$$\frac{\|\mathbf{X}^* - \mathbf{X}_q^*\|_F^2}{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{i \in [n_1], j \in [n_2]} (\mathbf{A}\mathbf{P}_q^*\mathbf{Q}_q^*\mathbf{B} - \mathbf{A}\mathbf{P}^*\mathbf{Q}^*\mathbf{B})_{i,j}^2$$

$$\leq \frac{36 \cdot (r \cdot r_1 \cdot r_2)^2 (A_{\max} B_{\max} Q_{\max})^2}{L_{\text{lev}}^2}$$

$$\leq \frac{X_{\max}^2}{m},$$

where the last line follows from the fact that our specific choice of $\beta$ in (9) also implies

$$L_{\text{lev}} \geq \frac{6\sqrt{m} \cdot (r \cdot r_1 \cdot r_2) \cdot A_{\max} B_{\max} Q_{\max}}{X_{\max}}. \tag{14}$$

Now, evaluating the oracle term at the candidate $\mathbf{X}_q^* = \mathbf{A}\mathbf{P}_q^*\mathbf{Q}_q^*\mathbf{B}$, and using the fact that $\|\mathbf{P}_q^*\|_0 = \|\mathbf{P}^*\|_0$ and $\|\mathbf{Q}_q^*\|_0 = \|\mathbf{Q}^*\|_0$, we have

$$\frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_{\mathcal{S}}}\left[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2\right]}{n_1 n_2} \leq \frac{70 X_{\max}^2 \log m}{m} +$$

$$48(\sigma^2 + 2X_{\max}^2)(\beta+4)\log(n_1 \vee n_2)\left(\frac{\|\mathbf{P}^*\|_0 + \|\mathbf{Q}^*\|_0}{m}\right).$$

$\blacksquare$

REFERENCES

[1] A. Soni, S. Jain, J. Haupt, and S. Gonella, "Noisy matrix completion under sparse factor models," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3636–3661, June 2016.

[2] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[3] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, pp. 2057–2078, 2010.

[4] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[5] A. Soni and J. Haupt, "Estimation error guarantees for poisson denoising with sparse and structured dictionary models," in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 2002–2006.

[6] A. Soni, S. Jain, J. Haupt, and S. Gonella, "Error bounds for maximum likelihood matrix completion under sparse factor models," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Dec 2014, pp. 399–403.

[7] A. Soni and J. Haupt, "On the fundamental limits of recovering tree sparse vectors from noisy linear measurements," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 133–149, Jan 2014.

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

[9] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 931–940.

[10] Ian Porteous, Arthur U Asuncion, and Max Welling, "Bayesian matrix factorization with side information and dirichlet process mixtures.," in *AAAI*, 2010.

[11] Prateek Jain and Inderjit S Dhillon, "Provable inductive matrix completion," *arXiv preprint arXiv:1306.0626*, 2013.

[12] Miao Xu, Rong Jin, and Zhi-Hua Zhou, "Speedup matrix completion with side information: Application to multi-label learning," in *Advances in Neural Information Processing Systems*, 2013, pp. 2301–2309.

[13] E. Acar, G. Grdeniz, M. A. Rasmussen, D. Rago, L. O. Dragsted, and R. Bro, "Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics," in *2012 IEEE 12th International Conference on Data Mining Workshops*, Dec 2012, pp. 1–8.

[14] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, March 2010.

[15] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra, "Computing a nonnegative matrix factorization–provably," *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162, 2012.

[16] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3239–3252, 2012.

[17] J. D. Haupt, N. D. Sidiropoulos, and G. B. Giannakis, "Sparse dictionary learning from 1-bit data," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7664–7668.

[18] R. Gribonval and K. Schnass, "Dictionary identification – Sparse matrix-factorization via $l_1$ minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[19] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[20] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 3087–3090.

[21] A. R. Barron, C. Huang, J. Q. Li, and X. Luo, "The MDL principle, penalized likelihoods, and statistical risk," *Festschrift for Jorma Rissanen. Tampere University Press, Tampere, Finland*, 2008.

[22] Deepak Agarwal and Bee-Chung Chen, "flda: matrix factorization through latent dirichlet allocation," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 91–100.

[23] Steffen Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.

[24] Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon, "Matrix completion with noisy side information," in *Advances in Neural Information Processing Systems*, 2015, pp. 3447–3455.

[25] Donghyuk Shin, Suleyman Cetintas, Kuang-Chih Lee, and Inderjit S Dhillon, "Tumblr blog recommendation with boosted inductive matrix completion," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 203–212.

[26] Si Si, Kai-Yang Chiang, Cho-Jui Hsieh, Nikhil Rao, and Inderjit S. Dhillon, "Goal-directed inductive matrix completion," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, aug 2016.

[27] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1665–1697, 2012.

[28] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2006.