

Learning Latent Features with Pairwise Penalties in Low-Rank Matrix Completion

Kaiyi Ji, Jian Tan, Jinfeng Xu and Yuejie Chi, *Senior Member, IEEE*

Abstract—Low-rank matrix completion has achieved great success in many real-world data applications. A matrix factorization model that learns latent features is usually employed and, to improve prediction performance, the similarities between latent variables can be exploited by pairwise learning using the graph regularized matrix factorization (GRMF) method. However, existing GRMF approaches often use the squared loss to measure the pairwise differences, which may be overly influenced by dissimilar pairs and lead to inferior prediction. To fully empower pairwise learning for matrix completion, we propose a general optimization framework that allows a rich class of (non-)convex pairwise penalty functions. A new and efficient algorithm is developed to solve the proposed optimization problem, with a theoretical convergence guarantee under mild assumptions. In an important situation where the latent variables form a small number of subgroups, its statistical guarantee is also fully considered. In particular, we theoretically characterize the performance of the complexity-regularized maximum likelihood estimator, as a special case of our framework, which is shown to have smaller errors when compared to the standard matrix completion framework without pairwise penalties. We conduct extensive experiments on both synthetic and real datasets to demonstrate the superior performance of this general framework.

Index Terms—matrix factorization, pairwise learning, non-convex pairwise penalty.

I. INTRODUCTION

Low-rank matrix completion [1], [2] has been widely used in many real-world data applications, e.g., image restoration and collaborative filtering. A typical optimization problem follows the form [3], [4]

$$\min_{\mathbf{X}, \mathbf{Y}} \frac{1}{2} \|\Psi_{\Omega}(\mathbf{M} - \mathbf{X}\mathbf{Y}^T)\|_F^2 + \frac{\alpha}{2} (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2), \quad (1)$$

where $\mathbf{M} \in \mathbb{R}^{n \times m}$ is the data matrix, $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{Y} \in \mathbb{R}^{m \times d}$ are the low-rank factors, with typically $d \ll \min\{n, m\}$. The projection operator $\Psi_{\Omega}(\mathbf{M})$ retains the entries of the matrix \mathbf{M} in the index set Ω that denotes the observed indices, and $\alpha > 0$ is some regularization parameter.

The row vectors $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ of $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and $\{\mathbf{y}_j \in \mathbb{R}^d\}_{j=1}^m$ of $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m]^T$, known as *latent variables*, usually represent the features of two classes of

interdependent objects, e.g., user features and movie features in recommender systems [3], [5], respectively. Beyond this basic model, many variants have been considered for different application scenarios to further improve the estimation accuracy. For example, the similarities between latent variables can be exploited by pairwise learning. Define a graph $\mathcal{G}_X = (V_X, E_X, \mathbf{W})$, where the feature vectors $\{\mathbf{x}_i\}_{i=1}^n$ correspond to the vertices $V_X = \{1, 2, \dots, n\}$ of the graph, and the edges $E_X = V_X \times V_X$ are weighted by entries in a non-negative matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$. Similarly, define the graph $\mathcal{G}_Y = (V_Y, E_Y, \mathbf{U})$ for the feature vectors $\{\mathbf{y}_j\}_{j=1}^m$. The graph regularized matrix factorization (GRMF) method [6], [7], [8] aims to solve the following problem:

$$\min_{\mathbf{X}, \mathbf{Y}} \frac{1}{2} \|\Psi_{\Omega}(\mathbf{M} - \mathbf{X}\mathbf{Y}^T)\|_F^2 + \frac{\alpha}{2} (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + \gamma_X \sum_{i < j} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + \gamma_Y \sum_{s < t} u_{st} \|\mathbf{y}_s - \mathbf{y}_t\|_2^2, \quad (2)$$

which adds a smoothing graph regularizer to (1), a technique also known as Laplacian smoothing [9]. This encourages the feature vectors $\mathbf{x}_i \approx \mathbf{x}_j$ (respectively $\mathbf{y}_s \approx \mathbf{y}_t$) if the weight w_{ij} (respectively u_{st}) is large. It has been demonstrated that GRMF has the potential to reduce the estimation error and improve prediction performance [6], [7], [8].

A. Our contributions

It is commonly known that Laplacian smoothing using the squared loss as in (2) tends to enforce smoothness property in a global manner, and does not adapt well to inhomogeneity across different nodes [10]. Indeed, as illustrated in Fig. 1, the squared loss in (2) tends to impose a large penalty when \mathbf{x}_i is highly different from \mathbf{x}_j ; correspondingly, if w_{ij} is not so small, GRMF tends to push the solution \mathbf{x}_i to be close to \mathbf{x}_j which can lead to biased estimates as shown in [11], and dramatically affect the recovery performance. This is exacerbated further when the weight matrices \mathbf{W} and \mathbf{U} are constructed using side information of the feature vectors, e.g., the friendships of users and the attributes of movies in recommender systems, which could be very noisy [12] and inappropriately selected.

In this work, we introduce a large family of (non-)convex pairwise penalty functions in the low-rank matrix factorization framework, by proposing a general optimization framework

$$\min_{\mathbf{X}, \mathbf{Y}} \frac{1}{2} \|\Psi_{\Omega}(\mathbf{M} - \mathbf{X}\mathbf{Y}^T)\|_F^2 + \frac{\alpha}{2} (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) + \sum_{i < j} w_{ij} p(\|\mathbf{x}_i - \mathbf{x}_j\|_2, \gamma_X) + \sum_{s < t} u_{st} p(\|\mathbf{y}_s - \mathbf{y}_t\|_2, \gamma_Y), \quad (3)$$

K. Ji and J. Tan are with Department of Electrical and Computer Engineering, The Ohio State University; e-mail: {ji.367, tan.252}@osu.edu. This work is supported in part by the National Science Foundation under Grant No. 1717060.

J. Xu is with Department of Statistics and Actuarial Science, The University of Hong Kong; e-mail: xujf@hku.hk.

Y. Chi is with Department of Electrical and Computer Engineering, Carnegie Mellon University; e-mail: yuejiechi@cmu.edu. The work of Y. Chi is supported in part by ONR under the grant N00014-18-1-2142, by ARO under the grant ARO W911NF-18-1-0303, and by NSF under CIF-1806154, CIF-1826519 and ECCS-1818571.

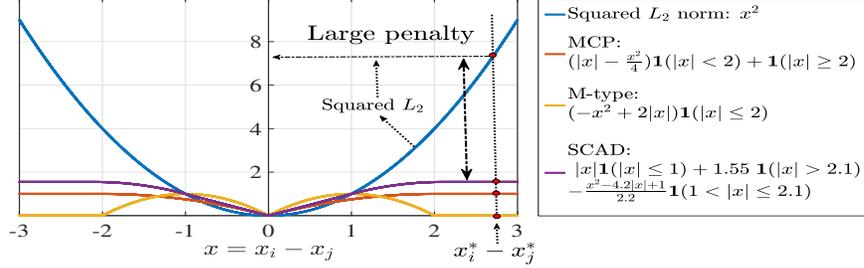


Fig. 1: Illustration of representative pairwise penalty functions $p(x, \gamma)$ w.r.t. a scalar x .

where $p(\cdot, \gamma)$ is a (non-)convex pairwise penalty function with a tuning parameter $\gamma \geq 0$. The pairwise penalty function takes Laplacian smoothing as a special case, where $p(x, \gamma) = \gamma x^2$. More importantly, it offers the opportunity to incorporate other important regularizers such as ℓ_1 penalty by setting $p(x, \gamma) = \gamma|x|$, which has proven its value in a different context for graph trend filtering [10], [13]. Furthermore, non-convex pairwise penalties can be used, such as the minimax concave penalty (MCP) [14], the smoothly clipped absolute deviation (SCAD) penalty [15] and the M-type function, as illustrated in Fig. 1 (the details will be given later). To the best of our knowledge, the power of general families of pairwise penalties in the context of low-rank matrix factorization has not been studied before and is the focus of this work.

With the optimization problem (3) in mind, our contributions in this paper are summarized below. First, we design a *novel and scalable* algorithm based on a modified alternating direction method of multipliers (ADMM) [16] to efficiently solve the proposed problem (3), by introducing Bregman divergence to certain subproblems. The allowed non-convex and non-smooth pairwise penalty functions complicate the optimization, which, if not handled carefully, can even result in divergent iterations. Theoretically, we characterize the convergence of our algorithm for a large family of pairwise penalty functions satisfying some mild properties. This requires new analysis since recent works on the convergence of ADMM for non-convex functions [17], [18], [19] cannot be directly applied to our problem. Moreover, the modified ADMM algorithm needs to solve an expensive linear system of equations that typically requires $O((m^2 + n^2) \times d^2)$ run time to invert two large matrices in each iteration. We provide a conjugate gradient (CG) based approach to obtain an *inexact* solution to this linear system of equations, using only $O(|\Omega| \times d)$ run time. Notably, we still guarantee the convergence using the inexact solutions during iterations.

Second, we study the statistical benefits of incorporating pairwise penalties. Fully characterizing the statistical guarantees for the proposed class of penalty functions is challenging. Instead, we restrict ourselves to an important situation where the latent variables form a small number of *subgroups*. Specifically, we investigate a subgroup-based model, where the feature vectors are considered in the same group if they are identical. For a partially observed matrix corrupted with Gaussian noise, we prove that the complexity-regularized maximum likelihood estimator, as a special case of the framework (3), can achieve a smaller error bound than

the one without considering the pairwise penalties, especially when the numbers of the subgroups are small. The class of sparsity-promoting pairwise penalty functions (possibly with a finite support) can be regarded as a computationally-efficient surrogate of the complexity-regularized maximum likelihood estimator. Interestingly, not only can we identify the subgroups automatically but also we significantly reduce the recovery error, as verified in Section V-B.

We propose a few heuristic rules to adaptively construct the weight matrices of the graph based on the partially observed matrix. As a result, our framework also applies to datasets that do not provide any side information. Our extensive experiments on both synthetic and real data demonstrate the superior performance of the proposed framework.

B. Organization of the paper and notations

The rest of the paper is organized as follows. Section II introduces the optimization framework. Section III develops the modified ADMM algorithm and states its convergence. Section IV describes the statistical properties of incorporating pairwise penalties in a scenario where latent variables form a small number of subgroups. Section V contains extensive experiments on both synthetic and real datasets to demonstrate the superior performance of this general framework. Finally, we conclude in Section VI. All proofs are delegated to the supplemental material.

Notations: Let $\mathbf{A} = [A_{ij}] \in \mathbb{R}^{n \times m}$ be a $n \times m$ matrix and $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^d$ be a d -dimensional vector. Recall the Frobenius norm $\|\mathbf{A}\|_F = (\sum_{ij} A_{ij}^2)^{1/2}$ and the infinity norm $\|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|$. Define ℓ_1 and ℓ_2 norms as $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$ and $\|\mathbf{x}\|_2 = (\sum_{i=1}^d x_i^2)^{1/2}$, respectively. The trace of a square matrix \mathbf{A} is denoted as $\text{tr}(\mathbf{A})$. Denote $|S|$ as the cardinality of a set S , and \otimes as the Kronecker product.

II. OPTIMIZATION FRAMEWORK

We develop a general optimization framework for incorporating pairwise penalties in low-rank matrix factorization, given in (3), that allows a wide class of (non-)convex pairwise penalty functions $p(z, \gamma)$, characterized by the following assumption.

Assumption 1. The penalty function $p(z, \gamma)$ satisfies

- 1) $p(z, \gamma) \geq 0$ for $z, \gamma \geq 0$,
- 2) For any given vector $\mathbf{v} \in \mathbb{R}^d$, there exists a constant $s_0 \geq 0$ independent of \mathbf{v} such that for $\forall \varsigma > s_0$, $f(\mathbf{u}) =$

$c\|\mathbf{u} - \mathbf{v}\|_2^2 + p(\|\mathbf{u}\|_2, \gamma)$ is strongly convex with respect to $\mathbf{u} \in \mathbb{R}^d$.

This assumption holds for a variety of (non-)convex pairwise penalty functions, including MCP [14] and SCAD [15], both of which are highly popular as sparsity-promoting non-convex regularizers in the statistics literature. For example, Assumption 1 holds for the MCP regularizer

$$p_{\text{MCP}}(z, \gamma) = \gamma|z| - \frac{z^2}{2t} \mathbf{1}(|z| \leq \gamma t) + \left(\frac{t\gamma^2}{2} - \gamma|z|\right) \mathbf{1}(|z| > \gamma t), \quad (4)$$

for $t > 0$ by setting $\varsigma_0 = 1/(2t)$, where $\mathbf{1}(\cdot)$ denotes the indicator function. Moreover, we introduce the following M-type function:

$$p_{\text{M}}(z, \gamma) = -\gamma(z^2 - 2b|z|) \mathbf{1}(|z| < 2b), \quad b > 0, \quad (5)$$

which satisfies Assumption 1 by letting $\varsigma_0 = \gamma$. In addition, Assumption 1 also holds for all convex functions, e.g., the ℓ_1 and the squared ℓ_2 norms. Thus, our framework unifies and generalizes existing GRMF approaches.

III. ALGORITHM AND CONVERGENCE

For the class of optimization problems introduced in Section II, we develop an efficient and general algorithm, dubbed **LLFMC**, with a strong theoretical convergence guarantee. We begin with the following equivalent constrained form of the main problem (3):

$$\begin{aligned} \min_{\mathbf{X}, \mathbf{Y}} \quad & \frac{1}{2} \|\Psi_{\Omega}(\mathbf{M} - \mathbf{X}\mathbf{Y}^T)\|_F^2 + \frac{\alpha}{2} (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) \\ & + \sum_{l \in \varepsilon_X} w_l p(\|\mathbf{p}_l\|_2, \gamma_X) + \sum_{j \in \varepsilon_Y} u_j p(\|\mathbf{q}_j\|_2, \gamma_Y) \end{aligned} \quad (6)$$

subject to $\mathbf{p}_l = \mathbf{x}_{l_1} - \mathbf{x}_{l_2}$, $\mathbf{q}_j = \mathbf{y}_{j_1} - \mathbf{y}_{j_2}$,

where the index set $\varepsilon_X = \{l = (l_1, l_2) : w_{l_1 l_2} > 0, l_1 < l_2\}$ and $\varepsilon_Y = \{j = (j_1, j_2) : u_{j_1 j_2} > 0, j_1 < j_2\}$. It can be seen that for every $l = (l_1, l_2) \in \varepsilon_X$, we can rewrite the first constraint in (6) as $\mathbf{p}_l = \mathbf{X}^T(\mathbf{e}_{l_1} - \mathbf{e}_{l_2})$, where \mathbf{e}_k is the k^{th} standard basis vector in \mathbb{R}^n . Let $\mathbf{P} \in \mathbb{R}^{d \times |\varepsilon_X|}$ be the matrix whose columns are given by \mathbf{p}_l^T with $l \in \varepsilon_X$, and let $\mathbf{E}_x \in \mathbb{R}^{n \times |\varepsilon_X|}$ be the matrix whose corresponding columns are given by $(\mathbf{e}_{l_1} - \mathbf{e}_{l_2})$. Symmetrically, we can define matrices $\mathbf{Q} \in \mathbb{R}^{d \times |\varepsilon_Y|}$ and $\mathbf{E}_y \in \mathbb{R}^{m \times |\varepsilon_Y|}$. The constraint in (6) can be rewritten in a matrix form as

$$\mathbf{P} - \mathbf{X}^T \mathbf{E}_x = \mathbf{0}, \quad \mathbf{Q} - \mathbf{Y}^T \mathbf{E}_y = \mathbf{0}. \quad (7)$$

Let $\mathcal{L}_{\eta}(\mathbf{P}, \mathbf{Q}, \mathbf{X}, \mathbf{Y}, \mathbf{\Lambda}, \mathbf{V})$ be the augmented Lagrangian function of the problem (6), which is given as

$$\begin{aligned} \mathcal{L}_{\eta} = \quad & \frac{1}{2} \|\Psi_{\Omega}(\mathbf{M} - \mathbf{X}\mathbf{Y}^T)\|_F^2 + \frac{\alpha}{2} (\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2) \\ & + \text{tr}(\mathbf{\Lambda}^T (\mathbf{P} - \mathbf{X}^T \mathbf{E}_x)) + \text{tr}(\mathbf{V}^T (\mathbf{Q} - \mathbf{Y}^T \mathbf{E}_y)) \\ & + \frac{\eta}{2} (\|\mathbf{P} - \mathbf{X}^T \mathbf{E}_x\|_F^2 + \|\mathbf{Q} - \mathbf{Y}^T \mathbf{E}_y\|_F^2) \\ & + \sum_{l \in \varepsilon_X} w_l p(\|\mathbf{p}_l\|_2, \gamma_X) + \sum_{j \in \varepsilon_Y} u_j p(\|\mathbf{q}_j\|_2, \gamma_Y), \end{aligned} \quad (8)$$

where $\mathbf{\Lambda} \in \mathbb{R}^{d \times |\varepsilon_X|}$ and $\mathbf{V} \in \mathbb{R}^{d \times |\varepsilon_Y|}$ are the dual variables.

We now describe our 2-step algorithm, LLFMC, based on a modified ADMM.

- 1) **Step 1.** Define an undirected graph $\mathcal{G}_X = (V_X, \varepsilon_X)$ with $V_X = \{1, 2, \dots, n\}$. We first use the standard depth-first-search algorithm to find cycles. Then, for each cycle of \mathcal{G}_X , we randomly cut one edge l off by letting $w_l = 0$. As a result, the graph \mathcal{G}_X becomes acyclic. Similar operations can be applied for \mathcal{G}_Y .
- 2) **Step 2.** In each iteration k , we do the following updates:
 - a) Primal updates:

$$\mathbf{P}^{k+1} = \arg \min_{\mathbf{P}} \mathcal{L}_{\eta}(\mathbf{P}, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k), \quad (9a)$$

$$\mathbf{Q}^{k+1} = \arg \min_{\mathbf{Q}} \mathcal{L}_{\eta}(\mathbf{P}^{k+1}, \mathbf{Q}, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k), \quad (9b)$$

$$\begin{aligned} \mathbf{X}^{k+1} = \arg \min_{\mathbf{X}} \quad & \left(\mathcal{L}_{\eta}(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k) \right. \\ & \left. + \frac{\alpha_x}{2} \|\mathbf{X} - \mathbf{X}^k\|_F^2 \right), \end{aligned} \quad (9c)$$

$$\begin{aligned} \mathbf{Y}^{k+1} = \arg \min_{\mathbf{Y}} \quad & \left(\mathcal{L}_{\eta}(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^{k+1}, \mathbf{Y}, \mathbf{\Lambda}^k, \mathbf{V}^k) \right. \\ & \left. + \frac{\alpha_y}{2} \|\mathbf{Y} - \mathbf{Y}^k\|_F^2 \right), \end{aligned} \quad (9d)$$

- b) Dual updates:

$$\begin{aligned} \mathbf{\Lambda}^{k+1} &= \mathbf{\Lambda}^k + \eta(\mathbf{P}^{k+1} - (\mathbf{X}^{k+1})^T \mathbf{E}_x), \\ \mathbf{V}^{k+1} &= \mathbf{V}^k + \eta(\mathbf{Q}^{k+1} - (\mathbf{Y}^{k+1})^T \mathbf{E}_y). \end{aligned} \quad (10)$$

Some discussions are in order. Step 1 is necessary to address the row-rank deficiency issue of the matrices \mathbf{E}_x and \mathbf{E}_y ¹ that prevents the application of convergence guarantees of standard ADMM, e.g. [17], [18], [19]. As shall be seen later, this step verifies Property 3 in Section III-B, which is needed to obtain the convergence guarantee of our algorithm.

Compared to standard ADMM, our algorithm further introduces Bregman divergences in Step 2 to the update of \mathbf{X} and \mathbf{Y} in (9c) and (9d) to improve the convergence performance. As shall be seen later, these modifications guarantee sufficient descents of \mathcal{L}_{η} during the subproblems (9c) and (9d) in each iteration.

In Fig. 2, we empirically investigate the impacts of Step 1 and the inclusion of Bregman divergence in Step 2 on the convergence performance of our algorithm. We can observe that without Bregman divergence in Step 2, ADMM will diverge after around 600 iterations, whereas ADMM with Bregman divergence in Step 2 has a better convergence performance as well as a lower relative error. In addition, without Step 1, ADMM still has a good convergence performance but achieves a larger relative error than that with Step 1. In sum, the Bregman divergence term in Step 2 indeed contributes to the better convergence performance of our modified ADMM algorithm, while Step 1 helps to attain a lower relative error but does not affect the convergence behavior much.

Remark 1. Compared to the Bregman ADMM [20], we do not apply Bregman divergence to \mathbf{P} and \mathbf{Q} in order to save $O((|\varepsilon_X| + |\varepsilon_Y|) \times d)$ space for tracking \mathbf{P}^k and \mathbf{Q}^k .

¹Since the summation of the entries in \mathbf{E}_x is equal to 0, the row rank of \mathbf{E}_x in the constraint (7) is at most $n - 1$. A similar argument holds for \mathbf{E}_y .

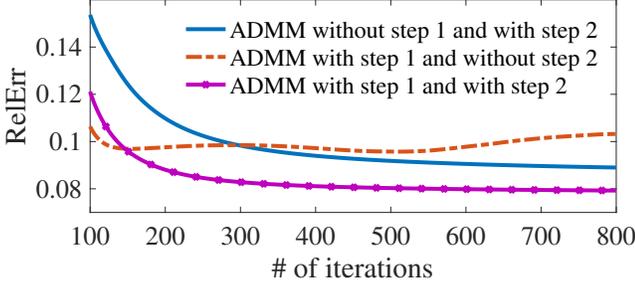


Fig. 2: Illustration of the impact of Step 1 and the inclusion of Bregman divergence Step 2 on the convergence performance. We use MCP as the pairwise penalty function, and randomly pick 30% of the pixels of a 512×512 image *Lena* to generate the partially observed matrix $\Psi_\Omega(\mathbf{M})$. The relative error is defined by $\text{RelErr} = \|\widehat{\mathbf{M}} - \mathbf{M}\|_F / \|\mathbf{M}\|_F$, where $\widehat{\mathbf{M}}$ is the estimate.

A. A Fast Inexact Solver via CG

We now discuss how to solve the subproblems in Step 2 efficiently. Based on (8), updating \mathbf{P} in (9a) is to solve the following proximal map for each of its column \mathbf{p}_l^T in parallel:

$$\mathbf{p}_l^{k+1} = \arg \min_{\mathbf{p}_l} \left(\frac{1}{2} \|\mathbf{p}_l^T - ((\mathbf{X}^k)^T \mathbf{E}_x[l] - \eta^{-1} \mathbf{\Lambda}^k[l])\|_2^2 + w_l \eta^{-1} p(\|\mathbf{p}_l\|_2, \gamma_X) \right), \quad (11)$$

where $\mathbf{E}_x[l]$ and $\mathbf{\Lambda}^k[l]$ are the column of \mathbf{E}_x and $\mathbf{\Lambda}^k$ corresponding to \mathbf{p}_l , respectively. In practice, we choose $\eta > 2\varsigma_0 \max_{l \in \varepsilon_X} w_l$, which, based on Assumption 1, implies that the problem in (11) has a unique solution. Notably, this solution has simple and explicit analytical forms for the penalty functions such as SCAD, MCP and the M-type function. We update \mathbf{Q} in a similar way by solving the proximal map for each column \mathbf{q}_j^T , given as

$$\mathbf{q}_j^{k+1} = \arg \min_{\mathbf{q}_j} \left(\frac{1}{2} \|\mathbf{q}_j^T - ((\mathbf{Y}^k)^T \mathbf{E}_y[j] - \eta^{-1} \mathbf{V}^k[j])\|_2^2 + u_j \eta^{-1} p(\|\mathbf{q}_j\|_2, \gamma_Y) \right), \quad (12)$$

where $\mathbf{E}_y[j]$ and $\mathbf{V}^k[j]$ are the column of \mathbf{E}_y and \mathbf{V}^k corresponding to \mathbf{q}_j , respectively.

Based on (9c), updating \mathbf{X} is to minimize

$$F(\mathbf{X}) = \frac{1}{2} \sum_{(i,j) \in \Omega} (M_{ij} - \mathbf{x}_i (\mathbf{y}_j^k)^T)^2 + \frac{\eta}{2} \|\mathbf{P}^{k+1} - \mathbf{X}^T \mathbf{E}_x\|_F^2 + \text{tr}((\mathbf{\Lambda}^k)^T (\mathbf{P}^{k+1} - \mathbf{X}^T \mathbf{E}_x)) + \frac{\alpha}{2} \|\mathbf{X}\|_F^2 + \frac{1}{2} \|\mathbf{X} - \mathbf{X}^k\|_F^2.$$

Recognizing that $F(\mathbf{X})$ is quadratic in \mathbf{X} , the optimizer can be found by solving a linear system of equations, where $\text{vec}(\mathbf{X}^T)$ satisfies

$$[\mathbf{G}_y^k + (\eta \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1) \mathbf{I}_n) \otimes \mathbf{I}_d] \text{vec}(\mathbf{X}^T) = \mathbf{c}, \quad (13)$$

where

$$\mathbf{c} = (\mathbf{b}_y^k)^T + \text{vec}((\mathbf{X}^k)^T + \eta \mathbf{P}^{k+1} \mathbf{E}_x^T + \mathbf{\Lambda}^k \mathbf{E}_x^T). \quad (14)$$

Algorithm 1 Learning latent features with pairwise penalties in low-rank matrix factorization (LLFMC)

- 1: **Input:** $\Psi_\Omega(\mathbf{M}) \in \mathbb{R}^{n \times m}$, $\text{rank } d \in \mathbb{N}^+$, $\text{MaxIter} > 0$, $\text{tol}_1, \text{tol}_2 > 0$.
- 2: **Initialize:** $\gamma_X, \gamma_Y, \alpha, \eta$. Set $\mathbf{X}^1, \mathbf{Y}^1, \mathbf{\Lambda}^1, \mathbf{V}^1$ as random matrices.
- 3: **for** $k = 1$ **to** MaxIter **do**
- 4: **for** $l \in \varepsilon_X$ **do**
- 5: Compute \mathbf{p}_l^{k+1} by solving the proximal map (11);
- 6: **end for**
- 7: **for** $j \in \varepsilon_Y$ **do**
- 8: Compute \mathbf{q}_j^{k+1} by solving the proximal map (12);
- 9: **end for**
- 10: Update $\mathbf{X}^{k+1}, \mathbf{Y}^{k+1}$ by solving (13) through CG;
- 11: $\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k + \eta (\mathbf{P}^{k+1} - (\mathbf{X}^{k+1})^T \mathbf{E}_x)$;
- 12: $\mathbf{V}^{k+1} = \mathbf{V}^k + \eta (\mathbf{Q}^{k+1} - (\mathbf{Y}^{k+1})^T \mathbf{E}_y)$;
- 13: **if** $D_k < \text{tol}_1$ or $|D_k - D_{k+1}| < \text{tol}_2$ **then**
- 14: break and output \mathbf{X}^{k+1} and \mathbf{Y}^{k+1} ;
- 15: **end if**
- 16: **end for**
- 17: **Output:** $\widehat{\mathbf{M}} = \mathbf{X}^{k+1} (\mathbf{Y}^{k+1})^T$.

Here, $\mathbf{G}_y^k = \text{diag}(\mathbf{G}_1, \dots, \mathbf{G}_n)$ is a block diagonal matrix with the i^{th} block given as $\mathbf{G}_i = \sum_{j \in \Omega_i} (\mathbf{y}_j^k)^T (\mathbf{y}_j^k) \in \mathbb{R}^{d \times d}$, where $\Omega_i = \{j : (i, j) \in \Omega\}$ is the index set of the observed entries in the i^{th} row. The row vector $\mathbf{b}_y^k = [\mathbf{b}_1, \dots, \mathbf{b}_n]$ concatenates the vectors $\mathbf{b}_i = \sum_{j \in \Omega_i} M_{ij} \mathbf{y}_j^k$, with \mathbf{y}_j^k being the j^{th} row of the matrix \mathbf{Y}^k .

Unfortunately, solving (13) requires inverting an $nd \times nd$ matrix, which is computationally demanding when dealing with large datasets. To address this issue, we apply the standard CG to directly minimize $f(\text{vec}(\mathbf{X}^T))$, where the function $f(\text{vec}(\mathbf{X}^T))$ is a re-write of $F(\mathbf{X})$ in terms of its vectorized argument $\text{vec}(\mathbf{X}^T)$, defined as

$$f(\mathbf{s}) = \frac{1}{2} \mathbf{s}^T [\mathbf{G}_y^k + (\eta \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1) \mathbf{I}_n) \otimes \mathbf{I}_d] \mathbf{s} - \mathbf{c}^T \mathbf{s}.$$

The most expensive part in each CG iteration is the Hessian-vector multiplication $\nabla^2 f(\mathbf{s}) \mathbf{s}$. Using the identity $(\mathbf{B}^T \otimes \mathbf{I}) \text{vec}(\mathbf{S}) = \text{vec}(\mathbf{S} \mathbf{B})$, we have

$$\nabla^2 f(\mathbf{s}) \mathbf{s} = \mathbf{G}_y^k \mathbf{s} + \text{vec}(\eta \mathbf{S} \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1) \mathbf{S}), \quad (15)$$

with $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_n] \in \mathbb{R}^{d \times n}$ and $\mathbf{s} = \text{vec}(\mathbf{S})$. Since \mathbf{G}_y^k is a block diagonal matrix, to compute $\mathbf{G}_y^k \mathbf{s}$ it suffices to obtain $\mathbf{G}_i \mathbf{s}_i$ for $1 \leq i \leq n$, which can be computed in $O(|\Omega_i|d)$ time by using $\mathbf{G}_i \mathbf{s}_i = \sum_{j \in \Omega_i} (\mathbf{y}_j^k \mathbf{s}_i) (\mathbf{y}_j^k)^T$.

The time complexity for computing a single CG iteration is $O((|\Omega| + \text{nnz}(\mathbf{E}_x \mathbf{E}_x^T)) \times d)$, where $\text{nnz}(\cdot)$ is the number of nonzeros. For our algorithm, we use a k_w -nearest neighbor method to select the weights w_{ij} , as introduced in section V-A. Thus, using the structure of the matrix \mathbf{E}_x , we have an upper bound for $\text{nnz}(\mathbf{E}_x \mathbf{E}_x^T)$, as shown in the following proposition.

Proposition 1. $\text{nnz}(\mathbf{E}_x \mathbf{E}_x^T) \leq n(k_w + 1)$.

The proof can be found in Appendix A. In most real-world applications, k_w is small and satisfies $n(k_w + 1) < |\Omega|$. Thus,

each CG iteration can be computed in $O(|\Omega| \times d)$ time. We stop CG if the approximation error of (13) is small enough:

$$\begin{aligned} & \left\| [\mathbf{G}_y^k + (\eta \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1) \mathbf{I}_n) \otimes \mathbf{I}_d] \text{vec}((\mathbf{X}^t)^T) - \mathbf{c} \right\|_2 \\ & \leq 10^3 \sqrt{dn} / (k + 1)^{1.2}, \end{aligned} \quad (16)$$

using the output of the t^{th} CG iteration \mathbf{X}^t . Note that we do not require CG to fully converge. Instead, we only need a very small number (empirically ≤ 5) of CG iterations to obtain an *inexact* solution \mathbf{X}^{k+1} that satisfies the stopping criteria (16), which still guarantees the convergence for our algorithm, as shown in Theorem 2 below. This step further speeds up our algorithm.

The main steps are summarized in Algorithm 1, where $D_k = \|\mathbf{X}^k - \mathbf{X}^{k+1}\|_F / (2\sqrt{dn}) + \|\mathbf{Y}^k - \mathbf{Y}^{k+1}\|_F / (2\sqrt{dm})$. In our implementation, we choose $\text{tol}_1 = 10^{-1}$ and $\text{tol}_2 = 10^{-4}$.

Remark 2. *The Graph Regularized Alternating Least Squares (GRALS) [6] algorithm is a GRMF method that is most relevant to ours, and can be regarded as solving a special case of our framework (2) when the objective function is convex and differentiable with respect to \mathbf{X} or \mathbf{Y} when fixing the other, and therefore much easier to optimize. In contrast, we propose an algorithm to solve the much more general optimization problem that is non-convex and non-smooth.*

B. Convergence Guarantees

This section proves the convergence of our algorithm. Let $\mathbf{U}^k = (\mathbf{P}^k, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k)$ be the exact solutions of the subproblems in (9). We first prove the convergence of the sequence $\{\mathbf{U}^k\}$. Then, we extend this convergence result to Algorithm 1 that solves (13) (or equivalently, (9c) and (9d)) inexactly.

Our convergence analysis adopts the following assumption on the boundedness of the generated sequence $(\mathbf{X}^k, \mathbf{Y}^k)$.

Assumption 2. *The sequence $(\mathbf{X}^k, \mathbf{Y}^k)$ generated by (9) is bounded, which means that there exist constants M_x, M_y such that $\|\mathbf{X}^k\|_F^2 \leq M_x$, $\|\mathbf{Y}^k\|_F^2 \leq M_y$ for all k .*

We note that directly proving the finiteness of the iterates is very challenging. Compared with previous analysis [19], [20], where the boundedness of the iterates is established typically based on stronger assumptions, e.g. the Lipschitz property of the gradient of the augmented Lagrangian function, which do not hold in our problem. Nonetheless, such an assumption does not seem uncommon in the optimization literature, particularly when dealing with functions that are non-smooth and non-convex, e.g. [21], [22]. In practice, we find that the generated sequence by our algorithm is well-bounded, and hence this assumption is rather mild.

Following [23], [24], we use the quantity $\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2$ as a measure of the convergence rate for the sequence $\{\mathbf{U}^k\}$. We have the following theorem, whose proof can be found in Appendix B.

Theorem 1. *Suppose Assumptions 1 and 2 are satisfied. Choosing a sufficiently large η , we have $\sum_{k=1}^{\infty} \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 < \infty$. Furthermore, if the penalty function $p(\cdot, \gamma)$ is*

semi-algebraic, then $\sum_{k=1}^{\infty} \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F < \infty$ and the sequence $\{\mathbf{U}^k\}$ converges to a stationary point of \mathcal{L}_η .

The first result $\sum_{k=1}^{\infty} \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 < \infty$ guarantees a subsequential convergence for our proposed algorithm under general penalty functions including MCP and SCAD. Furthermore, for all semi-algebraic penalty functions, which include all real polynomial functions and ℓ_p -norm with any $p > 0$ (see Examples 2 and 4 in [25]), we establish the convergence of the whole sequence $\{\mathbf{U}^k\}$ by exploring the Kurdyka-Łojasiewicz (KL) property of the augmented Lagrangian function in (8).

We further extend the convergence guarantee in Theorem 1 to the scenario when (9c) and (9d) are solved inexactly. Let $\hat{\mathbf{U}}^k = (\hat{\mathbf{P}}^k, \hat{\mathbf{Q}}^k, \hat{\mathbf{X}}^k, \hat{\mathbf{Y}}^k, \hat{\mathbf{\Lambda}}^k, \hat{\mathbf{V}}^k)$ be the sequence generated by Algorithm 1 using inexact solutions $\hat{\mathbf{X}}^k$ and $\hat{\mathbf{Y}}^k$ to (9c) and (9d). Define

$$\mathbf{t}_x^k = [\mathbf{G}_y^k + (\eta \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1) \mathbf{I}_n) \otimes \mathbf{I}_d] \text{vec}((\hat{\mathbf{X}}^k)^T) - \mathbf{c}$$

for $k \geq 1$, which measures the approximation error of solving (13). In a symmetric way, we can define \mathbf{t}_y^k . Based on (16), we have $\sum_{k=1}^{\infty} \|\mathbf{t}_x^k\|_2 < \infty$ and $\sum_{k=1}^{\infty} \|\mathbf{t}_y^k\|_2 < \infty$. We have the following theorem, whose proof can be found in Appendix C.

Theorem 2. *Suppose Assumption 1 and Assumption 2 hold with $(\hat{\mathbf{X}}^k, \hat{\mathbf{Y}}^k)$. Choosing a sufficiently large η , we have $\sum_{k=1}^{\infty} \|\hat{\mathbf{U}}^{k+1} - \hat{\mathbf{U}}^k\|_F^2 < \infty$. Furthermore, if the penalty function $p(\cdot, \gamma)$ is semi-algebraic, then $\sum_{k=0}^{\infty} \|\hat{\mathbf{U}}^{k+1} - \hat{\mathbf{U}}^k\|_F < \infty$ and the sequence $\{\hat{\mathbf{U}}^k\}$ converges to a stationary point of \mathcal{L}_η . Furthermore, the best running convergence rate of Algorithm 1 is $o(1/k)$.*

Different from using the exact solution in Theorem 1, Theorem 2 only requires $\hat{\mathbf{X}}^k$ to satisfy $\sum_{k=1}^{\infty} \|\mathbf{t}_x^k\|_2 < \infty$ (similarly for $\hat{\mathbf{Y}}^k$). This extension allows more efficient methods to find inexact solutions of (13), e.g., the CG approach used in our algorithm.

IV. STATISTICAL PROPERTIES

Fully characterizing the statistical properties of the proposed class of penalty functions in Section II is challenging. Instead, we restrict to an important class, where the latent vectors can be divided into subgroups, and rigorously derive the estimation error bound for this class. More importantly, our theoretical analysis motivates two interesting directions for our framework, i.e., subgroup identification and adaptive weights. We first introduce the subgroup-based model in Section IV-A. Then, the estimation error is characterized in Section IV-B. This characterization directly motivates us the use of adaptive weights in Section V-A.

A. Subgroup-based Model

Assume that there exists a ground truth low-rank matrix $\mathbf{M}^* \in \mathbb{R}^{n \times m}$, which can be factored as $\mathbf{M}^* = \mathbf{X}^* \mathbf{Y}^{*T}$, where $\mathbf{X}^* \in \mathbb{R}^{n \times d}$, and $\mathbf{Y}^* \in \mathbb{R}^{m \times d}$. The observations are given as $\Psi_\Omega(\mathbf{M}) = \Psi_\Omega(\mathbf{M}^* + \mathbf{N})$, which are corrupted by the additive noise \mathbf{N} composed of i.i.d. Gaussian entries $\mathcal{N}(0, \sigma^2)$ with zero mean and variance σ^2 . For an integer s with $s < nm$, assume that each index pair $(i, j) \in [n] \times [m]$ is included

in the observed index set Ω independently with probability $s/(nm)$. The elements of the matrix $\Psi_\Omega(\mathbf{M})$ are independent conditioned on the set Ω . Assume $\|\mathbf{X}^*\|_\infty \leq C_x$, $\|\mathbf{Y}^*\|_\infty \leq C_y$, and $\|\mathbf{M}^*\|_\infty \leq C_m$, respectively.

The feature vectors $\{\mathbf{x}_i^*\}$ and $\{\mathbf{y}_j^*\}$ form some latent subgroups, where two features \mathbf{x}_i^* and \mathbf{x}_j^* (respectively \mathbf{y}_s^* and \mathbf{y}_t^*) are considered in the same subgroup if $\mathbf{x}_i^* = \mathbf{x}_j^*$ (respectively $\mathbf{y}_s^* = \mathbf{y}_t^*$). One key difference with the existing group based models, e.g., the group sparsity model [26], is that the subgroup structure in our model is not assumed to be known a priori. Studies [27] have shown that the users with similar types in recommender systems often have the same features, implying a natural subgroup structure. Nevertheless, our subgroup-based model can be fairly general. For example, the number of subgroups can even be as large as the matrix dimension.

To facilitate analysis, we introduce additional useful notations. Let $\mathcal{G}(\mathbf{X}) = (\mathcal{G}_1^x, \dots, \mathcal{G}_{k_x}^x)$ and $\mathcal{G}(\mathbf{Y}) = (\mathcal{G}_1^y, \dots, \mathcal{G}_{k_y}^y)$ be two sets of mutually exclusive partitions of the indices $\{1, \dots, n\}$ and $\{1, \dots, m\}$, which satisfy $\mathbf{x}_i = \mathbf{x}_j$ for $\forall i, j \in \mathcal{G}_u^x$, $1 \leq u \leq k_x$ and $\mathbf{y}_s = \mathbf{y}_t$ for $\forall s, t \in \mathcal{G}_v^y$, $1 \leq v \leq k_y$. We use $|\mathcal{G}(\mathbf{X})| = k_x$ and $|\mathcal{G}(\mathbf{Y})| = k_y$ to denote the number of subgroups of $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$, respectively.

B. Estimation Error Bound

We first introduce the complexity-regularized maximum likelihood estimator [28], [29], and show it is a special case of our framework. Then, we prove that this estimator achieves a lower error bound compared to the standard trace-norm regularized matrix completion without pairwise penalties.

Consider the following complexity-regularized maximum likelihood estimator

$$\widehat{\mathbf{M}} = \arg \min_{\mathbf{H} \in \mathcal{H}} \|\Phi_\Omega(\mathbf{M} - \mathbf{H})\|_F^2 + \lambda(|\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|), \quad (17)$$

where the finite set \mathcal{H} is given as

$$\mathcal{H} \triangleq \{\mathbf{H} = \mathbf{X}\mathbf{Y}^T : \mathbf{X} \in \mathcal{X}, \mathbf{Y} \in \mathcal{Y}, \|\mathbf{H}\|_\infty \leq C_m\}, \quad (18)$$

with the set \mathcal{X} in (18) defined by

$$\mathcal{X} \triangleq \{\mathbf{X} \in \mathbb{R}^{n \times d} : X_{ij} \in \{-C_x + \frac{2C_x}{K}t, t = 0, \dots, K-1\}\}, \quad (19)$$

where $K = 2^{\lceil \mu \log_2(n \vee m) \rceil}$, $n \vee m = \max(n, m)$ and $\mu > 1$. In a symmetrical way, we can also define the set \mathcal{Y} .

Using the following proposition, proved in Appendix D, we show that the estimator (17) is a special case of our framework (3).

Proposition 2. *Let $w_{ij} = 1/(K_i(n - K_i))$, where K_i is the total number of feature vectors \mathbf{x}_s that satisfy $\|\mathbf{x}_s - \mathbf{x}_i\|_2 < 2C_x/K$. Then, we have for all $\mathbf{X} \in \mathcal{X}$,*

$$|\mathcal{G}(\mathbf{X})| = \sum_{i,j} w_{ij} \Upsilon(\|\mathbf{x}_i - \mathbf{x}_j\|_2), \quad (20)$$

where the indicator function $\Upsilon(z) = 0$ if $z < 2C_x/K$ and 1 otherwise. A symmetrical result holds for $|\mathcal{G}(\mathbf{Y})|$.

Moreover, the constraints $\|\mathbf{X}\|_\infty \leq C_x$ and $\|\mathbf{Y}\|_\infty \leq C_y$ in (19) play a similar role as the regularizer $\alpha(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2)/2$ in our framework (3). In the following theorem, we provide an error bound for the estimator (17).

Theorem 3. *Set $\lambda \geq 8\mu d(\sigma^2 + 4C_m^2/3) \log(n \vee m)$, then the estimator (17) satisfies the error bound*

$$\begin{aligned} \frac{1}{nm} \mathbb{E} \left[\|\widehat{\mathbf{M}} - \mathbf{M}^*\|_F^2 \right] &\leq \frac{C_1 \log s}{s} + \frac{dC_2}{(\mu \log(n \vee m))^2} + \\ &\left(\frac{6\lambda}{\log(n \vee m)} + \mu C_3 \right) \left(\frac{(m+n+2)}{s} \log(n \vee m) + \right. \\ &\left. \frac{(|\mathcal{G}(\mathbf{X}^*)| + |\mathcal{G}(\mathbf{Y}^*)|)d}{s} \log(n \vee m) \right), \end{aligned} \quad (21)$$

where the expectation is with respect to the joint distribution of $(\Omega, \Phi_\Omega(\mathbf{M}))$ and C_1, C_2, C_3 are positive constants related to C_x, C_y and C_m .

We refer to Appendix E for its proof. We compare the error bound (21) with the one for the trace-norm regularized matrix completion obtained by [30], which assumes the observed entries are corrupted by additive Gaussian noise. Instantiating [30, Corollary 2] using our notations, we obtain an error bound

$$\frac{\|\widehat{\mathbf{M}} - \mathbf{M}^*\|_F^2}{nm} \leq c(\sigma^2 + t) \frac{(n+m)d}{s} \log(n \vee m) \quad (22)$$

with high probability, where c and t are two constants related to C_x, C_y and C_m . Note that our error bound

$$O\left(\left(|\mathcal{G}(\mathbf{X}^*)| + |\mathcal{G}(\mathbf{Y}^*)|\right)d + m + n\right) \log(n \vee m)/s$$

can be much lower than (22) provided that the total number of subgroups $|\mathcal{G}(\mathbf{X}^*)| + |\mathcal{G}(\mathbf{Y}^*)| \ll n + m$. However, to directly optimize (17) with the regularizer (20) is undesirable in practice since the indicator function complicates the optimization and the sets \mathcal{X}, \mathcal{Y} are discretized. Nonetheless, Theorem 3 provides some theoretical justifications on the benefits of incorporating nonconvex pairwise penalties in reducing the estimation error.

V. NUMERICAL EXPERIMENTS

In this section, we first provide some heuristic rules to adaptively set the graph weights based on the partially observed data matrix in Section V-A. We then compare our algorithm **LLFMC** to the trace-norm regularized **MC** without pairwise penalties [31] and the state-of-the-art **GRMF** algorithm, i.e., the graph regularized alternating Least Squares (**GRALS**) [6] on different graphs and evaluate the effectiveness of the proposed adaptive weights in Section V-A. We use both synthetic and real datasets to verify the performance of the proposed framework. All experiments are performed using Matlab/Octave on a PC with macOS system, Intel 5 core 2.9 GHz CPU and 16G RAM.

A. Adaptive Weights Selection

Existing works [32], [33], [6], [7] often construct the weight matrices \mathbf{W} and \mathbf{U} in (3) by incorporating side information, which may not always be available. Motivated by the analysis

in Section IV-B, we introduce adaptive weights that are computed based on the partially observed matrix itself without any additional side information. Due to symmetry, we only explain how to adaptively select w_{ij} , and u_{st} follows similarly. We consider the following weighted nearest neighbor rule for generating the weights w_{ij} , where

$$w_{ij} = \mathbf{1}_{ij}^{k_w} / d(\mathbf{x}_i, \mathbf{x}_j), \quad i \neq j, \quad (23)$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ approximates the similarity between \mathbf{x}_i and \mathbf{x}_j , and the indicator function $\mathbf{1}_{ij}^{k_w}$ is 1 if \mathbf{x}_j belongs to one of the k_w -nearest neighbors of \mathbf{x}_i and 0 otherwise. The rule (23) makes w_{ij} large during the computation to enforce $\mathbf{x}_i \approx \mathbf{x}_j$, similar to the adaptive weight chosen in [34] for adaptive LASSO.

Below we provide two heuristic ways to approximate $d(\mathbf{x}_i, \mathbf{x}_j)$ using $\Psi_\Omega(\mathbf{M})$. Recall $\Omega_i = \{j : (i, j) \in \Omega\}$ is the index set of the observed entries in the i^{th} row of $\Psi_\Omega(\mathbf{M})$. The first rule is based on the observation that the i^{th} and j^{th} rows of \mathbf{M} should be close if $\mathbf{x}_i \approx \mathbf{x}_j$ when the noise is not too large. Therefore, we define

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \left(\frac{\sum_{t \in \Omega_i \cap \Omega_j} (M_{it} - M_{jt})^2}{|\Omega_i \cap \Omega_j|} \right)^{\frac{1}{2}}, & \mathcal{S}_i \cap \mathcal{S}_j \neq \emptyset \\ 0 & \text{otherwise} \end{cases}. \quad (24)$$

The key idea of $d_1(\mathbf{x}_i, \mathbf{x}_j)$ in (24) is to use the ratings of user \mathbf{x}_i and user \mathbf{x}_j on the same items to measure their similarity. However, the above rule does not work very well when the observations are very sparse, since the number of items simultaneously rated by both user \mathbf{x}_i and user \mathbf{x}_j can be quite small. In that case, we use

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{\sum_{t \in \Omega_i \cup \Omega_j} (\bar{M}_{it} - \bar{M}_{jt})^2}{|\Omega_i \cup \Omega_j|} \right)^{\frac{1}{2}}, \quad (25)$$

where

$$\bar{M}_{it} = \begin{cases} M_{it} & \text{if } t \in \Omega_i, \\ \sum_{r \in \mathcal{R}_t} M_{rt} / |\mathcal{R}_t|, & \text{if } t \in \Omega_j \setminus \Omega_i, \end{cases}$$

where \mathcal{R}_t is the set of indices of the observed entries for the t^{th} column of \mathbf{M} . In (25), for the items rated only by user \mathbf{x}_i , we estimate the ratings of user \mathbf{x}_j on these items by taking the average of the ratings on these items. This step is motivated by the observation that the ratings of each user are not very far away from their average. Using $d_2(\mathbf{x}_i, \mathbf{x}_j)$, we obtain a solution $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$ of the optimization problem (3) with weights computed by (23). Next, we use $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$ to compute the distance by $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i^{(1)} - \mathbf{x}_j^{(1)}\|_2$. The performance of these heuristic rules will be examined in the experiments below.

B. Synthetic Data with Subgroups

To verify Theorem 3, we generate a ground truth matrix $\mathbf{M}^* = \mathbf{X}^* \mathbf{Y}^{*T} \in \mathbb{R}^{n \times n}$, $\mathbf{X}^* \in \mathbb{R}^{n \times d}$, $\mathbf{Y}^* \in \mathbb{R}^{n \times d}$ in the following way. Let \mathbf{U}_1 be composed of entries drawn uniformly at random from $[0, 1]^d$ and $\{\mathbf{U}_{i+1} - \mathbf{U}_i, 1 \leq i < k_x\}$ be i.i.d.

and uniformly distributed on $[0, 10]^d$. Similarly we generate $\mathbf{V}_1, \dots, \mathbf{V}_{k_y}$. We evenly divide the feature vectors $\{\mathbf{x}_i^*\}$ into k_x subgroups and set the features in the i^{th} subgroup to be \mathbf{U}_i . We divide $\{\mathbf{y}_j^*\}$ in a similar way. When $k_x = k_y = n$, there exist no subgroups. Normalizing by a constant, we scale \mathbf{M}^* to satisfy $\|\mathbf{M}^*\|_F = 10^6$. The observed entries of $\Psi_\Omega(\mathbf{M})$ are sampled from $\mathbf{M} = \mathbf{M}^* + \mathbf{N}$ uniformly at random with a sample rate $0 < \rho < 1$, with the noise \mathbf{N} containing i.i.d. entries following $\mathcal{N}(0, \sigma^2)$. Moreover, we use the rule (24) to compute the graph weights by setting $k_w = 0.15n$. We fix $n = 200$ and $d = 5$. Let $\bar{\mathbf{M}}$ be the estimate of \mathbf{M}^* .

Relative error. We study three cases where the number of subgroups is small ($k_x = k_y = 20$), medium ($k_x = k_y = 50$) and large ($k_x = k_y = 200$). Fig. 3 shows that our method LLMFC with the MCP significantly reduces the recovery error compared to standard MC, even when no real subgroups exist.

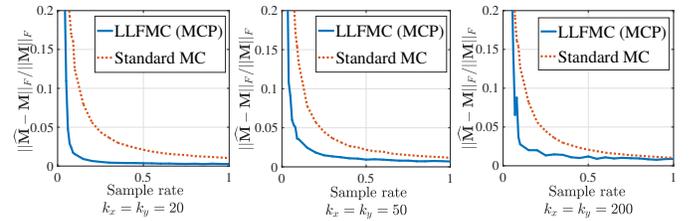


Fig. 3: The relative error w.r.t. sample rate: $n = 200$, $\sigma = 10^2$.

Subgroup identification. Similar to [35], we identify the subgroups for the recovered features $\{\hat{\mathbf{x}}_i\}$ and $\{\hat{\mathbf{y}}_j\}$ through the matrices \mathbf{S}^x , \mathbf{S}^y , where the features $\hat{\mathbf{x}}_u$ and $\hat{\mathbf{x}}_v$ (respectively $\hat{\mathbf{y}}_s$ and $\hat{\mathbf{y}}_t$) are in the same subgroup if and only if $S_{uv}^x = 1$ (respectively $S_{st}^y = 1$). We set $S_{uv}^x = 1$ if $\|\hat{\mathbf{x}}_u - \hat{\mathbf{x}}_v\|_2 < 0.01 \min\{\|\hat{\mathbf{x}}_u\|_2, \|\hat{\mathbf{x}}_v\|_2\}$ and 0 otherwise. Similarly, we define the matrix \mathbf{S}^y . Due to the limited space, we only visualize the matrix \mathbf{S}^x in Fig. 4. The results clearly show that our method recovers the subgroup structure quite well. In contrast, the standard MC cannot identify this latent structure.

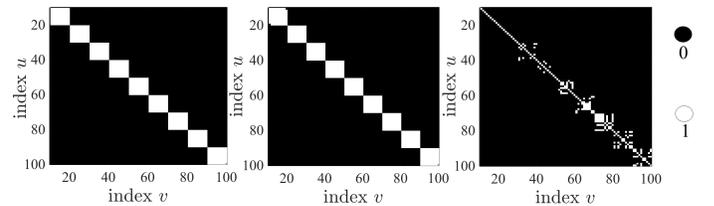


Fig. 4: Illustration of \mathbf{S}^x obtained by the ground truth (left), LLMFC (MCP) (middle) and the standard MC (right): $n = 100$, $\sigma = 100$, $k_x = k_y = 10$ and the sample rate $\rho = 0.3$.

C. Comparison to GRALS on Real Data

We conduct the experiments on two groups of real collaborative filtering datasets: Jester² [36] and MovieLens³ [37]. Their

²<http://goldberg.berkeley.edu/jester-data>

³<http://grouplens.org/datasets/movielens>

statistics are presented in Table I. The Jester datasets contain anonymous ratings of 100 jokes by users, where the ratings are real values ranging from -10.00 to 10.00 . For MovieLens100k and MotiveLens1M datasets, the ratings of the movies have 5 scores (1-5).

TABLE I: Summary statistics of datasets

Datasets	# of users (n)	# of items (m)	Density
Jester1	24983	100	0.7247
Jester2	23500	100	0.7272
Jester3	24938	100	0.2474
MovieLens100K	943	1682	0.0630
MovieLens1M	6040	3706	0.0447

Following [6], we evaluate MovieLens100K using the five provided data splits. For MovieLens1M, we randomly split the observed ratings into a training set (90%) and a test set (10%). For Jester datasets, we randomly choose 10% of the observed ratings as the training set and the remaining 90% as the test set. We use the root mean squared error (RMSE) to measure the prediction performance, given by $\text{RMSE} = (\sum_{i=1}^N (\hat{r}_i - r_i)^2 / N)^{1/2}$, where r_i is the observed rating in the test set, and \hat{r}_i is the predicted rating and N is the total number of ratings in the test set. When reporting the performance, the best performance in each scenario is bolded.

For our algorithm LLFMC, we choose the regularization parameters γ_X and γ_Y from $2^{\{-2, -1, \dots, 10\}}$, fix $\alpha = 1$ and set $\eta = 10^4$ to guarantee the convergence. In addition, we fix $d = 4$ for MovieLens datasets and $d = 100$ for Jester datasets. For MCP (4), we select t from $\{0.5, 2, 20\}$, and for the M-type penalty function, we fix $b = 3$. We downloaded the code of **GRALS** from the author’s website⁴, and changed the number of iterations from 10 (default) to 100 to ensure its convergence.

Comparison with GRALS with side information. We first use the same graph constructed in [6] based on side information obtained therein. Table II shows that our method still makes improvement over GRALS using the same graphs as in [6], which might be due to the bias reduction effect of using non-convex penalties.

TABLE II: RMSE on MovieLens100K, using the same graphs as in [6] that were constructed by side information.

Methods	MC	GRALS	LLFMC (M-type)	LLFMC (MCP)
RMSE	0.973	0.945	0.930	0.927

Adaptive weights. We test the performance of different algorithms under different adaptive schemes of weight selection. The first one is an unweighted 100-nearest neighbor graph, where the nearest neighbors are determined using the distance (25), and the edges all weighted by 1. Table III summarizes the performance of different methods using an unweighted 100-nearest neighbor graph for both users and movies. It shows that with MCP and M-type penalties, our method outperforms GRALS in all datasets.

TABLE III: RMSE on real datasets using unweighted 100-nearest neighbor graphs.

Datasets	GRALS	LLFMC (MCP)	LLFMC (M-type)
Jester1	4.801	4.712	4.710
Jester2	4.880	4.831	4.838
Jester3	6.701	6.651	6.634
MovieLens100K	0.950	0.934	0.936
MovieLens1M	0.873	0.859	0.861

TABLE IV: RMSE on real datasets using the graph based on the adaptive weights (23) using the distance (25).

Datasets	GRALS	LLFMC (MCP)	LLFMC (M-type)
Jester1	4.713	4.651	4.672
Jester2	4.789	4.801	4.792
Jester3	6.461	6.551	6.461
MovieLens100K	0.919	0.909	0.914
MovieLens1M	0.851	0.847	0.850

We next examine the performance using the graph constructed based on our adaptive weights defined in (23) using the distance (25). We set $k_w = k_u = 10$ in our wights (23) for MovieLens datasets and set $k_w = 20$, $k_u = 2$ for Jester datasets. Table IV summarizes the performance of different methods using the adaptive weights. Compared with Table III and Table II, it can be seen that the proposed adaptive weights significantly improve both the performance of GRALS and our methods on all datasets. In addition, our method still outperforms the competitors in almost all experiments.

Subgroup identification. We illustrate the subgroups of users and movies from the MovieLens100K dataset computed by LLFMC (MCP) following the similar manner of Fig. 4. Fig. 5 clearly shows that the users and movies exhibit subgroup structures; only 200 users and 200 movies are plotted for a better presentation. To get more useful information, we investigate the 2^{th} largest user subgroup in Fig. 6. Interestingly, we find out that most users in this subgroup have ages between 25 and 35, consisting of mainly engineers and students.

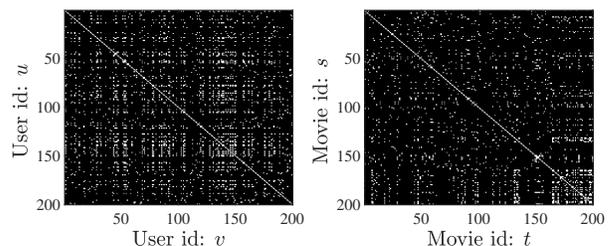


Fig. 5: Illustration of S^x (left) and S^y (right): white spots imply that the users/items are in the same subgroup.

VI. CONCLUSIONS

We propose a new optimization framework to learn latent features in low-rank matrix completion by a wide class of (non-)convex pairwise penalty functions. To efficiently solve this class of optimization problems, we develop an efficient algorithm and prove its convergence guarantee. On the statistical side, we characterize the complexity-regularized maximum

⁴<http://nikrao.github.io/Code.html>

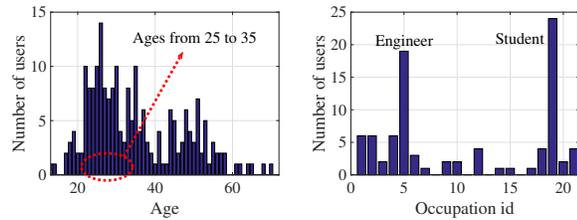


Fig. 6: Further insights into the 2th largest user subgroup. Left: the age distribution. Right: the occupation distribution for the users with age between 25 and 35 in the same subgroup.

likelihood estimator under the subgroup-based model, which sheds light to the statistical benefit of incorporating pairwise penalty functions. Extensive experiments on both synthetic and real datasets confirm the superior performance of our framework.

REFERENCES

- [1] Y. Chi, Y. M. Lu, and Y. Chen, “Nonconvex optimization meets low-rank matrix factorization: An overview,” *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.
- [2] Y. Chen and Y. Chi, “Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14–31, 2018.
- [3] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
- [4] P. K. Gopalan, L. Charlin, and D. Blei, “Content-based recommendations with poisson factorization,” in *Advances in Neural Information Processing Systems*, 2014, pp. 3176–3184.
- [5] J. Bennett, S. Lanning, and N. Netflix, “The netflix prize,” in *In KDD Cup and Workshop in conjunction with KDD*, 2007.
- [6] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, “Collaborative filtering with graph information: Consistency and scalable methods,” in *Advances in neural information processing systems*, 2015, pp. 2107–2115.
- [7] Z. Zhao, L. Zhang, X. He, and W. Ng, “Expert finding for question answering via graph regularized matrix completion,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 993–1004, 2015.
- [8] F. Monti, M. Bronstein, and X. Bresson, “Geometric matrix completion with recurrent multi-graph neural networks,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3700–3710.
- [9] A. J. Smola and R. Kondor, “Kernels and regularization on graphs,” in *Learning theory and kernel machines*. Springer, 2003, pp. 144–158.
- [10] Y.-X. Wang, J. Sharpnack, A. Smola, and R. Tibshirani, “Trend filtering on graphs,” in *Artificial Intelligence and Statistics*, 2015, pp. 1042–1050.
- [11] S. Ma and J. Huang, “A concave pairwise fusion approach to subgroup analysis,” *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 410–423, 2017.
- [12] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, “Matrix completion with noisy side information,” in *Advances in Neural Information Processing Systems*, 2015, pp. 3447–3455.
- [13] R. Varma, H. Lee, J. Kovacevic, and Y. Chi, “Vector-valued graph trend filtering with non-convex penalties,” *IEEE Transactions on Signal and Information Processing over Networks*, pp. 1–1, 2019.
- [14] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.
- [15] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
- [17] F. Wang, W. Cao, and Z. Xu, “Convergence of multi-block Bregman ADMM for nonconvex composite problems,” *Science China Information Sciences*, vol. 61, no. 12, pp. 101–122, 2018.
- [18] M. Hong, Z.-Q. Luo, and M. Razaviyayn, “Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems,” *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.
- [19] Y. Wang, W. Yin, and J. Zeng, “Global convergence of ADMM in nonconvex nonsmooth optimization,” *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.
- [20] H. Wang and A. Banerjee, “Bregman alternating direction method of multipliers,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2816–2824.
- [21] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, “Stochastic subgradient method converges on tame functions,” *Foundations of Computational Mathematics*, pp. 1–36.
- [22] J. C. Duchi and F. Ruan, “Stochastic methods for composite and weakly convex optimization problems,” *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3229–3259, 2018.
- [23] B. He, L. Hou, and X. Yuan, “On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming,” *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2274–2312, 2015.
- [24] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, “Parallel multi-block ADMM with $O(1/k)$ convergence,” *Journal of Scientific Computing*, vol. 71, no. 2, pp. 712–736, 2017.
- [25] J. Bolte, S. Sabach, and M. Teboulle, “Proximal alternating linearized minimization or nonconvex and nonsmooth problems,” *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
- [26] J. Kim, R. D. Monteiro, and H. Park, “Group sparsity in nonnegative matrix factorization,” in *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 2012, pp. 851–862.
- [27] J. Delporte, A. Karatzoglou, T. Matuszyk, and S. Canu, “Socially enabled preference learning from implicit feedback data,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 145–160.
- [28] J. Q. Li and A. R. Barron, “Mixture density estimation,” in *Advances in neural information processing systems*, 2000, pp. 279–285.
- [29] E. D. Kolaczyk and R. D. Nowak, “Multiscale generalised linear models for nonparametric function estimation,” *Biometrika*, vol. 92, no. 1, pp. 119–133, 2005.
- [30] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion,” *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.
- [31] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
- [32] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, “Recommender systems with social regularization,” in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 287–296.
- [33] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, “Matrix completion on graphs,” *arXiv preprint arXiv:1408.1717*, 2014.
- [34] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [35] L. Han and Y. Zhang, “Learning multi-level task groups in multi-task learning,” in *AAAI*, 2015, pp. 2638–2644.
- [36] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: A constant time collaborative filtering algorithm,” *information retrieval*, vol. 4, no. 2, pp. 133–151, 2001.
- [37] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *ACM Transactions on Interactive Intelligent Systems (TiIS)*, vol. 5, no. 4, p. 19, 2016.
- [38] R. B. Bapat, *Graphs and matrices*. Springer, 2010, vol. 27.
- [39] G. Li and T. K. Pong, “Global convergence of splitting methods for nonconvex composite optimization,” *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2434–2460, 2015.
- [40] 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, 2016.
- [41] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

APPENDIX

A. Proof of Proposition 1

Let \mathbf{e}_i be the i^{th} row vector of the matrix \mathbf{E}_x . Recalling the definition of the matrix \mathbf{E}_x given in Section III, we make the following three key observations for \mathbf{E}_x .

- 1) Each \mathbf{e}_i has exactly k_w non zeros, $1 \leq i \leq n$.
- 2) For any two row vectors \mathbf{e}_i and \mathbf{e}_j , there exists at most one coordinate at which \mathbf{e}_i and \mathbf{e}_j are both non-zero.
- 3) Each column vector of the matrix \mathbf{E}_x has exactly two non zeros.

Based on these observations, we claim that each row vector of $\mathbf{E}_x \mathbf{E}_x^T$, i.e., $\mathbf{e}_i \mathbf{E}_x^T = [\mathbf{e}_i \mathbf{e}_1^T, \mathbf{e}_i \mathbf{e}_2^T, \dots, \mathbf{e}_i \mathbf{e}_n^T]$ has at most $k_w + 1$ non zeros for any $1 \leq i \leq n$. We prove this result by contradiction. Suppose there exists a vector $\mathbf{e}_s \mathbf{E}_x^T = [\mathbf{e}_s \mathbf{e}_1^T, \dots, \mathbf{e}_s \mathbf{e}_n^T]$ that has $k_w + 2$ nonzeros. Without loss of generality, we assume that $s = 1$ and $\mathbf{e}_1 \mathbf{e}_t^T \neq 0$ for all $1 \leq t \leq k_w + 2$. Let $\mathcal{C}_1 = \{c_u : \mathbf{e}_1(c_u) \neq 0, 1 \leq u \leq k_w\}$ be a set of coordinates at which \mathbf{e}_1 is non-zero. Based on the observation 2), we have, for all $\mathbf{e}_t, 2 \leq t \leq k_w + 2$, there exists exactly one coordinate in $\{d_s, 1 \leq s \leq k_w\}$ at which \mathbf{e}_t is non-zero. Thus, there must exist one coordinate c_v in \mathcal{C}_1 such that $\mathbf{e}_1(c_v) \neq 0$, $\mathbf{e}_p(c_v) \neq 0$ and $\mathbf{e}_q(c_v) \neq 0$, for certain $2 \leq p, q \leq k_w + 2$. However, this contradicts the observation 3). By contradiction, we can conclude that each vector $\mathbf{e}_i \mathbf{E}_x^T$ has at most $k_w + 1$ non zeros. Then, the proof is completed by summing over all rows of $\mathbf{E}_x \mathbf{E}_x^T$.

B. Proof of Theorem 1

We first provide a useful property of our algorithm.

Proposition 3 (Full column rank). *After Step 1 of the LLFMC algorithm, the matrices \mathbf{E}_x and \mathbf{E}_y defined in (7) are both full column rank. Thus, there exist positive constants σ_x, σ_y such that $\mathbf{E}_x^T \mathbf{E}_x \succeq \sigma_x \mathbf{I}$ and $\mathbf{E}_y^T \mathbf{E}_y \succeq \sigma_y \mathbf{I}$.*

Proof. Recall that after Step 1, the graph becomes acyclic, where \mathbf{E}_x is the incidence matrix associated to this graph. By [38, Lemma 2.5], we have that all columns $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{|\varepsilon_X|}$ of \mathbf{E}_x are linearly independent and hence \mathbf{E}_x is full column rank. Similar result holds for \mathbf{E}_y . Then, using Remark 1 in [39], we finish the proof. \square

This proposition immediately leads to the following lemma.

Lemma 1. *For any $\mathbf{A}_x \in \mathbb{R}^{d \times |\varepsilon_X|}$ and $\mathbf{A}_y \in \mathbb{R}^{d \times |\varepsilon_Y|}$, we have $\|\mathbf{A}_x \mathbf{E}_x^T\|_F^2 \geq \sigma_x \|\mathbf{A}_x\|_F^2$ and $\|\mathbf{A}_y \mathbf{E}_y^T\|_F^2 \geq \sigma_y \|\mathbf{A}_y\|_F^2$.*

Proof. Recalling that $\mathbf{E}_x^T \mathbf{E}_x - \sigma_x \mathbf{I} \succeq 0$, we have $\|\mathbf{A} \mathbf{E}_x^T\|_F^2 - \sigma_x \|\mathbf{A}\|_F^2 = \text{tr}[\mathbf{A}_x (\mathbf{E}_x^T \mathbf{E}_x - \sigma_x \mathbf{I}) \mathbf{A}_x^T] = \sum_{i=1}^d \mathbf{A}_x^i (\mathbf{E}_x^T \mathbf{E}_x - \sigma_x \mathbf{I}) (\mathbf{A}_x^i)^T \geq 0$, where \mathbf{A}_x^i is the i^{th} row vector of the matrix \mathbf{A}_x . Similarly, we can prove $\|\mathbf{A}_y \mathbf{E}_y^T\|_F^2 \geq \sigma_y \|\mathbf{A}_y\|_F^2$. \square

We next establish the following Lemmas 2, 3 and 4 to lower bound the descent of \mathcal{L}_η during each subproblem.

Lemma 2 (Upper-bound the descents of multipliers). *Suppose that Assumption 2 holds. Then, For each $k \in \mathbb{N}^+$, there exists a constant $L_y > 0$ such that*

$$\begin{aligned} & \|\Lambda^{k+1} - \Lambda^k\|_F^2 \\ & \leq \frac{2}{\sigma_x} \|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F^2 + \frac{2(M_x + 1)L_y}{\sigma_x} \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2 \\ & \quad + \frac{2L_y + 2(\alpha + 1)^2}{\sigma_x} \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2. \end{aligned} \quad (26)$$

Proof. Noting that $\text{vec}((\mathbf{X}^{k+1})^T)$ is the solution to

$$\begin{aligned} & (\mathbf{G}_y^k + (\eta \mathbf{E}_x \mathbf{E}_x^T + (\alpha + 1)) \otimes \mathbf{I}_d) \text{vec}((\mathbf{X}^{k+1})^T) \\ & = (\mathbf{b}_y^k)^T + \text{vec}((\mathbf{X}^k)^T + \eta \mathbf{P}^{k+1} \mathbf{E}_x^T + \Lambda^k \mathbf{E}_x^T), \end{aligned} \quad (27)$$

which, together with $\Lambda^{k+1} - \Lambda^k = \eta(\mathbf{P}^{k+1} - (\mathbf{X}^{k+1})^T \mathbf{E}_x)$ and the identity $\text{vec}(\mathbf{A}\mathbf{B}) = (\mathbf{B}^T \otimes \mathbf{I}) \text{vec}(\mathbf{A})$, implies

$$\begin{aligned} \text{vec}(\Lambda^{k+1} \mathbf{E}_x^T) & = (\mathbf{G}_y^k + (\alpha + 1) \mathbf{I}_{nd}) \text{vec}((\mathbf{X}^{k+1})^T) \\ & \quad - (\mathbf{b}_y^k)^T - \text{vec}((\mathbf{X}^k)^T). \end{aligned} \quad (28)$$

Using (28), we further obtain

$$\begin{aligned} & \|(\Lambda^{k+1} - \Lambda^k) \mathbf{E}_x^T\|_F^2 = \|\text{vec}((\Lambda^{k+1} - \Lambda^k) \mathbf{E}_x^T)\|_2^2 \\ & = \|(\mathbf{G}_y^k - \mathbf{G}_y^{k-1}) \text{vec}((\mathbf{X}^{k+1})^T) + (\mathbf{b}_y^{k-1} - \mathbf{b}_y^k)^T \\ & \quad + \mathbf{G}_y^{k-1} \text{vec}((\mathbf{X}^{k+1} - \mathbf{X}^k)^T) + \text{vec}((\mathbf{X}^{k-1} - \mathbf{X}^k)^T) \\ & \quad + (\alpha + 1) \text{vec}((\mathbf{X}^{k+1} - \mathbf{X}^k)^T)\|_2^2 \\ & \leq 2M_x \|\mathbf{G}_y^k - \mathbf{G}_y^{k-1}\|_F^2 + 2\|\mathbf{X}^{k-1} - \mathbf{X}^k\|_F^2 \\ & \quad + 2(\alpha + 1)^2 \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 + 2\|\mathbf{b}_y^k - \mathbf{b}_y^{k-1}\|_2^2 \\ & \quad + 2\|\mathbf{G}_y^{k-1}\|_F^2 \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2, \end{aligned} \quad (29)$$

where the inequality follows from the facts that $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq (\|\mathbf{a}\|_2 + \|\mathbf{b}\|_2)^2 \leq 2(\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2)$, $\|\mathbf{A}\mathbf{a}\|_2 \leq \|\mathbf{A}\|_F \|\mathbf{a}\|_2$ and $\|\mathbf{a}\mathbf{b}\|_2 \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$. Recalling the definition of \mathbf{G}_y^k and \mathbf{b}_y^k in (14) and using Assumption 2, there exists a sufficiently large constant L_y such that

$$\begin{aligned} & \|\mathbf{G}_y^k\|_F^2 \leq L_y, \|\mathbf{G}_y^k - \mathbf{G}_y^{k-1}\|_F^2 \leq L_y \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2 \\ & \|\mathbf{b}_y^k\|_2^2 \leq L_y, \|\mathbf{b}_y^k - \mathbf{b}_y^{k-1}\|_2^2 \leq L_y \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2, \end{aligned}$$

which, by (29) and Lemma 1, yields the proof. \square

Using a similar approach as in Lemma 2, we have

$$\begin{aligned} & \|\mathbf{V}^{k+1} - \mathbf{V}^k\|_F^2 \\ & \leq \frac{2}{\sigma_y} \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2 + \frac{2(M_y + 1)L_x}{\sigma_y} \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ & \quad + \frac{2L_x + 2(\alpha + 1)^2}{\sigma_y} \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2. \end{aligned} \quad (30)$$

Lemma 3 (Lower bound descents of \mathcal{L}_η for \mathbf{P}, \mathbf{Q} subproblems). *Let η_0 be a constant satisfying $\eta_0 > 2\varsigma_0 \max_{i \in \varepsilon_X} (w_i)$ and $\eta_0 > 2\varsigma_0 \max_{j \in \varepsilon_Y} (u_j)$, where $\varepsilon_X, \varepsilon_Y, w_i, u_j$ are defined in (6). Then, for any $\eta > \eta_0$ and each $k \in \mathbb{N}^+$, we have*

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \Lambda^k, \mathbf{V}^k) - \mathcal{L}_\eta(\mathbf{P}^k, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \Lambda^k, \mathbf{V}^k) \\ & \leq -\frac{(\eta - \eta_0)}{2} \|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2. \end{aligned}$$

Proof. Define the function $\tilde{p}(\mathbf{p}_l, \lambda_X) = w_l p(\|\mathbf{p}_l\|_2, \gamma_X) + \eta_0 \|\mathbf{p}_l - (\mathbf{X}^k)^T \mathbf{E}_x[l]\|_2^2 / 2$. Recalling Assumption 1 and noting

$\eta_0 > 2\widehat{w}\zeta_0$, we have that the function \widetilde{p} is strongly convex in \mathbf{p}_l . Based on the optimality of \mathbf{p}_l^{k+1} for (11), we have

$$\mathbf{0} \in \partial_{\mathbf{p}_l} \widetilde{p}(\mathbf{p}_l^{k+1}, \gamma_X) + \mathbf{\Lambda}^k[l] + (\eta - \eta_0)(\mathbf{p}_l^{k+1} - (\mathbf{X}^k)^T \mathbf{E}_x[l]). \quad (31)$$

Let $\mathbf{d}_k := -\mathbf{\Lambda}^k[l] - (\eta - \eta_0)(\mathbf{p}_l^{k+1} - (\mathbf{X}^k)^T \mathbf{E}_x[l])$, which is a subgradient of the convex function \widetilde{p} at $\mathbf{p}_l = \mathbf{p}_l^{k+1}$. By the definition of the subgradient of a convex function, we have

$$\widetilde{p}(\mathbf{p}_l^k, \gamma_X) - \widetilde{p}(\mathbf{p}_l^{k+1}, \gamma_X) \geq \langle \mathbf{d}_k, \mathbf{p}_l^k - \mathbf{p}_l^{k+1} \rangle. \quad (32)$$

Let $F_l(\mathbf{p}_l) = \widetilde{p}(\mathbf{p}_l, \gamma_X) + (\mathbf{p}_l - (\mathbf{X}^k)^T \mathbf{E}_x[l])\mathbf{\Lambda}^k[l] + (\eta - \eta_0)\|\mathbf{p}_l - (\mathbf{X}^k)^T \mathbf{E}_x[l]\|_2^2/2$. Using (32), we have

$$\begin{aligned} & F_l(\mathbf{p}_l^{k+1}) - F_l(\mathbf{p}_l^k) \\ &= \widetilde{p}(\mathbf{p}_l^{k+1}, \gamma_X) - \widetilde{p}(\mathbf{p}_l^k, \gamma_X) + (\mathbf{p}_l^{k+1} - \mathbf{p}_l^k)\mathbf{\Lambda}^k[l] \\ & \quad - \frac{\eta - \eta_0}{2} (\|\mathbf{p}_l^k - (\mathbf{X}^k)^T \mathbf{E}_x[l]\|_2^2 - \|\mathbf{p}_l^{k+1} - (\mathbf{X}^k)^T \mathbf{E}_x[l]\|_2^2) \\ & \stackrel{(i)}{=} \widetilde{p}(\mathbf{p}_l^{k+1}, \gamma_X) - \widetilde{p}(\mathbf{p}_l^k, \gamma_X) - \langle \mathbf{d}_k, \mathbf{p}_l^{k+1} - \mathbf{p}_l^k \rangle \\ & \quad - \frac{\eta - \eta_0}{2} \|\mathbf{p}_l^{k+1} - \mathbf{p}_l^k\|_2^2 \leq -\frac{\eta - \eta_0}{2} \|\mathbf{p}_l^{k+1} - \mathbf{p}_l^k\|_2^2, \end{aligned} \quad (33)$$

where (i) follows from the fact that $\|\mathbf{a} + \mathbf{c}\|_2^2 - \|\mathbf{b} + \mathbf{c}\|_2^2 = \|\mathbf{a} - \mathbf{b}\|_2^2 + 2\langle \mathbf{b} + \mathbf{c}, \mathbf{a} - \mathbf{b} \rangle$. Using (33), we have

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k) - \mathcal{L}_\eta(\mathbf{P}^k, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k) \\ &= \sum_{l \in \varepsilon_X} (F_l(\mathbf{p}_l^{k+1}) - F_l(\mathbf{p}_l^k)) \leq -\frac{\eta - \eta_0}{2} \|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2, \end{aligned}$$

which finishes the proof. \square

By the symmetry of \mathbf{P} and \mathbf{Q} , we derive similar results for \mathbf{Q} by replacing \mathbf{P} in Lemma 3 by \mathbf{Q} .

Lemma 4 (Upper bound the descents of \mathcal{L}_η for \mathbf{X}, \mathbf{Y} subproblems). For $\forall k \in \mathbb{N}^+$, we have

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^{k+1}, \mathbf{Y}^{k+1}, \mathbf{\Lambda}^k, \mathbf{V}^k) \\ & \quad - \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k) \\ & \leq -\frac{1}{2} (\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2) \end{aligned}$$

Proof. Since \mathbf{X}^{k+1} is the minimizer of the \mathbf{X} -subproblems in (9) respectively, we obtain

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^{k+1}, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k) + \frac{1}{2} \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \\ & \leq \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k), \end{aligned} \quad (34)$$

Similar result holds for \mathbf{Y} . \square

Proof of Theorem 1: Based on the above established lemmas, we now prove Theorem 1. We first introduce some notations.

$$\begin{aligned} \sigma_0 &= 2\eta^{-1}\sigma_x^{-1}, \sigma_1 = 2(M_x + 1)L_y\eta^{-1}\sigma_x^{-1} + 2\eta^{-1}\sigma_y^{-1}, \\ \sigma_2 &= \frac{1}{2} - 2(L_y + (\alpha + 1)^2)\eta^{-1}\sigma_x^{-1} - 2(M_y + 1)L_x\eta^{-1}\sigma_y^{-1}, \\ \sigma_3 &= \frac{1}{2} - (2L_x + 2(\alpha + 1)^2)\eta^{-1}\sigma_y^{-1}, \\ \sigma_* &= \min\{(\eta - \eta_0)/2, \sigma_2 - \sigma_0, \sigma_3 - \sigma_1\}, \end{aligned} \quad (35)$$

where constants $\sigma_x, \sigma_y, L_x, L_y$ are given in Lemma 2 M_x, M_y are given in Assumption 2 and η_0 is given in Lemma 3. Based on the above notations, we also define

$$\eta_1 = \max\left\{\frac{4 + 4L_y + 4(\alpha + 1)^2}{\sigma_x} + \frac{4(M_y + 1)L_x}{\sigma_y}, \frac{4L_x + 4(\alpha + 1)^2}{\sigma_y} + \frac{4(M_x + 1)L_y}{\sigma_x}, \eta_0 + \frac{1}{4}\right\}. \quad (36)$$

Based on (8) and (9), we have

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^{k+1}, \mathbf{Y}^{k+1}, \mathbf{\Lambda}^{k+1}, \mathbf{V}^{k+1}) \\ & \quad - \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^{k+1}, \mathbf{Y}^{k+1}, \mathbf{\Lambda}^k, \mathbf{V}^k) \\ & = \|\mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^k\|_F^2/\eta + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|_F^2/\eta. \end{aligned} \quad (37)$$

Combining (26), (30), (37), Lemma 3 and Lemma 4, we have

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{U}^{k+1}) - \mathcal{L}_\eta(\mathbf{U}^k) \\ & \leq -\frac{\eta - \eta_0}{2} (\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2) \\ & \quad + \sigma_0 \|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F^2 + \sigma_1 \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2 \\ & \quad - \sigma_2 \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 - \sigma_3 \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2. \end{aligned} \quad (38)$$

Let $\widetilde{\mathbf{U}}^k = (\mathbf{P}^k, \mathbf{Q}^k, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1})$ and define $\widehat{\mathcal{L}}_\eta$ as

$$\widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k) = \mathcal{L}_\eta(\mathbf{U}^k) + \sigma_0 \|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F^2 + \sigma_1 \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2,$$

which, in conjunction with (38), indicates that

$$\begin{aligned} & \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^{k+1}) - \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k) \\ & \leq -\sigma_* \left(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 \right. \\ & \quad \left. + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2 \right). \end{aligned} \quad (39)$$

Combining (35) and (36), we have $\sigma_* > 0$ for any $\eta > \eta_1$, which implies that $\widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k)$ is non-increasing.

Next, we show that the sequences (\mathbf{U}^k) and $(\widetilde{\mathbf{U}}^k)$ are bounded. The boundedness of the sequence $(\mathbf{X}^k, \mathbf{Y}^k)$ follows from Assumption 2. Recalling (28) and using Lemma 1 yield

$$\|\mathbf{\Lambda}^k\|_F^2 \leq 2(M_x + 1)L_y + 2M_x + 2(\alpha + 1)^2 M_x/\sigma_x \quad (40)$$

A symmetric result holds for $\|\mathbf{V}^k\|_F^2$. Thus, the sequence $(\mathbf{\Lambda}^k, \mathbf{V}^k)$ is bounded. Since $p(\cdot, \gamma) \geq 0$, we further have

$$\begin{aligned} \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k) & \geq \frac{\eta}{2} \|\mathbf{P}^k - (\mathbf{X}^k)^T \mathbf{E}_x\|_F^2 + \frac{1}{\eta} \|\mathbf{\Lambda}^k\|_F^2 - \frac{1}{2\eta} (\|\mathbf{\Lambda}^k\|_F^2 + \|\mathbf{V}^k\|_F^2) \\ & \quad + \frac{\eta}{2} \|\mathbf{Q}^k - (\mathbf{Y}^k)^T \mathbf{E}_y\|_F^2 + \frac{1}{\eta} \|\mathbf{V}^k\|_F^2. \end{aligned} \quad (41)$$

Since $\widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k)$ is non-increasing, $\widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^1) \geq \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k)$ for $\forall k \in \mathbb{N}$, which, combined with (41), implies

$$\begin{aligned} & \frac{\eta}{2} \|\mathbf{P}^k - (\mathbf{X}^k)^T \mathbf{E}_x\|_F^2 + \frac{1}{\eta} \|\mathbf{\Lambda}^k\|_F^2 + \frac{\eta}{2} \|\mathbf{Q}^k - (\mathbf{Y}^k)^T \mathbf{E}_y\|_F^2 + \frac{1}{\eta} \|\mathbf{V}^k\|_F^2 \\ & \leq \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^1) + \frac{1}{2\eta} \|\mathbf{\Lambda}^k\|_F^2 + \frac{1}{2\eta} \|\mathbf{V}^k\|_F^2 < \infty. \end{aligned}$$

Then, using the boundedness of $(\mathbf{X}^k, \mathbf{Y}^k, \mathbf{\Lambda}^k, \mathbf{V}^k)$ and the properties of the Frobenius norm that $\|\mathbf{A} - \mathbf{B}\|_F \geq \|\mathbf{A}\|_F - \|\mathbf{B}\|_F$ and $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_F$, we further obtain

$$\begin{aligned} \|\mathbf{P}^k\|_F &\leq \|\mathbf{X}^k\|_F \|\mathbf{E}_x\|_F + \frac{1}{\eta} \|\mathbf{\Lambda}^k\|_F < \infty, \\ \|\mathbf{Q}^k\|_F &\leq \|\mathbf{Y}^k\|_F \|\mathbf{E}_y\|_F + \frac{1}{\eta} \|\mathbf{V}^k\|_F < \infty, \end{aligned}$$

which implies that the sequence $(\mathbf{P}^k, \mathbf{Q}^k)$ is bounded. Combining the results above yields the boundedness of the sequences (\mathbf{U}^k) and $(\tilde{\mathbf{U}}^k)$. Thus, there exists a convergent subsequence $(\tilde{\mathbf{U}}^{k_i})$. Suppose it converges to $\tilde{\mathbf{U}}^*$. By the continuity of the function $\hat{\mathcal{L}}_\eta$, we obtain $\liminf_{i \rightarrow \infty} \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k_i}) \geq \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)$, which, together with (39) that $\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k)$ is non-increasing, implies that the sequences $(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k_i}))$ and $(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k))$ are bounded below by $\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)$. Using the inequality (39), we have, for any n

$$\begin{aligned} \sigma_* \sum_{k=1}^n \left(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \right. \\ \left. + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2 \right) &\leq \sum_{k=1}^n \left(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) \right) \\ &\leq \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^1) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) < \infty, \end{aligned} \quad (42)$$

which, combined with Lemma 2, yields $\sum_{k=1}^\infty \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F^2 < \infty$ and $\lim_{k \rightarrow \infty} \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F \rightarrow 0$.

Let $\mathbf{U}^* = (\mathbf{P}^*, \mathbf{Q}^*, \mathbf{X}^*, \mathbf{Y}^*, \mathbf{\Lambda}^*, \mathbf{V}^*)$ denote any cluster point of sequence (\mathbf{U}^k) . Applying Fermat's rule to (9) yields

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{X}} \left(\frac{1}{2} \|\Psi_\Omega(\mathbf{X}^{k+1}(\mathbf{Y}^k)^T - \mathbf{M})\|_F^2 + \frac{\alpha}{2} \|\mathbf{X}^{k+1}\|_F^2 \right) \\ &\quad + \mathbf{E}_x(\mathbf{\Lambda}^k)^T + (\mathbf{X}^{k+1} - \mathbf{X}^k) - \eta \mathbf{E}_x \left((\mathbf{P}^{k+1})^T - \mathbf{E}_x^T \mathbf{X}^{k+1} \right), \\ \mathbf{0} &= \nabla_{\mathbf{Y}} \left(\frac{1}{2} \|\Psi_\Omega(\mathbf{X}^{k+1}(\mathbf{Y}^{k+1})^T - \mathbf{M})\|_F^2 + \frac{\alpha}{2} \|\mathbf{Y}^{k+1}\|_F^2 \right) \\ &\quad + \mathbf{E}_y(\mathbf{V}^k)^T + (\mathbf{Y}^{k+1} - \mathbf{Y}^k) - \eta \mathbf{E}_y \left((\mathbf{Q}^{k+1})^T - \mathbf{E}_y^T \mathbf{Y}^{k+1} \right), \end{aligned}$$

and

$$\begin{aligned} \mathbf{0} &\in \partial_{\mathbf{P}} \sum_{l \in \varepsilon_X} w_l p(\mathbf{p}_l^{k+1}, \gamma_X) + \mathbf{\Lambda}^k + \eta \left(\mathbf{P}^{k+1} - (\mathbf{X}^k)^T \mathbf{E}_x \right), \\ \mathbf{0} &\in \partial_{\mathbf{Q}} \sum_{j \in \varepsilon_Y} u_j p(\mathbf{q}_j^{k+1}, \gamma_Y) + \mathbf{V}^k + \eta \left(\mathbf{Q}^{k+1} - (\mathbf{Y}^k)^T \mathbf{E}_y \right), \end{aligned}$$

which, in conjunction with (9) and $\|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F \rightarrow 0$, implies $\mathbf{P}^* - (\mathbf{X}^*)^T \mathbf{E}_x = \mathbf{0}$, $\mathbf{Q}^* - (\mathbf{Y}^*)^T \mathbf{E}_y = \mathbf{0}$ and

$$\begin{aligned} \mathbf{0} &\in \partial_{\mathbf{P}} \sum_{l \in \varepsilon_X} w_l p(\mathbf{p}_l^*, \gamma_X) + \mathbf{\Lambda}^*, \quad \mathbf{0} \in \partial_{\mathbf{Q}} \sum_{j \in \varepsilon_Y} u_j p(\mathbf{q}_j^*, \gamma_Y) + \mathbf{V}^*, \\ \mathbf{0} &= \nabla_{\mathbf{X}} \left(\frac{1}{2} \|\Psi_\Omega(\mathbf{X}^*(\mathbf{Y}^*)^T - \mathbf{M})\|_F^2 + \frac{\alpha}{2} \|\mathbf{X}^*\|_F^2 \right) + \mathbf{E}_x(\mathbf{\Lambda}^*)^T, \\ \mathbf{0} &= \nabla_{\mathbf{Y}} \left(\frac{1}{2} \|\Psi_\Omega(\mathbf{X}^*(\mathbf{Y}^*)^T - \mathbf{M})\|_F^2 + \frac{\alpha}{2} \|\mathbf{Y}^*\|_F^2 \right) + \mathbf{E}_y(\mathbf{V}^*)^T, \end{aligned}$$

which implies that $\mathbf{0} \in \partial \mathcal{L}_\eta(\mathbf{U}^*)$ and thus \mathbf{U}^* is a stationary point of \mathcal{L}_η . The above result establishes a subsequential convergence for the proposed algorithm.

We next show that when the penalty function is semi-algebraic, we enable to provide a global convergence for the proposed algorithm by using the KL property. Let \mathcal{Q} be a set of all cluster points of the sequence $(\tilde{\mathbf{U}}^k)$. For any point $\tilde{\mathbf{U}}^*$

in \mathcal{Q} , let $(\tilde{\mathbf{U}}^{k_j})$ be a subsequence of $(\tilde{\mathbf{U}}^k)$ converging to the point $\tilde{\mathbf{U}}^*$. As we prove before (see (39) and the text above (42)), the sequence $(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}))$ is non-increasing and lower-bounded, and hence is convergent (without loss of generality, we suppose it converges to C). Then, we have

$$\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) = \lim_{j \rightarrow \infty} \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k_j}) = \lim_{k \rightarrow \infty} \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) = C, \quad (43)$$

which implies that the function $\hat{\mathcal{L}}_\eta(\cdot)$ is constant on \mathcal{Q} . Suppose that the penalty function $p(\cdot, \gamma)$ is semi-algebraic. Based on Example 4 in [25], we have $\|\cdot\|_2$ is also semi-algebraic. Then, according to the item 4 in Example 2 in [25] that composition of semi-algebraic functions are semi-algebraic, we have the function $p(\|\cdot\|_2, \gamma)$ is also semi-algebraic. Now, note that all individual terms in the augmented Lagrangian function $\hat{\mathcal{L}}_\eta(\cdot)$ are semi-algebraic. Based on item 3 in Example 2 in [25] that finite sums of semi-algebraic functions is semi-algebraic, we have that $\hat{\mathcal{L}}_\eta(\cdot)$ is semi-algebraic, and hence satisfies the KL property (see [25, Theorem 3]) at all points in \mathcal{Q} . Based on KL inequality in [17, Lemma 2.2], we have, there exist $r > 0, \delta > 0$ and a smooth concave function $\phi : [0, r] \rightarrow \mathbb{R}^+$ (which satisfies three conditions in [17, (a),(b),(c) in Definition 2.3]) such that

$$\text{dist}(0, \partial \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}})) \phi'(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) \geq 1, \quad (44)$$

for all $\tilde{\mathbf{U}}$ satisfying $\text{dist}(\tilde{\mathbf{U}}, \mathcal{Q}) < \delta$ and $\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) < \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}) < \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) + r$, where $\text{dist}(\mathbf{x}, S) := \inf\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in S\}$. Based on the definition of \mathcal{Q} , we have $\lim_{k \rightarrow \infty} \text{dist}(\tilde{\mathbf{U}}^k, \mathcal{Q}) = 0$, which, combined with the (43) that $\lim_{k \rightarrow \infty} \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) \rightarrow \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)$, implies that, there exists an integer k_0 such that for any $k \geq k_0$, $\text{dist}(\tilde{\mathbf{U}}^k, \mathcal{Q}) < \delta$ and $\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) < \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) < \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) + r$. Then, we have, for any $k > k_0$,

$$\text{dist}(0, \partial \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k)) \phi'(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) \geq 1, \quad (45)$$

and $\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) > \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)$, where $\text{dist}(\mathbf{x}_0, \partial f(\mathbf{x})) := \inf\{\|\mathbf{x}_0 - \mathbf{y}\| : \mathbf{y} \in \partial f(\mathbf{x})\}$.

Using an approach similar to [17, Lemma 3.2], we have, there exists a constant $\tau_0 > 0$ such that

$$\begin{aligned} \text{dist}(0, \partial \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k)) &\leq \tau_0 (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F \\ &\quad + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F \\ &\quad + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F), \end{aligned} \quad (46)$$

which, in conjunction with (45), implies that

$$\begin{aligned} \frac{1}{\phi'(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*))} &\leq \tau_0 (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F \\ &\quad + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F \\ &\quad + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F). \end{aligned} \quad (47)$$

Based on (39), we have

$$\begin{aligned} \sigma_* \left(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \right. \\ \left. + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2 \right) &\leq \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) \\ &\stackrel{(i)}{\leq} \frac{\phi(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) - \phi(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*))}{\phi'(\hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \hat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*))} \end{aligned} \quad (48)$$

where (i) follows from the concavity of the function $\phi(\cdot)$. To simplify notations, we define

$$D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1}) = \phi(\widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) - \phi(\widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)).$$

Then, combining (47) and (48) yields

$$\begin{aligned} & \sigma_* \left(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 \right. \\ & \quad \left. + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F^2 \right) \leq \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) \\ & \leq \tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1}) (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F \\ & \quad + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F \\ & \quad + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F), \end{aligned}$$

which, using $\sqrt{n}(\sum_{i=1}^n \mathbf{A}_i\|_F^2)^{1/2} \geq \sum_{i=1}^n \mathbf{A}_i\|_F$, yields

$$\begin{aligned} & 3(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) \leq \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^k) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k+1}) \\ & \leq 6\sqrt{\frac{\tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1})}{\sigma_*}} (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F \\ & \quad + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F \\ & \quad + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F)^{1/2}, \end{aligned}$$

which, in conjunction with the inequality $2ab \leq a^2 + b^2$, yields

$$\begin{aligned} & 3(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) \leq \frac{9\tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1})}{\sigma_*} \\ & \quad + \|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F \\ & \quad + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F. \end{aligned}$$

Telescoping the above inequality over k from k_0 to K yields

$$\begin{aligned} & 3 \sum_{k=k_0}^K (\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) \leq \sum_{k=k_0}^K \frac{9\tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1})}{\sigma_*} \\ & \quad + \sum_{k=k_0}^K (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F \\ & \quad + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F). \end{aligned}$$

Rearranging the terms in the above inequality, we obtain

$$\begin{aligned} & \sum_{k=k_0}^K (2\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + 2\|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) \leq \sum_{k=k_0}^K \frac{9\tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1})}{\sigma_*} + 2\|\mathbf{X}^{k_0} - \mathbf{X}^{k_0-1}\|_F \\ & \quad + \|\mathbf{X}^{k_0-1} - \mathbf{X}^{k_0-2}\|_F + \|\mathbf{P}^{k_0} - \mathbf{P}^{k_0-1}\|_F + \|\mathbf{Q}^{k_0} - \mathbf{Q}^{k_0-1}\|_F \\ & \quad + 2\|\mathbf{Y}^{k_0} - \mathbf{Y}^{k_0-1}\|_F + \|\mathbf{Y}^{k_0-1} - \mathbf{Y}^{k_0-2}\|_F, \end{aligned} \quad (49)$$

Note that

$$\begin{aligned} & \sum_{k=k_0}^K \frac{9\tau_0 D(\tilde{\mathbf{U}}^k, \tilde{\mathbf{U}}^{k+1})}{\sigma_*} \\ & = \frac{9\tau_0}{\sigma_*} (\phi(\widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k_0}) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) - \phi(\widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{K+1}) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*))) \\ & \stackrel{(i)}{<} \frac{9\tau_0}{\sigma_*} \phi(\widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{k_0}) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*)) \end{aligned} \quad (50)$$

where (i) follows from the fact that $r > \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^{K+1}) - \widehat{\mathcal{L}}_\eta(\tilde{\mathbf{U}}^*) \geq 0$ and $\phi(\mathbf{x}) \geq 0$ for $0 < \mathbf{x} < r$ (Definition of function ϕ before). Combining (49), (50) and letting $K \rightarrow \infty$, we obtain

$$\begin{aligned} & \sum_{k=0}^{\infty} (\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) < \infty, \end{aligned}$$

which, in conjunction with Lemma 2, also implies that $\sum_{k=0}^{\infty} (\|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F + \|\mathbf{V}^{k+1} - \mathbf{V}^k\|_F) < \infty$. Combining the above two facts, we have $\sum_{k=0}^{\infty} \|\mathbf{U}^{k+1} - \mathbf{U}^k\|_F < \infty$, which means that the sequence (\mathbf{U}^k) is convergent.

C. Proof of Theorem 2

Proof. Note that $\widehat{\mathbf{X}}^{k+1}$ can be regarded as the minimizer of the following subproblem

$$\widehat{\mathbf{X}}^{k+1} = \arg \min_{\mathbf{X}} (f(\text{vec}(\mathbf{X}^T)) - (\mathbf{t}_x^{k+1})^T \text{vec}(\mathbf{X}^T)), \quad (51)$$

where the function $f(\cdot)$ is given by (14). Then, using a similar approach to (28), we can obtain

$$\begin{aligned} \text{vec}(\mathbf{A}^{k+1} \mathbf{E}_x^T) & = (\mathbf{G}_y^k + (\alpha + 1)\mathbf{I}_{nd}) \text{vec}((\widehat{\mathbf{X}}^{k+1})^T) - (\mathbf{b}_y^k)^T \\ & \quad - \mathbf{t}_x^{k+1} - \text{vec}((\widehat{\mathbf{X}}^k)^T). \end{aligned} \quad (52)$$

Similarly to (26) and based on (52), we can obtain

$$\begin{aligned} & \|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F^2 \\ & \leq \frac{2}{\sigma_x} \|\widehat{\mathbf{X}}^k - \widehat{\mathbf{X}}^{k-1}\|_F^2 + \frac{2(M_x + 1)L_y}{\sigma_x} \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F^2 \\ & \quad + \frac{2L_y + 2(\alpha + 1)^2}{\sigma_x} \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F^2 \\ & \quad + \frac{2}{\sigma_x} (\|\mathbf{t}_x^k\|_2^2 + \|\mathbf{t}_x^{k+1}\|_2^2). \end{aligned} \quad (53)$$

A symmetric result holds for $\|\mathbf{V}^{k+1} - \mathbf{V}^k\|_F^2$. Using a similar approach to (34) and recalling (51), we have

$$\begin{aligned} & \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \widehat{\mathbf{X}}^{k+1}, \widehat{\mathbf{Y}}^k, \mathbf{A}^k, \mathbf{V}^k) \\ & \quad - \mathcal{L}_\eta(\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}, \mathbf{X}^k, \mathbf{Y}^k, \mathbf{A}^k, \mathbf{V}^k) \\ & \leq -\frac{1}{2} \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F^2 + \mathbf{t}_x^{k+1} (\text{vec}(\widehat{\mathbf{X}}^{k+1}) - \text{vec}(\widehat{\mathbf{X}}^k)) \\ & \stackrel{(i)}{\leq} -\frac{1}{4} \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F^2 + \frac{1}{2} \|\mathbf{t}_x^{k+1}\|_2^2, \end{aligned} \quad (54)$$

where (i) follows from the fact that $\|\mathbf{A} + \mathbf{B}\|_F^2 \leq 2(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$. A symmetrical result holds for $\widehat{\mathbf{Y}}^{k+1}$. Next, we show that the sequence $(\widehat{\mathbf{U}}^k)$ is bounded. The boundedness of $(\widehat{\mathbf{X}}^k, \widehat{\mathbf{Y}}^k)$ follows from Property 2. Similar to (40), using (52) and Property 1 yields the boundedness of $(\mathbf{A}^k, \mathbf{V}^k)$, which,

in conjunction with the last two equalities in (9), yields the boundedness of $(\mathbf{P}^k, \mathbf{Q}^k)$. These facts imply the boundedness of the sequence of $(\widehat{\mathbf{U}}^k)$.

Since the function $\widehat{\mathcal{L}}_\eta$ is continuous and coercive, by the boundedness of $(\widehat{\mathbf{U}}^k)$, we have, there exists $M_l > 0$ such that $|\widehat{\mathcal{L}}_\eta(\widehat{\mathbf{U}}^k)| < M_l$ for $\forall k$. Combining (54), Lemma 3 and Lemma 4, we have

$$\begin{aligned} & \left(\sigma_* - \frac{1}{4}\right) \sum_{k=1}^n \left(\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 \right. \\ & \quad \left. + \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F^2 + \|\widehat{\mathbf{Y}}^{k+1} - \widehat{\mathbf{Y}}^k\|_F^2 \right) \\ & \leq \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^1) - \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^{n+1}) + \frac{1}{2} \sum_{k=1}^{n+1} (\|\mathbf{t}_x^k\|_2^2 + \|\mathbf{t}_y^k\|_2^2) < \infty \end{aligned} \quad (55)$$

which, combined with (35), (36) and $\eta > 2\eta_1$, implies $\sigma_* > 1/4$ and hence

$$\begin{aligned} & \sum_{k=1}^{\infty} (\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F^2 + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F^2 + \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F^2 \\ & \quad + \|\widehat{\mathbf{Y}}^{k+1} - \widehat{\mathbf{Y}}^k\|_F^2) < \infty, \end{aligned} \quad (56)$$

Combining (53), (56) and the fact that $\sum_{k=1}^{\infty} \|\mathbf{t}_x^k\|_2^2 < \infty$ implies $\sum_{k=1}^{\infty} \|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F^2 < \infty$. Similarly, we can obtain $\sum_{k=1}^{\infty} \|\mathbf{V}^{k+1} - \mathbf{V}^k\|_F^2 < \infty$. Based on the above results, we conclude that $\lim_{k \rightarrow \infty} \|\widehat{\mathbf{U}}^{k+1} - \widehat{\mathbf{U}}^k\|_F = 0$. Let $\widehat{\mathbf{U}}^* = (\mathbf{P}^*, \mathbf{Q}^*, \widehat{\mathbf{X}}^*, \widehat{\mathbf{Y}}^*, \mathbf{A}^*, \mathbf{V}^*)$ be a limit point of $\{\widehat{\mathbf{U}}^k\}$. Using steps similar to those in the proof of Theorem 1 and using $\|\mathbf{t}_x^{k+1}\|_2 \rightarrow 0$ we have $\mathbf{0} \in \partial \mathcal{L}_\eta(\widehat{\mathbf{U}}^*)$ and $\widehat{\mathbf{U}}^*$ is a stationary point of \mathcal{L}_η .

Using an approach similar to (46), we obtain, there exists a positive constant such that

$$\begin{aligned} & \text{dist}(\mathbf{0}, \partial \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^k)) \leq \gamma_0 (\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F + \|\mathbf{Y}^k - \mathbf{Y}^{k-1}\|_F \\ & \quad + \|\mathbf{P}^k - \mathbf{P}^{k-1}\|_F + \|\mathbf{Q}^k - \mathbf{Q}^{k-1}\|_F + \|\mathbf{X}^{k-1} - \mathbf{X}^{k-2}\|_F \\ & \quad + \|\mathbf{Y}^{k-1} - \mathbf{Y}^{k-2}\|_F + \|\mathbf{t}_x^k\|_2 + \|\mathbf{t}_y^k\|_2), \end{aligned} \quad (57)$$

which, in conjunction with (45) and (55) and using an approach similar to (50), yields

$$\begin{aligned} & \sum_{k=k_0}^K (2\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + 2\|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \\ & \quad + \|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|_F) \leq \frac{9\gamma_0 \phi(\widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^{k_0}) - \widehat{\mathcal{L}}_\eta(\widetilde{\mathbf{U}}^*))}{\sigma_* - 1/4} \\ & \quad + 2\|\mathbf{X}^{k_0} - \mathbf{X}^{k_0-1}\|_F + \|\mathbf{X}^{k_0-1} - \mathbf{X}^{k_0-2}\|_F + \|\mathbf{P}^{k_0} - \mathbf{P}^{k_0-1}\|_F \\ & \quad + \|\mathbf{Q}^{k_0} - \mathbf{Q}^{k_0-1}\|_F + 2\|\mathbf{Y}^{k_0} - \mathbf{Y}^{k_0-1}\|_F + \|\mathbf{Y}^{k_0-1} - \mathbf{Y}^{k_0-2}\|_F, \\ & \quad + \sum_{k=k_0}^K (\|\mathbf{t}_x^k\|_2 + \|\mathbf{t}_y^k\|_2). \end{aligned}$$

Letting $K \rightarrow \infty$ in the above inequality and using the fact that $\sum_{k=0}^{\infty} (\|\mathbf{t}_x^k\|_2 + \|\mathbf{t}_y^k\|_2) < \infty$ yields

$$\begin{aligned} & \sum_{k=1}^{\infty} (\|\mathbf{P}^{k+1} - \mathbf{P}^k\|_F + \|\mathbf{Q}^{k+1} - \mathbf{Q}^k\|_F + \|\widehat{\mathbf{X}}^{k+1} - \widehat{\mathbf{X}}^k\|_F \\ & \quad + \|\widehat{\mathbf{Y}}^{k+1} - \widehat{\mathbf{Y}}^k\|_F) < \infty, \end{aligned} \quad (58)$$

which implies that $\sum_{k=1}^{\infty} \|\widehat{\mathbf{U}}^{k+1} - \widehat{\mathbf{U}}^k\|_F < \infty$.

In the meanwhile, if the sequence $(\|\widehat{\mathbf{U}}^{k+1} - \widehat{\mathbf{U}}^k\|_F)$ is non-increasing, then we have

$$2k \|\widehat{\mathbf{U}}^{2k+1} - \widehat{\mathbf{U}}^{2k}\|_F \leq 2 \sum_{i=k+1}^{2k} \|\widehat{\mathbf{U}}^{i+1} - \widehat{\mathbf{U}}^i\|_F \rightarrow 0$$

as $k \rightarrow \infty$. Thus, we have $\|\widehat{\mathbf{U}}^{k+1} - \widehat{\mathbf{U}}^k\|_F = o(1/k)$, which in conjunction with an argument similar to Proposition 2 in [19], implies that the best running convergence rate of Algorithm 1 is $o(1/k)$. \square

D. Proof of Proposition 2

For each i , there exists an integer $s_i \leq k_x$ such that $i \in \mathcal{G}_{s_i}^x$. Using the definition (19), we have $\|\mathbf{x}_j - \mathbf{x}_i\|_2 < 2C_x/K$ for any $j \in \mathcal{G}_{s_i}^x$ and $\|\mathbf{x}_j - \mathbf{x}_i\|_2 \geq 2C_x/K$ for any $j \in (\mathcal{G}_{s_i}^x)^c = \bigcup_{1 \leq t \leq k_x, t \neq s_i} \mathcal{G}_t^x$. Recalling the definition of K_i , we have $|(\mathcal{G}_{s_i}^x)^c| = n - K_i$. These facts imply

$$\begin{aligned} \sum_{j=1}^n \frac{\Upsilon(\mathbf{x}_i - \mathbf{x}_j)}{K_i(n - K_i)} &= \left(\sum_{j \in \mathcal{G}_{s_i}^x} + \sum_{j \in (\mathcal{G}_{s_i}^x)^c} \right) \frac{\Upsilon(\mathbf{x}_i - \mathbf{x}_j)}{K_i(n - K_i)} \\ &= \sum_{j \in (\mathcal{G}_{s_i}^x)^c} (K_i(n - K_i))^{-1} = K_i^{-1}, \end{aligned}$$

which implies

$$\sum_{i,j} w_{ij} \Upsilon(\mathbf{x}_i - \mathbf{x}_j) = \sum_{i=1}^n K_i^{-1} = k_x = |\mathcal{G}(\mathbf{X})| \quad (59)$$

where the last equality follows from the fact that $K_i = K_j$ for any $i, j \in \mathcal{G}_s^x, 1 \leq s \leq k_x$.

E. Proof of Theorem 3

Let $p(x)$ and $q(x)$ denote probability density functions (pdf) of a random variable. Recall the Kullback-Leibler divergence $D(p||q) \triangleq \mathbb{E}_p[\log(p/q)]$ and the Hellinger affinity $A(p, q) \triangleq \mathbb{E}_p[\sqrt{q/p}]$. We first quote the following lemma from [40, Lemma A.1].

Lemma 5. *Let \mathbf{M}^* be an $n \times m$ target matrix and let \mathcal{H} be a finite collection of candidate estimates \mathbf{H} of \mathbf{M}^* , each of which has a penalty $\text{pen}(\mathbf{H}) \geq 1$ such that $\sum_{\mathbf{H} \in \mathcal{H}} 2^{-\text{pen}(\mathbf{H})} \leq 1$. Assume each $(i, j) \in [n] \times [m]$ is included in the observed index set Ω with probability $\gamma = s/(nm)$ for certain integer $s < nm$ and the joint pdf of the observations $\Phi_\Omega(\mathbf{M})$ follows $p_{\mathbf{M}^*}(\Phi_\Omega(\mathbf{M})) = \prod_{(i,j) \in \Omega} p_{M_{ij}^*}(M_{ij})$, which are independent conditioned on Ω . Given a constant C_D satisfying*

$$C_D \geq \max_{\mathbf{H} \in \mathcal{H}} \max_{(i,j) \in [n] \times [m]} D(p_{M_{ij}^*}(M_{ij}) || p_{H_{ij}}(M_{ij})), \quad (60)$$

we have that for any $\xi \geq (1 + 2C_D/3) 2 \log 2$, the complexity regularized maximum likelihood estimator

$$\widehat{\mathbf{M}}(\Omega, \Phi_\Omega(\mathbf{M})) = \arg \min_{\mathbf{H} \in \mathcal{H}} \{-\log p_{\mathbf{H}}(\Phi_\Omega(\mathbf{M})) + \xi \text{pen}(\mathbf{H})\}$$

satisfies the estimation error bound

$$\begin{aligned} & \mathbb{E} \left[-2 \log A(p_{\widehat{\mathbf{M}}}(\mathbf{M}), p_{\mathbf{M}^*}(\mathbf{M})) \right] / (nm) \\ & \leq 8C_D \log s / s + 3 \min_{\mathbf{H} \in \mathcal{H}} \left\{ D(p_{\mathbf{M}^*}(\mathbf{M}) \| p_{\mathbf{H}}(\mathbf{M})) / (nm) \right. \\ & \quad \left. + (\xi + 4C_D \log 2/3) \text{pen}(\mathbf{H}) / s \right\}, \end{aligned} \quad (61)$$

where the expectation is with respect to the joint distribution of $(\Omega, \Phi_{\Omega}(\mathbf{M}))$ and the shorthands

$$\begin{aligned} A(p_{\widehat{\mathbf{M}}}(\mathbf{M}), p_{\mathbf{M}^*}(\mathbf{M})) &= \prod_{i,j} A(p_{\widehat{M}_{ij}}(M_{ij}), p_{M_{ij}^*}(M_{ij})), \\ D(p_{\mathbf{M}^*}(\mathbf{M}) \| p_{\mathbf{H}}(\mathbf{M})) &= \sum_{i,j} D(p_{M_{ij}^*}(M_{ij}) \| p_{H_{ij}}(M_{ij})). \end{aligned} \quad (62)$$

In order to apply this lemma, we need to construct penalties $\text{pen}(\mathbf{H})$ for each $\mathbf{H} \in \mathcal{H}$ defined in (18) such that $\sum_{\mathbf{H} \in \mathcal{H}} 2^{-\text{pen}(\mathbf{H})} \leq 1$. Let $\mathcal{H}_1 \triangleq \{\mathbf{H} = \mathbf{X}\mathbf{Y}^T : \mathbf{X} \in \mathcal{X}, \mathbf{Y} \in \mathcal{Y}\}$ with \mathcal{X} and \mathcal{Y} defined in (19). Based on [41], the inequality $\sum_{\mathbf{H} \in \mathcal{H}_1} 2^{-\text{pen}(\mathbf{H})} \leq 1$ can be satisfied by choosing $\text{pen}(\mathbf{H})$ to be the code length of some uniquely decodable binary code. Thus, we use the following three steps to design $\text{pen}(\mathbf{H})$. Recall $K = 2^{\lceil \mu \log_2(n \vee m) \rceil}$ and let $n_0 = n \vee m$.

- 1) First encode the number of subgroups of $\{\mathbf{x}_i, 1 \leq i \leq n\}$, i.e. $|\mathcal{G}(\mathbf{X})|$, with $\lceil \log_2(n) \rceil$ bits. Similarly, encode $|\mathcal{G}(\mathbf{Y})|$ with $\lceil \log_2(m) \rceil$ bits.
- 2) For each $\mathbf{x}_i, 1 \leq i \leq n$, encode the index of the subgroup that \mathbf{x}_i belongs to with $\lceil \log_2(n) \rceil$ bits. Similarly, for each $\mathbf{y}_j, 1 \leq j \leq m$, encode the index of the subgroup that \mathbf{y}_j belongs to with $\lceil \log_2(m) \rceil$ bits. Thus, the total number of bits used in this step is $n \lceil \log_2(n) \rceil + m \lceil \log_2(m) \rceil$.
- 3) Finally, using (19), we encode X_{ij} with $\log_2(K)$ bits. Similarly, encode Y_{ij} with $\log_2(K)$ bits. Since all row vectors in a subgroup are identical, the total number of bits used in this step is $d \log_2(K) (|\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|)$.
- 4) We assign each $\mathbf{H} \in \mathcal{H}_1$ with a code that is concatenation of the code for \mathbf{X} followed by the code for \mathbf{Y} . Thus, the penalty

$$\begin{aligned} \text{pen}(\mathbf{H}) &= (n+1) \lceil \log_2(n) \rceil + (m+1) \lceil \log_2(m) \rceil \\ & \quad + d \log_2(K) (|\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|). \end{aligned} \quad (63)$$

Since such codes are uniquely decodable, we can obtain that $\sum_{\mathbf{H} \in \mathcal{H}_1} 2^{-\text{pen}(\mathbf{H})} \leq 1$.

Since $\mathcal{H} \subseteq \mathcal{H}_1$, $\sum_{\mathbf{H} \in \mathcal{H}} 2^{-\text{pen}(\mathbf{H})} \leq \sum_{\mathbf{H} \in \mathcal{H}_1} 2^{-\text{pen}(\mathbf{H})} \leq 1$. Then, replacing $\text{pen}(\mathbf{H})$ in Lemma 5 by (63) implies

$$\begin{aligned} \widehat{\mathbf{M}} &= \arg \min_{\mathbf{H} \in \mathcal{H}} \left\{ -\log p_{\mathbf{H}}(\Phi_{\Omega}(\mathbf{M})) \right. \\ & \quad \left. + \xi d \log_2(K) (|\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|) \right\}, \end{aligned} \quad (64)$$

which, based on the fact that $\max\{\lceil \log_2(n) \rceil, \lceil \log_2(m) \rceil\} \leq \log_2(K) = \lceil \mu \log_2(n_0) \rceil$, satisfies

$$\begin{aligned} & \mathbb{E} \left[-2 \log A(p_{\widehat{\mathbf{M}}}(\mathbf{M}), p_{\mathbf{M}^*}(\mathbf{M})) \right] / (nm) \\ & \leq \frac{8C_D \log s}{s} + 3 \min_{\mathbf{H} \in \mathcal{H}} \left\{ \frac{D(p_{\mathbf{M}^*}(\mathbf{M}) \| p_{\mathbf{H}}(\mathbf{M}))}{nm} + \left(\xi + \frac{4C_D \log 2}{3} \right) \right. \\ & \quad \left. d \log_2(K) \times \frac{(m+n+2)d^{-1} + |\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|}{s} \right\}. \end{aligned} \quad (65)$$

Recall that $\xi \geq (1 + 2C_D/3) 2 \log 2$. Using the fact that $\log_2(K) = \lceil \mu \log_2(n_0) \rceil \leq 2\mu \log(n_0) / \log 2$ and letting $\lambda = 2\sigma^2 \xi d \log_2(K)$, for any

$$\lambda \geq 8\sigma^2 \mu d (1 + 2C_D/3) \log(n_0), \quad (66)$$

the estimate of

$$\begin{aligned} \widehat{\mathbf{M}}(\Omega, \Phi_{\Omega}(\mathbf{M})) &= \arg \min_{\mathbf{H} \in \mathcal{H}} \left\{ -\log p_{\mathbf{H}}(\Phi_{\Omega}(\mathbf{M})) \right. \\ & \quad \left. + \frac{\lambda}{2\sigma^2} (|\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|) \right\} \end{aligned} \quad (67)$$

satisfies

$$\begin{aligned} & \mathbb{E} \left[-2 \log A(p_{\widehat{\mathbf{M}}}(\mathbf{M}), p_{\mathbf{M}^*}(\mathbf{M})) \right] / (nm) \\ & \leq 8C_D \frac{\log s}{s} + 3 \min_{\mathbf{H} \in \mathcal{H}} \left\{ D(p_{\mathbf{M}^*}(\mathbf{M}) \| p_{\mathbf{H}}(\mathbf{M})) / (nm) + \left(\frac{\lambda}{2\sigma^2} + \frac{8\mu d C_D \log(n_0)}{3} \right) \times \frac{(m+n+2)/d + |\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|}{s} \right\}. \end{aligned} \quad (68)$$

Recall that the noise matrix \mathbf{N} contains i.i.d. $\mathcal{N}(0, \sigma^2)$ elements and the pdf of the observations $\Phi_{\Omega}(\mathbf{M})$ is

$$p_{\mathbf{H}}(\Phi_{\Omega}(\mathbf{M})) = \frac{1}{(2\pi\sigma^2)^{|\Omega|/2}} \exp \left(-\frac{\|\Phi_{\Omega}(\mathbf{M} - \mathbf{H})\|_F^2}{2\sigma^2} \right),$$

we have (17) is of the same form as (67). Furthermore, we have

$$\begin{aligned} & D(p_{M_{ij}^*}(M_{ij}) \| p_{H_{ij}}(M_{ij})) \\ &= \int_{-\infty}^{\infty} \frac{H_{ij}^2 - (M_{ij}^*)^2 + 2(M_{ij}^* - H_{ij})x}{2\sigma^2} p_{M_{ij}^*}(x) dx \\ &= \frac{1}{2\sigma^2} \left(H_{ij}^2 - (M_{ij}^*)^2 + 2(M_{ij}^* - H_{ij}) \int_{-\infty}^{\infty} x p_{M_{ij}^*}(x) dx \right) \\ &\stackrel{(i)}{=} 2\sigma^{-2} (H_{ij} - M_{ij}^*)^2, \end{aligned} \quad (69)$$

where (i) follows from the fact that $\int_{-\infty}^{\infty} x p_{M_{ij}^*}(x) dx = M_{ij}^*$. Recall (60) and note that $\|\mathbf{H}\|_{\infty} \leq C_m$. Then picking $C_D = 2C_m^2/\sigma^2$ in (66), we have

$$\lambda \geq 8\mu d (\sigma^2 + 4C_m^2/3) \log(n_0). \quad (70)$$

Using a similar approach, we obtain

$$-\log A(p_{M_{ij}^*}(M_{ij}), p_{H_{ij}}(M_{ij})) = (H_{ij} - M_{ij}^*)^2 / (8\sigma^2),$$

which, in conjunction with (69) and (68), implies

$$\begin{aligned} & \mathbb{E} \left[\|\widehat{\mathbf{M}} - \mathbf{M}^*\|_F^2 \right] / (nm) \\ & \leq \frac{64C_m^2 \log s}{s} + 6 \min_{\mathbf{H} \in \mathcal{H}} \left\{ \frac{\|\mathbf{H} - \mathbf{M}^*\|_F^2}{nm} + \left(\lambda + \frac{32\mu d C_m^2 \log(n_0)}{3} \right) \right. \\ & \quad \left. \times \frac{(m+n+2)d^{-1} + |\mathcal{G}(\mathbf{X})| + |\mathcal{G}(\mathbf{Y})|}{s} \right\}. \end{aligned} \quad (71)$$

Next, we prove an upper bound for $\min_{\mathbf{H} \in \mathcal{H}} \|\mathbf{H} - \mathbf{M}^*\|_F^2$. Let $\mathbf{H}_0 = \mathbf{X}_0 \mathbf{Y}_0^T \in \mathcal{M}$ be a candidate reconstruction such that the entries of \mathbf{X}_0 and \mathbf{Y}_0 are the closest discretized surrogates of the entries of \mathbf{X}^* and \mathbf{Y}^* , respectively. Recalling (18) and (19), we have, $\|\mathbf{X}_0 - \mathbf{X}^*\|_F^2 \leq 4dnC_x^2/K^2$ and $\|\mathbf{Y}_0 - \mathbf{Y}^*\|_F^2 \leq 4dmC_y^2/K^2$, which implies

$$\begin{aligned} \|\mathbf{H}_0 - \mathbf{M}^*\|_F &= \|\mathbf{X}_0 \mathbf{Y}_0^T - \mathbf{X}^* (\mathbf{Y}^*)^T\|_F \\ &\leq \|\mathbf{X}_0 - \mathbf{X}^*\|_F \|\mathbf{Y}_0\|_F + \|\mathbf{X}^*\|_F \|\mathbf{Y}_0 - \mathbf{Y}^*\|_F \\ &\leq \frac{4dC_x C_y \sqrt{nm}}{K} \leq \frac{4dC_x C_y \sqrt{nm}}{\mu \log(n_0)}, \end{aligned}$$

and hence

$$\|\mathbf{H}_0 - \mathbf{M}^*\|_F^2 / (nm) \leq 16dC_x^2C_y^2 / (\mu \log(n_0))^2,$$

which, in conjunction with the fact that $|\mathcal{G}(\mathbf{X}_0)| + |\mathcal{G}(\mathbf{Y}_0)| = |\mathcal{G}(\mathbf{X}^*)| + |\mathcal{G}(\mathbf{Y}^*)|$, implies

$$\begin{aligned} & \mathbb{E} \left[\|\widehat{\mathbf{M}} - \mathbf{M}^*\|_F^2 \right] / (nm) \\ & \leq \frac{64C_m^2 \log s}{s} + \frac{96dC_x^2C_y^2}{(\mu \log(n_0))^2} + (6\lambda + 64\mu C_m^2 \log(n_0)) \times \\ & \quad \frac{(m+n+2) + (|\mathcal{G}(\mathbf{X}^*)| + |\mathcal{G}(\mathbf{Y}^*)|)d}{s}. \end{aligned} \quad (72)$$

Letting $C_1 = 64C_m^2$, $C_2 = 96C_x^2C_y^2$, $C_3 = 64C_m^2$ and using (70), (72) finishes the proof.