Journal of Machine Learning Research 25 (2024) 1-76 Submitted 3/24; Revised 10/24; Published 11/24

Triple Component Matrix Factorization: Untangling Global, Local, and Noisy Components

Naichen Shi NAICHENS@UMICH.EDU
Department of Industrial & Operations Engineering
University of Michigan
Ann Arbor, MI 48109, USA

Salar Fattahi FATTAHI@UMICH.EDU
Department of Industrial & Operations Engineering
University of Michigan
Ann Arbor, MI 48109, USA

Raed Al Kontar * ALKONTAR@UMICH.EDU
Department of Industrial & Operations Engineering
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Mahdi Soltanolkotabi

Abstract

In this work, we study the problem of common and unique feature extraction from noisy data. When we have N observation matrices from N different and associated sources corrupted by sparse and potentially gross noise, can we recover the common and unique components from these noisy observations? This is a challenging task as the number of parameters to estimate is approximately thrice the number of observations. Despite the difficulty, we propose an intuitive alternating minimization algorithm called triple component matrix factorization (`TCMF`) to recover the three components exactly. `TCMF` is distinguished from existing works in literature thanks to two salient features. First, `TCMF` is a principled method to separate the three components given noisy observations provably. Second, the bulk of the computation in `TCMF` can be distributed. On the technical side, we formulate the problem as a constrained nonconvex nonsmooth optimization problem. Despite the intricate nature of the problem, we provide a Taylor series characterization of its solution by solving the corresponding Karush–Kuhn–Tucker conditions. Using this characterization, we can show that the alternating minimization algorithm makes significant progress at each iteration and converges into the ground truth at a linear rate. Numerical experiments in video segmentation and anomaly detection highlight the superior feature extraction abilities of `TCMF`.

Keywords: Matrix Factorization, Heterogeneity, Alternating minimization, Sparse noise, Outlier identification

1. Introduction

In the era of Big Data, an important task is to find low-rank features from high-dimensional observations. Methods including principal component analysis (Hotelling, 1933), low-rank

∗. Corresponding author

matrix factorization (Koren et al., 2009), and dictionary learning (Aharon et al., 2006), have found success in numerous fields of statistics and machine learning (Wright and Ma, 2022). Among them, matrix factorization (MF) is an efficient method to identify the features that best explain the observation matrices.

Despite the wide popularity, standard MF methods are known to be brittle in the presence of outliers with huge noise (Candès et al., 2011). These noises are often sparse but can have large norms. A series of methods (e.g., (Candès et al., 2011; Netrapalli et al., 2014; Wong and Lee, 2017; Fattahi and Sojoudi, 2020; Chen et al., 2021)) have been developed to estimate low-rank features from data that contain outliers. When the portion of outliers is not too large, one can *provably* identify the outliers and the low-rank components with convex programming (Candès et al., 2011) or nonconvex optimization algorithms equipped with convergence guarantees (Netrapalli et al., 2014).

Recently, there has been a growing number of applications where data are acquired from diverse but connected sources, such as smartphones, car sensors, or medical records from different patients. This type of data displays both mutual and individual characteristics. For instance, in biostatistics, the measurements of different miRNA and gene expressions from the same set of samples can reveal co-varying patterns yet exhibit heterogeneous trends (Lock et al., 2013). Statistical modeling of the common information among all data sources and the specific information for each source is of central interest in these applications. Multiple works propose to recover such common and unique features by minimizing the square norm of the residuals of fitting (Lock et al., 2013; Zhou et al., 2015; Gaynanova and Li, 2019; Park and Lock, 2020; Lee and Choi, 2009; Yang and Michailidis, 2016; Shi and Kontar, 2024; Shi et al., 2023; Liang et al., 2023). These methods prove to be useful in aligning genetics features (Lock et al., 2013), visualizing bone and soft tissues in X-ray images (Zhou et al., 2015), functional magnetic resonance imaging (Kashyap et al., 2019), surveillance video segmentation (Shi and Kontar, 2024), stocks market analysis (Shi et al., 2023), and many more.

Though these algorithms achieve decent performance on multiple applications, they rely on least square estimates, which are not robust to outliers in data. Real-world data are commonly corrupted by outliers (Tan et al., 2005). Factors including measurement errors or sensor malfunctions can give rise to large noise in data. These outliers can substantially skew the estimation of low-rank features. As such, we attempt to answer the following question.

Question: How can one *provably* identify low-rank common and unique information robustly from data corrupted by outlier noise?

A natural thought is to borrow techniques in robust PCA to handle outlier noise. Indeed, there exist a few heuristic methods in literature (Sagonas et al., 2017; Panagakis et al., 2015; Ponzi et al., 2021) to find robust estimates of shared and unique features. These methods often use ℓ_1 regularization (Sagonas et al., 2017; Panagakis et al., 2015) or Huber loss (Ponzi et al., 2021) to accommodate the sparsity of noise. However, these algorithms are mainly based on heuristics and lack theoretical guarantees, thus potentially compromising the quality of their outputs. A theoretically justifiable method to identify low-rank shared and unique components from outlier noise is still lacking. In this paper, we will study the question rigorously and develop an efficient algorithm to solve it.

2. Problem Statement

We consider the framework where N observation matrices $\mathbf{M}_{(1)}, \mathbf{M}_{(2)}, \cdots, \mathbf{M}_{(N)}$ come from $N \in \mathbb{N}^+$ different but associated sources. These matrices $\mathbf{M}_{(i)} \in \mathbb{R}^{n_1 \times n_{2,(i)}}$ have the same number of features n_1. To model their commonality and uniqueness, we assume each matrix is driven by r_1 shared factors and $r_{2,(i)}$ unique factors and contaminated by potentially gross noise. More specifically, we consider the model where the observation $\mathbf{M}_{(i)}$ from source i is generated by,

$$\mathbf{M}_{(i)} = \mathbf{U}^\star{}_g \mathbf{V}^{\star T}_{(i),g} + \mathbf{U}^\star{}_{(i),l} \mathbf{V}^{\star T}_{(i),l} + \mathbf{S}^\star{}_{(i)}, \tag{1}$$

where $\mathbf{U}^\star{}_g \in \mathbb{R}^{n_1 \times r_1}$, $\mathbf{V}^\star{}_{(i),g} \in \mathbb{R}^{n_{2,(i)} \times r_1}$, $\mathbf{U}^\star{}_{(i),l} \in \mathbb{R}^{n_1 \times r_{2,(i)}}$, $\mathbf{V}^\star{}_{(i),l} \in \mathbb{R}^{n_{2,(i)} \times r_{2,(i)}}$, $\mathbf{S}^\star{}_{(i)} \in \mathbb{R}^{n_1 \times n_{2,(i)}}$. We use * to denote the ground truth. r_1 is the rank of global (shared) feature matrices, and $r_{2,(i)}$ is the rank of local (unique) feature matrix from source i. The matrix $\mathbf{U}^\star{}_g \mathbf{V}^{\star T}_{(i),g}$ models the shared low-rank part of the observation matrix, as the column space is the same across different sources. $\mathbf{U}^\star{}_{(i),l} \mathbf{V}^{\star T}_{(i),l}$ models the unique low-rank part. $\mathbf{S}^\star{}_{(i)}$ models the noise from source i.

In matrix factorization problems, the representations $\mathbf{U}^\star$ and $\mathbf{V}^\star$ often correspond to latent data features. For instance, in recommender systems, $\mathbf{U}^\star$ can be interpreted as user features that reveal their preferences on different items in the latent space (Koren et al., 2009). For better interpretability, it is often desirable to have the underlying features disentangled so that each feature can vary independently of others (Higgins et al., 2017). Under this rationale, we consider the model where shared and unique factors are orthogonal,

$$\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_{(i),l} = 0, \ \forall i \in [N], \tag{2}$$

where $[N]$ denotes the set $\{1, 2, \cdots, N\}$. The orthogonality of features implies that the shared and unique features span different subspaces, thus describing different patterns in the observation. The orthogonal condition (2) is thus an inductive bias that reflects our prior belief about the independence between common and unique factors and naturally models a diverse range of applications, such as miRNA and gene expression (Lock et al., 2013), human faces (Zhou et al., 2015), and many more (Sagonas et al., 2017; Shi and Kontar, 2024).

We should note that the orthogonality (2) does not limit the model representation power. Suppose $\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_{(i),l} \neq 0$ otherwise, we can decompose $\mathbf{U}^\star{}_{(i),l}$ into the two parts, $\mathbf{U}^\star{}_{(i),l} = \mathbf{U}^\star{}_g \left(\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_g\right)^{-1} \mathbf{U}^{\star T}_g \mathbf{U}^\star{}_{(i),l} + \left(\mathbf{I} - \mathbf{U}^\star{}_g \left(\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_g\right)^{-1} \mathbf{U}^{\star T}_g\right) \mathbf{U}^\star{}_{(i),l}$. The first part is in the column subspace of $\mathbf{U}^\star{}_g$, while the second part is in the orthogonal space of the column subspace of $\mathbf{U}^\star{}_g$. If we define $\widetilde{\mathbf{U}^\star}_{(i),l} = \left(\mathbf{I} - \mathbf{U}^\star{}_g \left(\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_g\right)^{-1} \mathbf{U}^{\star T}_g\right) \mathbf{U}^\star{}_{(i),l}$, and $\widetilde{\mathbf{V}^\star}_{(i),g} = \mathbf{V}^\star{}_{(i),g} + \mathbf{V}^\star{}_{(i),l} \mathbf{U}^{\star T}_{(i),l} \mathbf{U}^\star{}_g \left(\mathbf{U}^{\star T}_g \mathbf{U}^\star{}_g\right)^{-1}$, we have, $\mathbf{M}_{(i)} = \mathbf{U}^\star{}_g \widetilde{\mathbf{V}^\star}^T_{(i),g} + \widetilde{\mathbf{U}^\star}_{(i),l} \mathbf{V}^{\star T}_{(i),l} + \mathbf{S}^\star{}_{(i)}$ where $\mathbf{U}^{\star T}_g \widetilde{\mathbf{U}^\star}_{(i),l} = 0$. This formulation admits the form of model (1) with constraint (2).

The noise term $\mathbf{S}^\star{}_{(i)}$ in (1) models the sparse and large noise, where only a small fraction of $\mathbf{S}^\star{}_{(i)}$ registers as nonzero. The noise sparsity is extensively invoked in literature, particularly when datasets are plagued by outliers (Candès et al., 2011; Netrapalli et al., 2014; Chen et al., 2020, 2021).

2.1 Challenges

Given data generation model (1), our task is to separate common, individual, and noise components. The task seems Herculean as the problem is under-definite: we need to estimate three sets of parameters from one set of observations. There are two major challenges associated with the problem,

Challenge 1: New identifiability conditions are needed. Standard analysis in robust PCA (Candès et al., 2011; Netrapalli et al., 2014) often uses the incoherence condition to distinguish low-rank components from sparse noise. However, the incoherence condition alone is insufficient to guarantee the separability between common and unique features. Since there are infinitely many ways in which shared, unique, and noise components can form the observation matrices, it is not apparent whether untangling them is even feasible. Thus, the crux of our investigation is to understand when the separation is possible.

Fortunately, we show that a group of conditions–known here as identifiability conditions–exists that can ensure the precise retrieval of the shared, unique, and sparse noise. Intuitively, these identifiability conditions require the three components to have "little overlaps".

Based on these conditions, we will develop an alternating minimization algorithm called `TCMF` to iteratively update the three components. An illustration of the algorithm is shown in the left graph in Figure 1. The hard-thresholding step finds the closest sparse matrix for the data noise. We use `JIMF` to denote a subroutine that represents a group of algorithms (e.g., (Lock et al., 2013; Shi and Kontar, 2024)) to identify common and unique low-rank features. In essence, `JIMF` solves a sub-problem in `TCMF`. It is worth noting that there exist multiple algorithms in literature to implement `JIMF`, many of which can produce high-quality outputs. With the implemented `JIMF`, `TCMF` applies hard thresholding and `JIMF` alternatively to estimate the sparse, as well as common and unique low-rank components. The left graph of Figure 1 offers an intuitive understanding of how estimates of various components progress toward the ground truth with each iterative step.

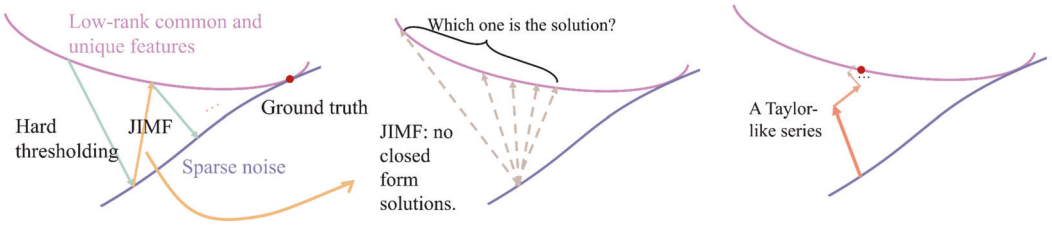


Figure 1: *Left*: An illustration of `TCMF`'s update trajectory. The purple and blue curves represent the spaces for the low-rank and sparse matrices. The algorithm alternatively performs hard thresholding and `JIMF`, making the updates closer and closer to the ground truth. *Middle*: An illustration showing why insufficient understanding about the output of `JIMF` can be problematic in the convergence analysis. *Right*: Our contribution to represent the solution into a Taylor-like series.

Challenge 2: New analysis tools are needed. Showing the exact recovery of low-rank and sparse components is not easy. Even in standard robust PCA, one needs to apply highly nontrivial analytical techniques to provide theoretical guarantees. For example, Robust PCA

(Candès et al., 2011) relies on a "golfing scheme" to construct dual variables that ensure the uniqueness of a convex optimization problem. Nonconvex robust PCA (Netrapalli et al., 2014) applies a perturbation analysis of SVD to quantify the improvement of the algorithm per iteration. These techniques are tailored for standard robust PCA and cannot be directly extended to the case where both common and unique features are involved, which increases the complexity of the analysis. The major difficulty stems from the fact that `TCMF` updates the low-rank common and unique components by another iterative algorithm `JIMF`. Unlike robust PCA, the output of `JIMF` does not have a closed-form formula. This conceptual hurdle is illustrated in the middle graph of Figure 1. As a result, novel analysis tools are needed to justify the convergence of the proposed `TCMF`.

One of our key contributions in tackling the challenge is to develop innovative analysis tools by solving the Karush–Kuhn–Tucker conditions of the objective of `JIMF` and express the solutions into a Taylor-like series. From the Taylor-like series, we can precisely characterize the output of `JIMF`, thereby showing the series converge to a close estimate of the ground truth shared and unique features.

The Taylor-like series is depicted in the right graph of Figure 1. Perhaps surprisingly, regardless of the choice of the subroutine `JIMF`, as long as `JIMF` finds a close estimate of the optimal solutions to a subproblem, its output can be represented by an infinite series. The series describes the optimal solution of the subproblem and is independent of the intermediate steps in `JIMF`. The derivation and analysis of Taylor-like series have stand-alone values in the theoretical research of the sensitivity analysis of matrix factorization. With the new analysis tool, we are able to show that even if the `JIMF` only outputs a reasonable approximate solution, the meta-algorithm `TCMF` can still take advantage of the information in such an inexact solution to refine the estimates of the three components. We will elaborate on the Taylor-like series in greater detail in Section 6 and the Appendix.

We summarize our contributions in the following.

2.2 Summary of Contributions

Identifiability conditions. We discover a group of identifiability conditions sufficient for the almost exact recovery of common, unique, and sparse components from noisy observation matrices. Essentially, the identifiability conditions require that the fraction of nonzero entries in the noise not be too large, the factor matrices be incoherent, and unique factors be misaligned. The first two conditions are needed even in the standard analysis of the robust PCA, while the third condition is essential for the disentanglement of the shared and unique components.

Efficient and distributed algorithm. We propose a constrained nonconvex nonsmooth matrix factorization problem to solve the shared, unique, and sparse components. Despite the nonconvexity of the problem, we design a meta-algorithm called **T**riple **C**omponent **M**atrix **F**actorization (`TCMF`) to solve the problem. Our approach is able to leverage a wide range of existing methods for separating the common and unique components to precision ϵ. Furthermore, `JIMF` can be distributed if the subroutine `JIMF` is distributed.

Convergence guarantee. We show that, under the identifiability conditions, our proposed `TCMF` has a convergence guarantee. To the best of our knowledge, such a guarantee is the first of its kind, as it ensures the recovery of common, unique, and noise components

to high precision. Our theoretical analysis introduces new techniques to solve the KKT conditions in Taylor-like series and bound each term in the series. It sheds light on the sensitivity analysis with the ℓ_∞ norm.

Case studies. We use a wide range of numerical experiments to demonstrate the application of `TCMF` in different case studies, as well as the effectiveness of our proposed method. Numerical experiments corroborate theoretical convergence results. Also, the case studies on video segmentation and anomaly detection showcase the benefits of untangling shared, unique, and noisy components.

In the rest of the paper, we provide a comprehensive review of the literature in Section 3. Then, we elaborate on the conditions sufficient for the separation of the three components in Section 4. In Section 5, we introduce the alternating minimization algorithm. We present our convergence theorem in Section 6 and discuss the key insights in the proof and how they solve challenge 2. In Section 7, we demonstrate the numerical experiment results. The detailed proofs are relegated to the Appendix for brevity of the main paper.

3. Related Work

Matrix Factorization There are numerous works that analyze the theoretical and practical properties of first-order algorithms that solve the (asymmetric) matrix factorization problem $\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{M}-\mathbf{U}\mathbf{V}^T\right\|_F^2$ or its variants (Li et al., 2018; Ye and Du, 2021; Sun and Luo, 2016; Park et al., 2017; Tu et al., 2016). Among them, Sun and Luo (2016) analyzes the local landscape of the optimization problem and establishes the local linear convergence of a series of first-order algorithms. Park et al. (2017); Ge et al. (2017) study the global geometry of the optimization problem. Tu et al. (2016) proposes the Rectangular Procrustes Flow algorithm that is proved to converge linearly into the ground truth under proper initialization and a balancing regularization. Recently, Ye and Du (2021) shows that gradient descent with small and random initialization can converge to the ground truth.

Robust PCA When the observation is corrupted by sparse and potentially large noise, several approaches can still identify the low-rank components. An exemplary work is Robust PCA (Candès et al., 2011), which proposes an elegant convex optimization problem called principal component pursuit that uses nuclear norm and ℓ_1 norm to promote the sparsity and low-rankness of the solutions. It is proved that under incoherence assumptions, the solution of the convex optimization is unique and corresponds to the ground truth. Several works also consider the problem of matrix completion under outlier noise (Wong and Lee, 2017; Chen et al., 2021). Nonconvex robust PCA (Netrapalli et al., 2014) improves the computational efficiency of principal component pursuit by proposing a nonconvex formulation and using an alternating projection algorithm to solve it. Though the formulation is nonconvex, the alternating projection algorithm is also proved to recover the ground truth exactly under incoherence and sparsity requirements. For the special case of rank-1 robust PCA, Fattahi and Sojoudi (2020) show that a simple sub-gradient method applies directly to the nonsmooth ℓ_1-loss provably recovers the low-rank component and sparse noise, under the same incoherence and sparsity requirements. To model a broader set of noise, Meng and De La Torre (2013) consider a mixture of Gaussian noise models and exploit the EM algorithm to estimate the low-rank components. Robust PCA has found successful applications in video segmentation (Bouwmans and Zahzah, 2014), image processing (Vaswani et al., 2018), change point

detection (Xiao et al., 2020), and many more. Nevertheless, the formulations of robust PCA focus on shared low-rank features among all data and neglect unique components.

Distributed matrix factorization The emergence of edge computation has prompted the research on distributed matrix factorization. Gemulla et al. (2011) exploits distributed gradient descent to factorize large matrices. Chai et al. (2021) proposes a cryptographic framework where multiple clients use their local data to collaboratively factorize a matrix without leaking private information to the server. These works use one set of feature matrices $\mathbf{U}$ and $\mathbf{V}$ to fit data from all clients, thus also neglecting the feature differences from different sources as well as the possible outliers in data. Our method `TCMF` is distributed when its subroutine `JIMF` is distributed. Different from conventional distributed matrix factorization, `TCMF` can find common and unique components simultaneously while remaining robust in the presence of outliers.

Joint and individual feature extraction The literature on using MF to identify shared and unique features abound (Lock et al., 2013; Zhou et al., 2015; Gaynanova and Li, 2019; Park and Lock, 2020; Lee and Choi, 2009; Yang and Michailidis, 2016; Shi and Kontar, 2024; Liang et al., 2023; Shi et al., 2023). Among them, JIVE (Lock et al., 2013), COBE (Zhou et al., 2015), PerPCA (Shi and Kontar, 2024), and HMF (Shi et al., 2023) uses mutually orthogonal features to model the shared and unique components. SLIDE (Gaynanova and Li, 2019) and BIDIFAC (Park and Lock, 2020) do not pose orthogonality constraints but use regularizations to encourage the unique features to have small norms. GNMF (Lee and Choi, 2009) and iNMF (Yang and Michailidis, 2016) further add nonnegativity constraints to the factor matrices. In particular, PerPCA (Shi and Kontar, 2024) and HMF (Shi et al., 2023) are two distributed algorithms that are guaranteed to converge to the optimal solutions under proper conditions. It is worth mentioning that these methods do not account for sparse noise in the observations. In this work, we remedy this challenge by utilizing existing methods as basic building blocks for our approach, which focuses on simultaneously separating shared and unique features as well as noise components.

Robust shared and unique feature extraction As discussed, a few heuristic methods also attempt to find the shared and unique features when data are corrupted by large noise (Sagonas et al., 2017; Ponzi et al., 2021; Panagakis et al., 2015). Amid them, RaJIVE (Ponzi et al., 2021) employs robust SVD (Zhang et al., 2013) to remove noise from the observations and then uses a variant of JIVE (Feng et al., 2018) to separate common and unique components. RJIVE (Sagonas et al., 2017) proposes a constrained optimization formulation to minimize the ℓ_1 norm of the fitting residuals and exploits ADMM to solve the problem. RCICA (Panagakis et al., 2015) adopts a similar optimization objective but uses a regularization to encourage the similarity of common subspaces and only works for $N = 2$ cases. Though these methods can achieve decent performance in applications including facial expression synthesis and audio-visual fusion, they are based on heuristics and it is not clear whether their output converges to the ground truth common and unique factors. In contrast, we prove that `TCMF` is guaranteed to recover the ground truths and use a few numerical examples to show that `TCMF` indeed recover more meaningful components.

4. Identifiability Conditions

Our goal is to decouple the common components, unique components, and the sparse noise, given a group of data observations $\{\mathbf{M}_{(i)}\}_{i=1}^{N}$. At first glance, such decoupling may seem impossible or even ill-defined: roughly speaking, the number of unknown variables, namely global components, local components, and noise, are thrice the number of observed data matrices $\{\mathbf{M}_{(i)}\}_{i=1}^{N}$, and hence, there are infinite number of decouplings that can give rise to the same $\mathbf{M}_{(i)}$.

The very first question to ask is whether such decoupling is possible and, if so, which properties can ensure the identifiability of three components. Intriguingly, we are able to prove that the exact decoupling of shared features, unique features, and noise is possible if there is "little overlap" among the three components. Below, we will formalize this intuition in more detail. Though intuitive, it turns out that these conditions can guarantee the *identifiability* of the shared components, unique components, and the sparse noise.

4.1 Sparsity

As discussed, identifying arbitrarily dense and large noise from signals is not possible. Hence, we consider sparse noise where only a small fraction of observations are corrupted. To characterize the sparsity of ${\mathbf{S}^{\star}}_{(i)}$, we use the following definition of α-sparsity.

Definition 1 *(α-sparsity) A matrix $\mathbf{S} \in \mathbb{R}^{n_1 \times n_2}$ is α-sparse if at most αn_1 entries in each column and at most αn_2 entries in each row are nonzero.*

The definition follows from that of Netrapalli et al. (2014). In Definition 1, α characterizes the maximum portion of corrupted entries in each row and each column. Intuitively, if a matrix is α-sparse with small α, then its nonzero entries are "spread out" instead of concentrated on specific columns or rows.

4.2 Incoherence

It is shown that distinguishing sparse components from arbitrary low-rank components is also hard (Candès et al., 2011; Netrapalli et al., 2014). As a simple counterexample, the matrix $\mathbf{M} = \mathbf{e}_i \mathbf{e}_j^T$, where we use $\mathbf{e}_i$ to denote the basis vector of axis i, has its ij-th entry to be 1 and all other entries to be 0. This matrix has rank 1, and is also sparse since it has only one nonzero entry. Thus, deciding whether it is sparse or low rank is difficult as it satisfies both requirements.

From the above analysis, one can see that the low-rank components should not be sparse. In other words, to be distinguishable from the sparse noise, their elements should be sufficiently spread out. In the literature, this requirement is often characterized by the so-called incoherence condition (Candès et al., 2011; Netrapalli et al., 2014).

Definition 2 *(μ-incoherence) A matrix $\mathbf{U} \in \mathbb{R}^{n \times r}$ is μ-incoherent if*

$$\max_i \left\| \mathbf{e}_i^T \mathbf{U} \right\|_2 \leq \frac{\mu \sqrt{r}}{\sqrt{n}},$$

where $\mathbf{e}_i \in \mathbb{R}^n$ is the standard basis vector of axis i, defined as $\mathbf{e}_i = (0, 0, ...0, 1, 0, ...0)^T$

The incoherence condition restricts the maximum row-wise ℓ_2 norm of a matrix $\mathbf{U}$, thus preventing the entries of $\mathbf{U}$ from being too concentrated on a few specific axes.

Remember that in model (1), $\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g}$ and $\mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}$ represent the global (shared) and local (unique) factors. For any $n \geq r$, we use $\mathbb{O}^{n\times r}$ to denote the set of n by r matrices whose column vectors are orthonormal, $\mathbb{O}^{n\times r} = \{\mathbf{W} \in \mathbb{R}^{n\times r} | \mathbf{W}^T\mathbf{W} = \mathbf{I}\}$. We assume the SVD of $\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g}$ and $\mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}$ has the following form,

$$\begin{cases} \mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g} = \mathbf{H}^\star{}_g\mathbf{\Sigma}^\star{}_{(i),g}\mathbf{W}^{\star T}_{(i),g} \\ \mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l} = \mathbf{H}^\star{}_{(i),l}\mathbf{\Sigma}^\star{}_{(i),l}\mathbf{W}^{\star T}_{(i),l} \end{cases}, \tag{3}$$

where $\mathbf{H}^\star{}_g \in \mathbb{O}^{n_1\times r_1}$, $\mathbf{W}^\star{}_{(i),g} \in \mathbb{O}^{n_{2,(i)}\times r_1}$, $\mathbf{H}^\star{}_{(i),l} \in \mathbb{O}^{n_1\times r_{2,(i)}}$. Moreover, $\mathbf{W}^\star{}_{(i),l} \in \mathbb{O}^{n_{2,(i)}\times r_{2,(i)}}$ are orthogonal matrices, $\mathbf{\Sigma}^\star{}_{(i),g} \in \mathbb{R}^{r_1\times r_1}$ and $\mathbf{\Sigma}^\star{}_{(i),l} \in \mathbb{R}^{r_{2,(i)}\times r_{2,(i)}}$ are positive diagonal matrices. In (3), we consider the case where the global and local column singular vectors are orthogonal, i.e., $\mathbf{H}^{\star T}_g\mathbf{H}^\star{}_{(i),l} = 0$. We use $r_2 = \max_i r_{2,(i)}$ throughout the paper.

To avoid overlapping between sparse and low-rank components, we assume the row and column singular vectors $\mathbf{H}^\star{}_g$, $\mathbf{H}^\star{}_{(i),l}$, $\mathbf{W}^\star{}_{(i),g}$, and $\mathbf{W}^\star{}_{(i),l}$ are all μ-incoherent. This assumption ensures that the low-rank components do not have entries too concentrated on specific rows or columns. As a result, the incoherence on singular vectors encourages the low-rank components to distribute evenly on all entries, which is distinguished from sparse noises that are nonzero on a small fraction of entries.

4.3 Misalignment

As discussed in (3), we use orthogonality between shared and unique features $\hat{\mathbf{H}}^{\star T}_g\hat{\mathbf{H}}^\star{}_{(i),l} = 0$ to encode our prior belief about the independence of different features. This is equivalent to $\mathbf{U}^{\star T}_g\mathbf{U}^\star{}_{(i),l} = 0$. Such orthogonality, however, is still insufficient to guarantee the identifiability of shared and unique factors.

To see this, consider a counterexample where all $\mathbf{U}^\star{}_{(i),l}$'s are equal, i.e., $\mathbf{U}^\star{}_{(1),l} = \mathbf{U}^\star{}_{(2),l} = \cdots = \mathbf{U}^\star{}_{(N),l}$. In this case, "unique" factors are also shared among all observation matrices. Thus, separating them from the ground truth $\mathbf{U}^\star{}_g$ is not possible. From this counterexample, we can see that it is essential for the local features not to be perfectly aligned with each other. Next, we formally introduce the notion of misalignment. For a full column-rank matrix $\mathbf{U} \in \mathbb{R}^{d\times n}$, we define the projection matrix $\mathbf{P}_\mathbf{U} \in \mathbb{R}^{d\times d}$ as $\mathbf{P}_\mathbf{U} = \mathbf{U}\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T$.

Definition 3 *(θ-misalignment) We say $\{\mathbf{U}^\star{}_{(i),l}\}$ are θ-misaligned if there exists a positive constant $\theta \in (0,1)$ such that:*

$$\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{P}_{\mathbf{U}^\star{}_{(i),l}}\right) \leq 1-\theta. \tag{4}$$

By the triangular inequality of $\lambda_{\max}(\cdot)$, we know $\lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{P}_{\mathbf{U}^\star{}_{(i),l}}\right) \leq \frac{1}{N}\sum_{i=1}^{N}\lambda_{\max}\left(\mathbf{P}_{\mathbf{U}^\star{}_{(i),l}}\right) = 1$. Thus, the introduced θ is always nonnegative. Indeed, all $\mathbf{P}_{\mathbf{U}^\star{}_{(i),l}}$'s have a common nonempty eigenspace with eigenvalue 1 if and only if $\theta = 0$. Thus, the θ-misalignment condition requires that the subspaces spanned by all unique factors do

not contain a common subspace. On the contrary, all global features are shared; hence, the subspaces spanned by these features are also identical. This comparison shows that the misalignment condition unequivocally distinguishes unique features from shared ones.

As a concrete example, consider $N = 2$ and $\mathbf{U}_{(1),l} = (\cos\vartheta, \sin\vartheta)^T$, $\mathbf{U}_{(2),l} = (\cos\vartheta, -\sin\vartheta)^T$ for $\vartheta \in [0, \frac{\pi}{4}]$. Indeed, the angle between $\mathbf{U}_{(1),l}$ and $\mathbf{U}_{(2),l}$ is 2ϑ and

$$\frac{1}{2}\left(\mathbf{P}_{\mathbf{U}_{(1),l}} + \mathbf{P}_{\mathbf{U}_{(2),l}}\right) = \begin{pmatrix} \cos^2\vartheta, & 0 \\ 0, & \sin^2\vartheta \end{pmatrix}.$$

Hence, by definition, $\theta = \sin^2\vartheta$. We can thus clearly see that when ϑ increases, the $\mathbf{U}_{(1),l}$ and $\mathbf{U}_{(2),l}$ become more misaligned.

The notion of θ-misalignment is first proposed by Shi and Kontar (2024) and intimately related to the uniqueness conditions in Lock et al. (2013).

5. Algorithm

The introduced identifiability conditions restrict the overlaps between shared, unique, and sparse components. It remains to develop algorithms to untangle the three parts from N matrices. In Section 5.1, we introduce a constrained optimization formulation, and in Section 5.2, we propose an alternating minimization program to decouple the three parts. The alternating minimization requires solving subproblems to distinguish shared features from unique ones.

Throughout the paper, we use $\|\mathbf{A}\|$ or $\|\mathbf{A}\|_2$ to denote the operator norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\|\mathbf{A}\|_F$ to denote the Frobenius norm of $\mathbf{A}$. We use r to denote $r = r_1 + r_2$.

5.1 Constrained Nonconvex Nonsmooth Optimization

We design a constrained optimization problem to decouple the three components. The decision variables $\mathbf{x}$ include features, coefficients, and sparse noise estimates: $\mathbf{x} = \left(\mathbf{U}_g, \{\mathbf{U}_{(i),l}, \mathbf{V}_{(i),g}, \mathbf{V}_{(i),l}, \mathbf{S}_{(i)}\}_{i=1}^N\right)$. The constrained optimization is formulated as,

$$\begin{aligned} \min_{\mathbf{x}} \quad & \sum_{i=1}^{N} h_i(\mathbf{U}_g, \mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}, \mathbf{S}_{(i)}; \lambda) \\ \text{s.t.} \quad & \mathbf{U}_g^T \mathbf{U}_{(i),l} = 0, \ \forall i \in [N]. \end{aligned} \tag{5}$$

Here, h_i is a regularized fitting residual consisting of two parts:

$$\begin{aligned} & h_i(\mathbf{U}_g, \mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}, \mathbf{S}_{(i)}; \lambda) \\ & = f_i(\mathbf{U}_g, \mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}, \mathbf{S}_{(i)}) + \Phi_i(\mathbf{S}_{(i)}; \lambda) \\ & = \frac{1}{2}\left\|\mathbf{M}_{(i)} - \mathbf{U}_g\mathbf{V}_{(i),g}^T - \mathbf{U}_{(i),l}\mathbf{V}_{(i),l}^T - \mathbf{S}_{(i)}\right\|_F^2 && (f_i) \\ & + \lambda^2 \left\|\mathbf{S}_{(i)}\right\|_0. && (\Phi_i) \end{aligned}$$

Term (f_i) measures the distance between the sum of shared, unique, and sparse components and the observation matrix $\mathbf{M}_{(i)}$. It denotes the residual of fitting. A common approach for

solving this problem is based on convex relaxation (Candès et al., 2011). However, convex relaxation increases the number of variables to $\mathcal{O}(n_1 n_2)$, while our nonconvex formulation keeps it in the order of $\mathcal{O}(\max\{n_1, n_2\}(r_1 + r_2))$, which is significantly smaller.

Term (Φ_i) is an ℓ_0 regularization term that promotes the sparsity of matrix $\mathbf{S}_{(i)}$. The parameter λ mediates the balance between the ℓ_0 penalty with the residual of fitting. A large value of λ leads to sparser $\mathbf{S}_{(i)}$ with only large nonzero elements. Conversely, a small value of λ yields a denser $\mathbf{S}_{(i)}$ with potentially small nonzero elements. Therefore, to identify both large and small nonzero values of $\mathbf{S}_{(i)}$, while correctly filtering out its zero elements, we propose to gradually decrease the value of λ during the optimization of objective (5). We use the notation $h_i(\mathbf{U}_g, \mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}, \mathbf{S}_{(i)}; \lambda)$ to explicitly show that the objective h_i is dependent on the regularization parameter λ.

At first glance, the proposed optimization problem (5) may appear daunting due to its inherent nonconvexity and nonsmoothness. Notably, it exhibits two distinct sources of nonconvexity: firstly, both terms (f_i) and (Φ_i) are nonconvex, and secondly, the feasible set corresponding to the constraint $\mathbf{U}_g^T \mathbf{U}_{(i),l} = 0$ is also nonconvex. Furthermore, the ℓ_0 regularization term in (Φ_i) introduces nonsmoothness into the problem. However, we will introduce an intuitive and efficient algorithm designed to alleviate these challenges and effectively solve the problem. Surprisingly, under our identifiability conditions introduced in Section 4, this algorithm can be proven to converge to the ground truth.

5.2 Alternating Minimization

One efficient approach to solving a ℓ_0 regularized objective is alternating minimization. We divide the decision variables $\mathbf{x}$ into 2 blocks, $\left(\mathbf{U}_g, \{\mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}\}\right)$ and $\left(\{\mathbf{S}_{(i)}\}\right)$, and alternatively minimize one block with the block of variables fixed.

More specifically, the alternating minimization proceeds by epochs, each comprised of two steps. For ease of exposition, we use $\hat{\mathbf{U}}_{g,t-1}, \{\hat{\mathbf{V}}_{(i),g,t-1}, \hat{\mathbf{U}}_{(i),l,t-1}, \hat{\mathbf{V}}_{(i),l,t-1}, \hat{\mathbf{S}}_{(i),t-1}\}_{i=1}^N$ to denote the values of $\mathbf{x}$ at the end of epoch $t-1$. The $\hat{}$ notation represents the estimated values of the variables.

In the first step, we fix the values of $\left(\hat{\mathbf{U}}_{g,t-1}, \{\hat{\mathbf{V}}_{(i),g,t-1}, \hat{\mathbf{U}}_{(i),l,t-1}, \hat{\mathbf{V}}_{(i),l,t-1}\}\right)$, and optimize over $\{\mathbf{S}_{(i)}\}$. The optimal $\hat{\mathbf{S}}_{(i),t}$ has a simple closed-form solution given by hard-thresholding,

$$\begin{aligned}\hat{\mathbf{S}}_{(i),t} &= \arg\min_{\mathbf{S}_{(i)}} \left\|\mathbf{M}_{(i)} - \hat{\mathbf{U}}_{g,t-1}\hat{\mathbf{V}}_{(i),g,t-1}^T - \hat{\mathbf{U}}_{(i),l,t-1}\hat{\mathbf{V}}_{(i),l,t-1}^T - \mathbf{S}_{(i)}\right\|_F^2 + \lambda_t^2 \left\|\mathbf{S}_{(i)}\right\|_0 \\ &= \mathrm{Hard}_{\lambda_t}\left[\mathbf{M}_{(i)} - \hat{\mathbf{U}}_{g,t-1}\hat{\mathbf{V}}_{(i),g,t-1}^T - \hat{\mathbf{U}}_{(i),l,t-1}\hat{\mathbf{V}}_{(i),l,t-1}^T\right],\end{aligned}$$

where $\mathrm{Hard}_\lambda(\cdot)$ is the hard-thresholding operator. For a matrix $X \in \mathbb{R}^{m\times n}$, the hard thresholding operator is defined as:

$$\left[\mathrm{Hard}_\lambda(X)\right]_{ij} = \begin{cases} X_{ij} \text{ , if } |X_{ij}| > \lambda \\ 0 \text{ , if } X_{ij} \in [-\lambda, \lambda] \end{cases}. \tag{7}$$

The coefficient λ is a thresholding parameter that controls the sparsity of the output. To recover the correct sparsity pattern of $\hat{\mathbf{S}}_{(i),t}$, our approach is to maintain a small false

positive rate (elements that are incorrectly identified as nonzero), while gradually improving the true positive rate (elements that are correctly identified as nonzero). To this goal, we start with a large λ to obtain a conservative estimate of $\hat{\mathbf{S}}_{(i),t}$. Then, we decrease λ to refine the estimate.

In the second step, we fix $\hat{\mathbf{S}}_{(i),t}$ and optimize $\big(\mathbf{U}_g, \{\mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}\}\big)$ under the constraint $\mathbf{U}_g^T\mathbf{U}_{(i),l} = 0$. Removing the ℓ_0 regularization term that is independent of $\big(\mathbf{U}_g, \{\mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}\}\big)$, the optimization subproblem takes the following form,

$$\begin{aligned} \min_{\left(\mathbf{U}_g, \{\mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l}\}\right)} \quad & \sum_{i=1}^{N} \left\| \hat{\mathbf{M}}_{(i)} - \mathbf{U}_g\mathbf{V}_{(i),g}^T - \mathbf{U}_{(i),l}\mathbf{V}_{(i),l}^T \right\|_F^2 \\ \text{s.t.} \quad & \mathbf{U}_g^T\mathbf{U}_{(i),l} = 0, \forall i \in [N], \end{aligned} \tag{8}$$

where $\hat{\mathbf{M}}_{(i)} = \mathbf{M}_{(i)} - \hat{\mathbf{S}}_{(i),t}$.

Despite its nonconvexity, there exist several iterative algorithms to solve the above optimization problem, including but not limited to JIVE (Lock et al., 2013), COBE (Zhou et al., 2015), PerPCA (Shi and Kontar, 2024), PerDL (Liang et al., 2023), and HMF (Shi et al., 2023). Given the similarity of these methods, we employ the name Joint and Individual Matrix Factorization (`JIMF`) to encapsulate the subroutine addressing problem (8).

The versatile `JIMF` is a meta-algorithm that can be implemented using any of the aforementioned methods, provided that they generate good solutions. Among these algorithms, PerPCA and HMF are of special interest as they are proved to converge to the optimal solutions of (8) under suitable conditions. They are also intrinsically federated as most of the computation can be distributed on N sources where the data are generated.

As problem (8) does not have a simple closed-form solution, the algorithms discussed above are iterative. The iterative algorithms do not output exact optimal solutions. Instead, they refine the estimates at every iteration. Therefore, there will be a difference between our algorithm-generated solutions and the optimal solution. To characterize the degree of such difference, we resort to employing the concept of ϵ-optimality, a notion well-known in the optimization community.

Definition 4 *(ϵ-optimality) Given $(\hat{\mathbf{U}}_g, \{\hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l}\})$ as any global optimal solution to the problem* (8) *and a constant $\epsilon > 0$, we say $(\hat{\mathbf{U}^\epsilon}_g, \{\hat{\mathbf{V}^\epsilon}_{(i),g}, \hat{\mathbf{U}^\epsilon}_{(i),l}, \hat{\mathbf{V}^\epsilon}_{(i),l}\})$ is an ϵ-optimal solution to* (8) *if it satisfies,*

$$\left\| \hat{\mathbf{U}^\epsilon}_g\hat{\mathbf{V}^{\epsilon T}}_{(i),g} + \hat{\mathbf{U}^\epsilon}_{(i),l}\hat{\mathbf{V}^{\epsilon T}}_{(i),l} - \hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T - \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T \right\|_\infty \leq \epsilon, \quad \forall i$$

and

$$\hat{\mathbf{U}}_g^{\epsilon T}\hat{\mathbf{U}^\epsilon}_{(i),l} = 0, \quad \forall i.$$

The nonconvexity of (8) gives rise to multiple global optimal solutions. Our definition of ϵ-optimality only emphasizes the closeness between the *product* of features and the coefficients, and the product of *any* set of global optimal solutions. As discussed, there exist multiple methods proposed to solve (8) that demonstrate decent practical performance. In particular, PerPCA, PerDL, and HMF are proved to converge to the optimal solutions of (8) at linear rates when initialized properly. Hence, under suitable initializations, PerPCA

and HMF can reach ϵ-optimality of (8) within $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ iterations for any value of ϵ. The details of the two algorithms will be discussed in Appendix A.1.

With the help of subroutine `JIMF`, the main alternating minimization algorithm proceeds by optimizing two blocks of variables iteratively. We present the pseudo-code in Algorithm 1.

Algorithm 1 `TCMF`: alternating minimization

Input observation matrices from N sources $\{\mathbf{M}_{(i)}\}_{i=1}^{N}$, constant λ_1, multiplicative factor $\rho \in (0,1)$, precision ϵ.
Initialize $\hat{\mathbf{U}}^{\epsilon}{}_{g,0}, \hat{\mathbf{V}}^{\epsilon}{}_{(i),g,0}, \hat{\mathbf{U}}^{\epsilon}{}_{(i),l,0}, \hat{\mathbf{V}}^{\epsilon}{}_{(i),l,0}, \hat{\mathbf{S}}_{(i),0}$ to be zero matrices.
for Epoch $t = 1, ..., T$ **do**
  for Source $i = 1, \cdots, N$ **do**
    $\hat{\mathbf{S}}_{(i),t} = \text{Hard}_{\lambda_t}\left[\mathbf{M}_{(i)} - \hat{\mathbf{U}}^{\epsilon}{}_{g,t-1}\hat{\mathbf{V}}^{\epsilon T}{}_{(i),g,t-1} - \hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t-1}\hat{\mathbf{V}}^{\epsilon T}{}_{(i),l,t-1}\right]$
  end for
  $(\hat{\mathbf{U}}^{\epsilon}{}_{g,t}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),g,t}\}, \{\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t}\}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),l,t}\}) = \texttt{JIMF}\left(\{\hat{\mathbf{M}}_{(i)}\} = \{\mathbf{M}_{(i)} - \hat{\mathbf{S}}_{(i),t}\}, \epsilon\right)$
  Set $\lambda_{t+1} = \rho\lambda_t + \epsilon$
end for
Return $\{\hat{\mathbf{U}}^{\epsilon}{}_{g,T}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),g,T}\}, \{\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,T}\}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),l,T}\}\}$.

In Algorithm 1, we use $\texttt{JIMF}\left(\{\hat{\mathbf{M}}_{(i)}\}, \epsilon\right)$ to denote the call for a subroutine to solve (8) to ϵ-optimality. In each epoch, sparse matrices $\hat{\mathbf{S}}_{(i),t}$ are firstly estimated by hard thresholding. Then $\hat{\mathbf{M}}_{(i)} = \mathbf{M}_{(i)} - \hat{\mathbf{S}}_{(i),t}$ are calculated, which are subsequently decoupled into the shared and unique components via a `JIMF` call. The output of this subroutine is represented as $(\hat{\mathbf{U}}^{\epsilon}{}_{g,t}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),g,t}\}, \{\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t}\}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),l,t}\})$, where the superscript $^{\epsilon}$ signifies ϵ-optimality. The outputs $(\hat{\mathbf{U}}^{\epsilon}{}_{g,t}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),g,t}\}, \{\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t}\}, \{\hat{\mathbf{V}}^{\epsilon}{}_{(i),l,t}\})$ are used to improve the estimate of $\hat{\mathbf{S}}_{(i)}$ in the next epoch. After each epoch, we decrease the thresholding parameter λ_t by a constant $\rho < 1$, then add a constant ϵ. The inclusion of ϵ in λ_{t+1} is necessary to ensure that the estimated $\hat{\mathbf{S}}_{(i),t+1}$ does not contain any false positive entries. By incorporating ϵ into λ_{t+1}, we guarantee that the inexactness of the `JIMF` outputs does not undermine the false positive rate of the entries in $\hat{\mathbf{S}}_{(i),t+1}$.

Then, per-epoch computational complexity of Algorithm 1 is $O(n_1 n_2 N)$ when the rank r_1 and r_2 are small $r_1, r_2 \ll n_1, n_2$. To see this, we can add up the computation complexity complexity of hard-thresholding and `JIMF`. Element-wise hard-thresholding requires $O(n_1 n_2)$ computations for each source. Efficient implementations of the subroutine `JIMF`, such as `PerPCA` and `HMF`, can converge into the $\epsilon-$optimal solutions within $\mathcal{O}(\log \frac{1}{\epsilon})$ steps, where each step require $\mathcal{O}(n_1 n_2)$ computations. Therefore, the per-epoch computational complexity of `TCMF` is $\mathcal{O}(n_1 n_2 N)$, where log factors are omitted.

Furthermore, if the `JIMF` and hard-thresholding are distributed among N sources, `TCMF` can further exploit parallel computation to reduce the running time. In the regime where communication cost is negligible, the per-iteration total running time scales as $\mathcal{O}\left(n_1 n_2 + N n_1\right)$.

We will show later that such a design can ensure that the estimation error diminishes linearly. A pictorial representation of Algorithm 1 is plotted in the left graph of Figure 1.

6. Convergence Analysis

In this section, we will analyze the convergence of Algorithm 1. Our theorem characterizes the conditions under which Algorithm 1 converges linearly to the ground truth. We additionally introduce $\sigma_{\max} > 0$ to denote an upper bound of the singular values of $\{\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g} + \mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}\}_{i=1}^N$, and $\sigma_{\min} > 0$ to denote a lower bound on the smallest nonzero singular values of $\{\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g} + \mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}\}_{i=1}^N$. For simplicity we assume $n_{2,(i)} = n_2$, $r_{2,(i)} = r_2$, and $r = r_1 + r_2$ in this section.

Theorem 5 (Convergence of Algorithm 1) *Consider the true model* (1) *with SVD defined in* (3)*, where nonzero singular values of* $\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g} + \mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}$ *are lower bounded by* $\sigma_{\min} > 0$ *and upper bounded by* $\sigma_{\max} \geq \sigma_{\min}$ *for each source* i*. Suppose that the following conditions are satisfied:*

- **($\boldsymbol{\mu}$-incoherency)** *The matrices* $\mathbf{H}^\star{}_g$ *and* $\left\{\mathbf{H}^\star{}_{(i),l}, \mathbf{W}^\star{}_{(i),g}, \mathbf{W}^\star{}_{(i),l}\right\}_{i=1}^N$ *are* μ*-incoherent for a constant* $\mu > 0$*.*
- **($\boldsymbol{\theta}$-misalignment)** *The local feature matrices* $\left\{\mathbf{U}^\star{}_{(i),l}\right\}_{i=1}^N$ *are* θ*-misaligned for a constant* $0 < \theta < 1$*.*
- **($\boldsymbol{\alpha}$-sparsity)** *The matrices* $\left\{\mathbf{S}^\star{}_{(i)}\right\}_{i=1}^N$*'s are* α*-sparse for some* $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\right)$*, where* $r = r_1 + r_2$*.*

Then, there exist constants $C_{g,1}, C_{g,2}, C_{l,1}, C_{l,2}, C_{s,1}, C_{s,2} > 0$ *and* $\rho_{\min} = \mathcal{O}\left(\sqrt{\alpha}\frac{\mu^2 r}{\sqrt{\theta}}\right) < 1$ *such that the iterations of Algorithm 1 with* $\lambda_1 = \frac{\sigma_{\max}\mu^2 r}{\sqrt{n_1 n_2}}$*,* $\epsilon \leq \lambda_1(1 - \rho_{\min})$*, and* $1 - \frac{\epsilon}{\lambda_1} > \rho \geq \rho_{\min}$ *satisfy*

$$\left\|\hat{\mathbf{U}}^\epsilon{}_{g,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),g,t} - \mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g}\right\|_\infty \leq C_{g,1}\rho^t + C_{g,2}\epsilon \tag{9}$$

$$\left\|\hat{\mathbf{U}}^\epsilon{}_{(i),l,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),l,t} - \mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}\right\|_\infty \leq C_{l,1}\rho^t + C_{l,2}\epsilon \tag{10}$$

$$\left\|\hat{\mathbf{S}}_{(i),t} - \mathbf{S}^\star{}_{(i)}\right\|_\infty \leq C_{s,1}\rho^t + C_{s,2}\epsilon. \tag{11}$$

Theorem 5 presents a set of sufficient conditions under which the model is identifiable, and Algorithm 1 converges to the ground truth at a linear rate. As discussed in Section 4, these conditions are indeed sufficient to guarantee the identifiability of the true model. In particular, μ-incoherency is required for disentangling global and local components from noise, whereas θ-misalignment is needed to separate local and global components. Moreover, there is a natural trade-off between the parameters μ, θ, and α: the upper bound on the sparsity level α, $\mathcal{O}\left(\frac{\theta^2}{\mu^4 r^2}\right)$, is proportional to θ^2, implying that more alignment among local feature matrices can be tolerated only at the expense of sparser noise matrices. Similarly, α is inversely proportional to μ^4, indicating that more coherency in the local and global components is only possible with sparser noise matrices. We also highlight the dependency of α on the rank r; such dependency is required even in the standard settings of robust PCA (Netrapalli et al., 2014; Chandrasekaran et al., 2011; Hsu et al., 2011), albeit with a milder condition on r. Finally, the scaling of α does not depend on the the number of

sources N, suggesting that the convergence guarantees provided by Theorem 5 are valid for extremely large datasets.

Two important observations are in order. First, we do not impose any constraint on the norm or sign of the sparse noise $\mathbf{S}^\star{}_{(i)}$. Thus, Theorem 5 holds for arbitrarily large noise values. Second, at every epoch, Algorithm 1 solves the inner optimization problem (8) via `JIMF` to ϵ-optimality. Also, the convergence of Algorithm 1 is contingent upon the precision of `JIMF` output: the ℓ_∞ norm of the optimization error should not be larger than $\mathcal{O}(\lambda_1)$. Such requirement is not strong as even the trivial solution $\mathbf{U}_g, \mathbf{V}_{(i),g}, \mathbf{U}_{(i),l}, \mathbf{V}_{(i),l} = 0$ is λ_1-optimal. One should expect many algorithms to perform much better than the trivial solution. Indeed, methods including PerPCA and PerDL are proved to output ϵ-optimal solutions for arbitrary small ϵ within logarithmic iterations, thus satisfying the requirement. In practice, heuristic methods including JIVE or COBE can output reasonable solutions that may also satisfy the requirement in Theorem 5.

In the next section, we provide the sketch of the proof for Theorem 5.

6.1 Proof Sketch of Theorem 5

Algorithm 1 is essentially an alternating minimization algorithm comprising a hard-thresholding step, followed by a joint and individual matrix factorization step. Our overarching goal is to control the estimation error at each iteration of the algorithm, showing that it decreases by a constant factor after every epoch. To this goal, we make extensive use of the error matrix $\mathbf{E}_{(i),t}$ defined as $\mathbf{E}_{(i),t} = \mathbf{S}^\star{}_{(i)} - \hat{\mathbf{S}}_{(i),t}$ for every client i.

In the ideal case where $\mathbf{E}_{(i),t} = 0$, the global solution of (5) coincides with the true shared and unique components, which is guaranteed by Theorem 1 in Shi and Kontar (2024). Therefore, it is crucial to control the behavior of $\{\mathbf{E}_{(i),t}\}_{i=1}^N$ and its effect on the recovered solution throughout the course of the algorithm. We define $\mathbf{L}^\star_{(i)} = \mathbf{U}^\star{}_g \mathbf{V}^{\star T}_{(i),g} + \mathbf{U}^\star{}_{(i),l} \mathbf{V}^{\star T}_{(i),l}$ and $\hat{\mathbf{L}}_{(i),t} = \hat{\mathbf{U}}_{g,t} \hat{\mathbf{V}}^T_{(i),g,t} + \hat{\mathbf{U}}_{(i),l,t} \hat{\mathbf{V}}^T_{(i),l,t}$ as the true and estimated low-rank components of client i. Similarly, $\hat{\mathbf{L}}^\epsilon_{(i),t} = \hat{\mathbf{U}}^\epsilon{}_{g,t} \hat{\mathbf{V}}^{\epsilon T}_{(i),g,t} + \hat{\mathbf{U}}^\epsilon{}_{(i),l,t} \hat{\mathbf{V}}^{\epsilon T}_{(i),l,t}$ is the reconstructed low-rank component from ϵ-optimal estimates. The following steps outline the sketch of our proof:

Step 1: α-sparsity of the initial error: At the first iteration, the threshold level λ_1 is large, enforcing $\operatorname{supp}\left(\hat{\mathbf{S}}_{(i),1}\right) \subseteq \operatorname{supp}\left(\mathbf{S}^\star{}_{(i)}\right)$, which in turn implies $\operatorname{supp}\left(\mathbf{E}_{(i),1}\right) \subseteq \operatorname{supp}\left(\mathbf{S}^\star{}_{(i)}\right)$. Therefore, the initial error matrix $\mathbf{E}_{(i),1}$ is also α-sparse.

Step 2: Error reduction via `JIMF`. Suppose that $\mathbf{E}_{(i),t}$ is α-sparse. In Step 7, $\hat{\mathbf{L}}_{(i),t}$ is obtained by applying `JIMF` on $\mathbf{M}_{(i)} - \hat{\mathbf{S}}_{(i),t} = \mathbf{L}^\star_{(i)} + \mathbf{E}_{(i),t}$. Note that the input to `JIMF` is the true low-rank component perturbed by an α-sparse matrix $\mathbf{E}_{(i),t}$. One of our key contributions is to show that $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}_{(i),t}\right\|_\infty$ is much smaller than $\left\|\mathbf{E}_{(i),t}\right\|_\infty$, provided that the true local and global components are μ-incoherent and $\mathbf{E}_{(i),t}$ is α-sparse. This fact is delineated in the following key lemma.

Lemma 6 (Error reduction via `JIMF` (informal)) *Suppose that the conditions of Theorem 5 are satisfied. Moreover, suppose that $\mathbf{E}_{(i),t}$ is α-sparse for each client i. We have*

$$\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}_{(i),t}\right\|_\infty \leq C \cdot \frac{\sqrt{\alpha}\mu^2 r}{\sqrt{\theta}} \cdot \max_j \left\{\left\|\mathbf{E}_{(j),t}\right\|_\infty\right\},$$

where $C > 0$ is a constant.

Indeed, proving Lemma 6 is particularly daunting since $\hat{\mathbf{L}}_{(i),t}$ does not have a closed-form solution. We will elaborate on the major techniques to prove Lemma 6 in Section 6.2.

Suppose that α is small enough such that $C \cdot \frac{\sqrt{\alpha}\mu^2 r}{\sqrt{\theta}} \leq \frac{\rho}{2}$. Then, Lemma 6 implies that $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}_{(i),t}\right\|_\infty \leq \frac{\rho}{2}\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\}$. From the definition of ϵ-optimality, we know $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right\|_\infty \leq \frac{\rho}{2}\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\} + \epsilon$. This implies that the ℓ_∞ norm of the error in the output of `JIMF` shrinks by a factor of $\frac{\rho}{2}$ compared with the error in the input $\left\|\mathbf{E}_{(i),t}\right\|_\infty$ (modulo an additive factor ϵ). As will be discussed next, this shrinkage in the ℓ_∞ norm of the error is essential for the exact sparsity recovery of the noise matrix.

Step 3: Preservation of sparsity via hard-thresholding. Given that $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right\|_\infty \leq \frac{\rho}{2}\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\} + \epsilon$, our next goal is to show that $\mathrm{supp}\left(\mathbf{E}_{(i),t+1}\right) \subseteq \mathrm{supp}\left(\mathbf{S}^\star{}_{(i)}\right)$ (i.e., $\mathbf{E}_{(i),t+1}$ remains α-sparse) and $\max_i\left\{\left\|\mathbf{E}_{(i),t+1}\right\|_\infty\right\} \leq 2\lambda_{t+1}$. To prove $\mathrm{supp}\left(\mathbf{E}_{(i),t+1}\right) \subseteq \mathrm{supp}\left(\mathbf{S}^\star{}_{(i)}\right)$, suppose that $\left(\mathbf{S}^\star_{(i)}\right)_{kl} = 0$ for some (k,l), we have $\left(\hat{\mathbf{S}}_{(i),t+1}\right)_{kl} \neq 0$ only if $\left|\left(\mathbf{M}_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right)_{kl}\right| = \left|\left(\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right)_{kl}\right| > \lambda_{t+1}$. On the other hand, in the Appendix, we show that $\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\} \leq 2\lambda_t$. This implies that $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right\|_\infty \leq \frac{\rho}{2}\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\} + \epsilon \leq \rho\lambda_t + \epsilon = \lambda_{t+1}$. This in turn leads to $\left(\hat{\mathbf{S}}_{(i),t+1}\right)_{kl} = \left(\mathbf{E}_{(i),t+1}\right)_{kl} = 0$, and hence, $\mathrm{supp}\left(\mathbf{E}_{(i),t+1}\right) \subseteq \mathrm{supp}\left(\mathbf{S}^\star{}_{(i)}\right)$. Finally, according to the definition of hard-thresholding, we have $\left|\left(\hat{\mathbf{S}}_{(i),t+1} - \left(\mathbf{S}^\star_{(i)} + \mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i)}\right)\right)_{kl}\right| \leq \lambda_{t+1}$, which, by triangle inequality, yields $\left|\left(\mathbf{E}_{(i),t+1}\right)_{kl}\right| \leq \left|\left(\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i)}\right)_{kl}\right| + \lambda_{t+1} \leq 2\lambda_{t+1}$.

Step 4: Establishing linear convergence. Repeating Steps 2 and 3, we have $\max_i\left\{\left\|\mathbf{E}_{(i),t+1}\right\|_\infty\right\} \leq 2\lambda_{t+1}$ and $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right\|_\infty \leq \lambda_{t+1}$ for all t. Noting that $\lambda_t = \rho\lambda_{t-1} + \epsilon = \epsilon + \rho\epsilon + \rho^2\lambda_{t-2} = \cdots = \epsilon + \rho\epsilon + \rho^3\epsilon + \cdots + \rho^{t-1}\lambda_1 \leq \frac{\epsilon}{1-\rho} + \rho^{t-1}\lambda_1$, we establish that $\max_i\left\{\left\|\mathbf{E}_{(i),t}\right\|_\infty\right\} = \mathcal{O}(\epsilon)$ and $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right\|_\infty = \mathcal{O}(\epsilon)$ in $\mathcal{O}\left(\frac{\log(\lambda_1/\epsilon)}{\log(1/\rho)}\right)$ iterations.

Step 5: Untangling global and local components. Under the misalignment condition, a small error of the joint low-rank components $\left\|\mathbf{L}^\star_{(i)} - \hat{\mathbf{L}}_{(i),t}\right\|_F$ indicates that both the shared component and the unique component is small. More specifically, Theorem 1 in Shi and Kontar (2024) indicates $\left\|\mathbf{U}^\star{}_g\mathbf{V}^{\star T}_{(i),g} - \hat{\mathbf{U}}_{g,t}\hat{\mathbf{V}}^T_{(i),g,t}\right\|_F, \left\|\mathbf{U}^\star{}_{(i),l}\mathbf{V}^{\star T}_{(i),l} - \hat{\mathbf{U}}_{(i),l,t}\hat{\mathbf{V}}^T_{(i),l,t}\right\|_F = \mathcal{O}\left(\left\|\mathbf{L}^\star{}_{(i)} - \hat{\mathbf{L}}_{(i),l,t}\right\|_F\right)$. Since $\left\|\mathbf{L}^\star{}_{(i)} - \hat{\mathbf{L}}_{(i),l,t}\right\|_F$ shrinks linearly to a small constant, we can conclude that the estimation errors for shared and unique features also decrease linearly to $\mathcal{O}(\epsilon)$.

6.2 Proof of Lemma 6

At the crux of our proof for Theorem 5 lies Lemma 6. In its essence, Lemma 6 seeks to answer the following question: if the input to `JIMF` is corrupted by α-sparse noise matrices

$\{\mathbf{E}_{(i)}\}$, how will the recovered solutions change in terms of ℓ_∞ norm? We highlight that the standard matrix perturbation analysis, such as the classical Davis-Kahan bound (Bhatia, 2013) as well as the more recent ℓ_∞ bound (Fan et al., 2018), fall short of answering this question for two main reasons. First, these bounds are overly pessimistic and cannot take into account the underlying sparsity structure of the noise. Second, they often control the singular vectors and singular values of the perturbed matrices, whereas the optimal solutions to problem (8) generally do not correspond to the singular vectors of $\hat{\mathbf{M}}_{(i)}$.

To address these challenges, we characterize the optimal solutions of (8) by analyzing its Karush–Kuhn–Tucker (KKT) conditions. We establish the KKT condition and ensure the linear independence constraint qualification (LICQ). Afterward, we obtain closed-form solutions for the KKT conditions in the form of convergent series and use these series to control the element-wise perturbation of the solutions.

KKT conditions. The following lemma shows two equivalent formulations for the KKT conditions. For convenience, we drop the subscript t in our subsequent arguments.

Lemma 7 *Suppose that* $\{\hat{\mathbf{U}}_g,\ \hat{\mathbf{U}}_{(i),l},\ \hat{\mathbf{V}}_{(i),g},\ \hat{\mathbf{V}}_{(i),l}\}$ *is the optimal solution to problem* (8) *and* $\hat{\mathbf{M}}_{(i)}$ *has rank at least* $r_1 + r_2$*. We have*

$$\sum_{i=1}^{N}\left(\hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T + \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T - \hat{\mathbf{M}}_{(i)}\right)\hat{\mathbf{V}}_{(i),g} = 0 \tag{12a}$$

$$\left(\hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T + \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T - \hat{\mathbf{M}}_{(i)}\right)\hat{\mathbf{V}}_{(i),l} = 0 \tag{12b}$$

$$\left(\hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T + \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T - \hat{\mathbf{M}}_{(i)}\right)^T\hat{\mathbf{U}}_{(i),l} = 0 \tag{12c}$$

$$\left(\hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T + \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T - \hat{\mathbf{M}}_{(i)}\right)^T\hat{\mathbf{U}}_g = 0 \tag{12d}$$

$$\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l} = \mathbf{I}, \hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g = \mathbf{I}, \hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_g = 0. \tag{12e}$$

Moreover, there exist positive diagonal matrices $\mathbf{\Lambda}_1 \in \mathbb{R}^{r_1\times r_1}$, $\mathbf{\Lambda}_{2,(i)} \in \mathbb{R}^{r_2\times r_2}$, *and* $\mathbf{\Lambda}_{3,(i)} \in \mathbb{R}^{r_1\times r_2}$ *such that the optimality conditions imply:*

$$\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_{(i),l} = \hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)} + \hat{\mathbf{H}}_g\mathbf{\Lambda}_{3,(i)} \tag{13a}$$

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_g = \hat{\mathbf{H}}_g\mathbf{\Lambda}_1 + \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{3,(i)}^T \tag{13b}$$

$$\hat{\mathbf{H}}_g^T\hat{\mathbf{H}}_g = \mathbf{I}, \hat{\mathbf{H}}_{(i),l}^T\hat{\mathbf{H}}_{(i),l} = \mathbf{I}, \hat{\mathbf{H}}_g^T\hat{\mathbf{H}}_{(i),l} = 0, \tag{13c}$$

for some $\hat{\mathbf{H}}_g \in \mathbb{O}^{n_1\times r_1}$ *that spans the same subspaces as* $\hat{\mathbf{U}}_g$*, and some* $\hat{\mathbf{H}}_{(i),l} \in \mathbb{O}^{n_1\times r_2}$ *that spans the same subspaces as* $\hat{\mathbf{U}}_{(i),l}$*.*

The $\mathbf{\Lambda}_{3,(i)}$ term in (13) complicates the relation between $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$. When $\mathbf{\Lambda}_{3,(i)}$ is nonzero, one can see that neither $\hat{\mathbf{H}}_g$ nor $\hat{\mathbf{H}}_{(i),l}$ span a invariant subspace of $\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T$. As a consequence, perturbation analysis from Netrapalli et al. (2014) based on characteristic equations is not applicable. To alleviate this issue, we provide a more delicate control over the solution set of (13).

Solutions to KKT conditions The characterization (13) contains structural information for $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$ that can be exploited for the perturbation analysis. To see this, recall the definition $\hat{\mathbf{M}}_{(i)} = \mathbf{M}_{(i)} - \hat{\mathbf{S}}_{(i)}$. Combining this definition with (13) leads to

$$\begin{cases} \left(\mathbf{E}_{(i),t}\mathbf{L}^{\star T}_{(i)} + \mathbf{L}^{\star}{}_{(i)}\mathbf{E}^T_{(i),t} + \mathbf{E}_{(i),t}\mathbf{E}^T_{(i),t}\right)\hat{\mathbf{H}}_{(i),l} - \hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)} = -\left(\mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\hat{\mathbf{H}}_{(i),l} - \hat{\mathbf{H}}_g\mathbf{\Lambda}_{3,(i)}\right) \\ \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{E}_{(i),t}\mathbf{L}^{\star T}_{(i)} + \mathbf{L}^{\star}{}_{(i)}\mathbf{E}^T_{(i),t} + \mathbf{E}_{(i),t}\mathbf{E}^T_{(i),t}\right)\hat{\mathbf{H}}_g - \hat{\mathbf{H}}_g\mathbf{\Lambda}_1 \\ \qquad\qquad = -\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\hat{\mathbf{H}}_g - \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{3,(i)}\right). \end{cases} \tag{14}$$

We can show that the norms of input errors $\mathbf{E}_{(i),t}$ $\mathbf{\Lambda}_{3,(i)}$ are upper bounded by $\mathcal{O}(\sqrt{\alpha})$. Thus, when the sparsity parameter α is not too large, we can write the solutions to (14) as a series of α.

In the limit $\alpha = 0$, we have $\mathbf{E}_{(i),t} = 0$, thus we can solve the leading terms of $\hat{\mathbf{H}}_{(i),l}$ and $\hat{\mathbf{H}}_g$ from (14). When α is not too large, we can prove the following lemma,

Lemma 8 *(informal) If α is not too large, then $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$ introduced in Lemma 7 satisfy,*

$$\begin{cases} \hat{\mathbf{H}}_g = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\left(\hat{\mathbf{H}}_g - \hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)}\mathbf{\Lambda}^T_{3,(i)}\right)\right)\mathbf{\Lambda}_6^{-1} + \mathcal{O}(\sqrt{\alpha}) \\ \hat{\mathbf{H}}_{(i),l} = \mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)} \\ \quad - \left(\frac{1}{N}\sum_{j=1}^{N}\mathbf{L}^{\star}{}_{(j)}\mathbf{L}^{\star T}_{(j)}\left(\hat{\mathbf{H}}_g - \hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}^T_{3,(j)}\right)\right)\mathbf{\Lambda}_6^{-1}\mathbf{\Lambda}_{3,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} + \mathcal{O}(\sqrt{\alpha}), \end{cases} \tag{15}$$

where $\mathcal{O}(\sqrt{\alpha})$'s are terms whose Frobenius norm and ℓ_∞ norm is upper bounded by $\mathcal{O}(\sqrt{\alpha})$, and $\mathbf{\Lambda}_6$ is defined as $\mathbf{\Lambda}_6 = \mathbf{\Lambda}_1 - \frac{1}{N}\sum_{j=1}^{N}\mathbf{\Lambda}_{3,(j)}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}^T_{3,(j)}$.

The formal version of lemma 8 and its proof are relegated to the appendix. We now briefly introduce our methodology for deriving the solutions in Lemma 8. For matrices $\mathbf{A}, \mathbf{B}, \mathbf{X}, \mathbf{Y}$ satisfying the Sylvester equation $\mathbf{AX} - \mathbf{XB} = -\mathbf{Y}$, if the spectra of $\mathbf{A}$ and $\mathbf{B}$ are separated, i.e., $\sigma_{\max}(\mathbf{A}) < \sigma_{\min}(\mathbf{B})$, then the solution can be written as $\mathbf{X} = \sum_{p=0}^{\infty}\mathbf{A}^p\mathbf{Y}\mathbf{B}^{-1-p}$. We apply this solution form to (14) and iteratively expand $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$. The exact forms of the resulting series are shown in (41) and (42) in the appendix. In the series, each term is a product of a group of sparse matrices, an incoherent matrix, and some remaining terms. Based on the special structure of the series, we can calculate upper bounds on the Frobenius norm and maximum row norm of each term in the series. The leading terms are simply $\frac{1}{N}\sum_{i=1}^{N}\mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\left(\hat{\mathbf{H}}_g - \hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)}\mathbf{\Lambda}^T_{3,(i)}\right)\mathbf{\Lambda}_6^{-1}$ and $\mathbf{L}^{\star}{}_{(i)}\mathbf{L}^{\star T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)} - \frac{1}{N}\sum_{j=1}^{N}\mathbf{L}^{\star}{}_{(j)}\mathbf{L}^{\star T}_{(j)}\left(\hat{\mathbf{H}}_g - \hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}^T_{3,(j)}\right)\mathbf{\Lambda}_6^{-1}\mathbf{\Lambda}_{3,(i)}\mathbf{\Lambda}^{-1}_{2,(i)}$. By summing up the norm of all remaining higher-order terms in the series and applying a few basic inequalities in geometric series, we can prove the result in Lemma 8.

Perturbations on the optimal solutions to (8) Lemma 8 then allows us to establish the inequality in Lemma 6. In epoch t, we know $\hat{\mathbf{L}}_{(i),t} = \hat{\mathbf{U}}_{g,t}\hat{\mathbf{V}}^T_{(i),g,t} + \hat{\mathbf{U}}_{(i),l,t}\hat{\mathbf{V}}^T_{(i),l,t}$, where $\hat{\mathbf{U}}_{g,t}, \hat{\mathbf{V}}_{(i),g,t}, \hat{\mathbf{U}}_{(i),l,t}, \hat{\mathbf{V}}_{(i),l,t}$ are the optimal solutions to the subproblem (8). By Lemma 7, one can replace $\hat{\mathbf{U}}_{g,t}, \hat{\mathbf{V}}_{(i),g,t}, \hat{\mathbf{U}}_{(i),l,t}, \hat{\mathbf{V}}_{(i),l,t}$ by $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$, and rewrite $\hat{\mathbf{L}}_{(i),t}$ as,

$$\hat{\mathbf{L}}_{(i),t} = \hat{\mathbf{H}}_{g,t}\hat{\mathbf{H}}^T_{g,t}\hat{\mathbf{M}}_{(i)} + \hat{\mathbf{H}}_{(i),l,t}\hat{\mathbf{H}}^T_{(i),l,t}\hat{\mathbf{M}}_{(i)}.$$

Then, we can replace $\hat{\mathbf{H}}_{g,t}$ and $\hat{\mathbf{H}}_{(i),l,t}$ by the Taylor-like series described in Lemma 8. The error between ${\mathbf{L}^\star}_{(i)}$ and $\hat{\mathbf{L}}_{(i),t}$ can be written as the summation of a few terms. The leading term is ${\mathbf{H}^\star}_g{\mathbf{H}^\star}^T_g{\mathbf{L}^\star}_{(i)} + {\mathbf{H}^\star}_{(i),l}{\mathbf{H}^\star}^T_{(i),l}{\mathbf{L}^\star}_{(i)}$, which is identical to ${\mathbf{L}^\star}_{(i)}$ because of the SVD (3). The remaining terms are errors resulting from $\mathcal{O}(\sqrt{\alpha})$ terms in (15) and $\mathbf{E}_{(i)}$. Each of the error terms possesses a special structure that allows us to derive an upper bound on its ℓ_∞ norm. By summing up all these bounds, we can show that $\left\|\hat{\mathbf{L}}_{(i),t} - {\mathbf{L}^\star}_{(i)}\right\|_\infty \leq \mathcal{O}(\sqrt{\alpha}\max_j \left\|\mathbf{E}_{(j),t}\right\|_\infty)$. The detailed calculations on the upper bounds on the ℓ_∞ norm of error terms are long and repetitive, thus relegated to the proof of Lemma 20 in the Appendix.

7. Numerical Experiments

In this section, we investigate the numerical performance of `TCMF` on several datasets. We first use synthetic datasets to verify the convergence in Theorem 5 and validate `TCMF`'s capability in recovering the common and individual features from noisy observations. Then, we use two examples of noisy video segmentation and anomaly detection to illustrate the utility of common, unique, and noise components. We implement Algorithm 1 with HMF (Shi et al., 2023) as its subroutine `JIMF`. Experiments in this section are performed on a desktop with 11th Gen Intel(R) i7-11700KF and NVIDIA GeForce RTX 3080. Code is available in the linked Github repository.

7.1 Exact Recovery on Synthetic Data

On the synthetic dataset, we simulate the data generation process in (1). We use $N = 100$ sources and set the data dimension of $\mathbf{M}_{(i)}$ to 15×1000 in each source. We randomly generate $r_1 = 3$ global features and $r_2 = 3$ local features for each source. The local features are first generated randomly, then deflated to be orthogonal to the global ones. The sparse noise matrix ${\mathbf{S}^\star}_{(i)}$ is randomly generated from the Bernoulli model, i.e., each entry of ${\mathbf{S}^\star}_{(i)}$ is nonzero with probability p and zero with probability $1 - p$. We use p as a proxy of the sparsity parameter $\alpha \triangleq p$. The value of each entry in ${\mathbf{S}^\star}_{(i)}$ is randomly sampled from $\{-100, 100\}$ with equal probability. Next, we use (1) to construct the observation matrix.

With the generated $\{\mathbf{M}_{(i)}\}$, we run Algorithm 1 with $\rho = 0.99$ to estimate local, global, and sparse components. The subroutine `JIMF` in Algorithm 1 is implemented by `HMF` with spectral initialization. As discussed in Appendix A.1, `HMF` is an iterative algorithm. In practice, for each call of `HMF`, we run 500 iterations with constant stepsize 0.005, which take around 62 seconds in our machine and generate satisfactory outputs. To quantitatively evaluate the convergence error, we calculate the ℓ_∞ error of local, global, and sparse components as specified in Theorem 5. More specifically, we calculate the ℓ_∞ global error at

epoch t as

$$\ell_\infty - \text{global error} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{U}}^{\epsilon}{}_{g,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),g,t} - \mathbf{U}^{\star}{}_{(i),g}\mathbf{V}^{\star T}_{(i),g}\right\|_\infty,$$

the ℓ_∞ local error at epoch t as

$$\ell_\infty - \text{local error} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),l,t} - \mathbf{U}^{\star}{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}\right\|_\infty,$$

and the ℓ_∞ sparse noise error at epoch t as

$$\ell_\infty - \text{sparse error} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{S}}_{(i),t} - \mathbf{S}^{\star}{}_{(i)}\right\|_\infty.$$

We show the error plot for three different sparsity parameters α in Figure 2.

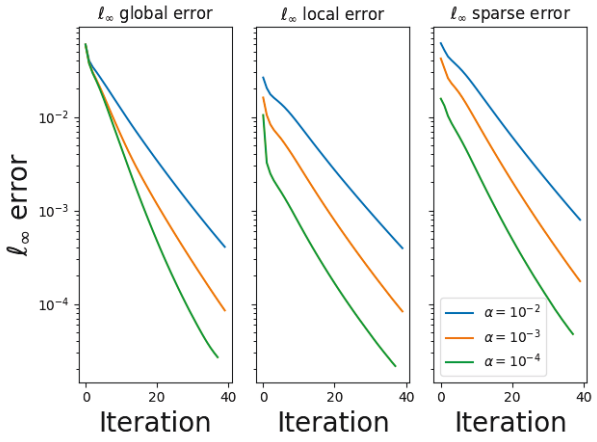


Figure 2: Error plots of Algorithm 1. The x-axis denotes the iteration index, and the y-axis shows the ℓ_∞ error at the corresponding iteration. The y-axis is in log scale.

From Figure 2, it is clear that the global, local, and sparse components indeed converge linearly to the ground truth.

Further, we compare the feature extraction performance of `TCMF` with benchmark algorithms, including JIVE (Lock et al., 2013), RJIVE (Sagonas et al., 2017), RaJIVE (Ponzi et al., 2021), and HMF (Shi et al., 2023). We do not include the comparison with RCICA (Panagakis et al., 2015) because RCICA is designed only for $N = 2$, while we have 100 different sources. Since the errors of different methods vary drastically, we calculate and report the logarithm of global error as $\texttt{g-error} = \log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{U}}^{\epsilon}{}_{g,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),g,t} - \mathbf{U}^{\star}{}_{(i),g}\mathbf{V}^{\star T}_{(i),g}\right\|_F^2\right)$, the logarithm of local error as $\texttt{l-error} = \log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{U}}^{\epsilon}{}_{(i),l,t}\hat{\mathbf{V}}^{\epsilon T}_{(i),l,t} - \mathbf{U}^{\star}{}_{(i),l}\mathbf{V}^{\star T}_{(i),l}\right\|_F^2\right)$, and the logarithm of sparse noise error as $\texttt{s-error} = \log_{10}\left(\sum_{i=1}^{N}\left\|\hat{\mathbf{S}}_{(i),t} - \mathbf{S}^{\star}{}_{(i)}\right\|_F^2\right)$ at

$t = 20$. We run experiments from 5 different random seeds and calculate the mean and standard deviation of the log errors. Results are reported in Table 1.

Table 1: Recovery error of different algorithms. The columns `g-error`, `l-error`, and `s-error` stand for the log recovery errors of global components, local components, and sparse components.

	$\alpha = 0.01$			$\alpha = 0.1$		
	`g-error`	`l-error`	`s-error`	`g-error`	`l-error`	`s-error`
JIVE	5.52 ± 0.01	5.64 ± 0.01	-	6.52 ± 0.01	6.58 ± 0.01	-
HMF	5.49 ± 0.01	5.62 ± 0.01	-	6.48 ± 0.01	6.55 ± 0.01	-
RaJIVE	5.46 ± 0.01	5.36 ± 0.05	5.71 ± 0.05	6.48 ± 0.00	6.25 ± 0.14	6.59 ± 0.12
RJIVE	5.49 ± 0.01	5.44 ± 0.01	5.77 ± 0.01	6.48 ± 0.00	6.47 ± 0.00	6.78 ± 0.00
`TCMF`	**-3.38** ± 0.14	**-3.37** ± 0.13	**-2.94** ± 0.08	**-1.93** ± 0.09	**-1.95** ± 0.06	**-1.54** ± 0.04

Table 1 shows that `TCMF` outperforms benchmark algorithms by several orders. This is understandable as `TCMF` is provably convergent into the ground truth, while benchmark algorithms either neglect sparse noise or rely on instance-dependent heuristics.

7.2 Video Segmentation from Noisy Frames

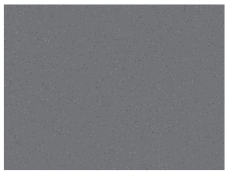
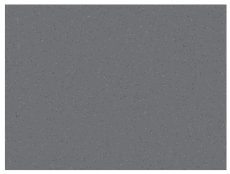
An important task in video segmentation is background-foreground separation. There are several matrix factorization algorithms that can achieve decent performance in video segmentation, including robust PCA (Candès et al., 2011), PerPCA (Shi and Kontar, 2024), and HMF (Shi et al., 2023). However, the separation is much more challenging when the videos are corrupted by large noise (Shen et al., 2022). `TCMF` can naturally handle such tasks with its power to recover global and local components from highly noisy measurements.

We use a surveillance video from Vacavant et al. (2013) as an example. In the video, multiple vehicles drive through the circle. We add large and sparse noise to the frames to simulate the effects of large measurement errors. More specifically, similar to 7.1, we sample each entry of noise from i.i.d. Bernoulli distribution that is zero with probability 0.99 and nonzero with probability 0.01. And each entry is sampled from $\{-500, 500\}$ with equal probability. Then we apply `TCMF` on the noisy frames to recover $\hat{\mathbf{U}}_g, \{\hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l}, \hat{\mathbf{S}}_{(i)}\}$. We set ρ to 0.95 and use $T = 15$ epochs. The subroutine `JIMF` is still implemented by `HMF` with spectral initialization. We also set the number of iterations for `HMF` to 500. This is a conservative choice to ensure small optimization error ϵ in the subroutine. To visualize the results, we plot global components $\hat{\mathbf{U}}_g \hat{\mathbf{V}}_{(i),g}^T$ and local components $\hat{\mathbf{U}}_{(i),l} \hat{\mathbf{V}}_{(i),l}^T$. They are shown in Table 2.

In Table 2, the background and foreground are clearly separated from the noise. The result highlights `TCMF`'s ability to extract features in high-dimensional noisy data.

We compare `TCMF` to several benchmark methods, namely JIVE, HMF, RJIVE, RaJIVE, and Robust PCA. These algorithms, including JIVE, HMF, RJIVE, and RaJIVE, are capable of producing joint and individual components of video frames. In our evaluation, we consider the joint component as the background and the individual component as the foreground. As for robust PCA, we flatten each image into a row vector and create a large matrix $\mathbf{M}_{\text{stack}}$ by stacking these row vectors. We then utilize the nonconvex robust PCA (Netrapalli et al.,

Table 2: Foreground Background separation

Frame	1	2	3
Original noisy frames			
Noise			
Global components			
Local components			

2014) to extract the sparse and low-rank components from $\mathbf{M}_{\text{stack}}$. The low-rank component is regarded as the background, while the sparse component captures the foreground.

To assess the performance of these methods, we calculate the differences between the recovered background and foreground compared to the ground truths. Specifically, we estimate the mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) of the recovered foreground and background with respect to the true foreground and background. The comparison results are presented in Table 3.

Table 3: Background and foreground recovery quality metrics for different algorithms.

	Background			Foreground			Wall-clock
	MSE ↓	PSNR ↑	SSIM ↑	MSE ↓	PSNR ↑	SSIM ↑	time (s) ↓
JIVE	415	-26	0.08	2521	14	0.03	1.6×10^3
HMF	198	-22	0.18	2413	14	0.05	2.3×10^1
PerPCA	236	-23	0.14	2389	14	0.07	9.8×10^1
RJIVE	277	-24	0.13	1309	16	0.22	9.2×10^1
RaJIVE	170	-22	0.22	166	26	0.18	1.2×10^4
Robust PCA	0.0016	31	**1.00**	5105	11	0.61	$\mathbf{3.3} \times 10^{-1}$
`TCMF`	**0.0003**	**33**	**1.00**	**98**	**31**	**0.98**	3.5×10^2

In Table 3, a lower MSE, a higher PSNR, and a higher SSIM signify superior recovery quality. In terms of background recovery, both `TCMF` and robust PCA exhibit low MSE, high PSNR, and high SSIM, surpassing other methods. This suggests that both algorithms effectively reconstruct the background. This outcome was anticipated as `TCMF` and robust PCA possess the capability to differentiate between significant noise and low-rank components. In contrast, other benchmarks either neglect large noise in the model or rely on heuristics. Furthermore, `TCMF` showcases marginally superior performance in MSE and PSNR compared to robust PCA, signifying higher-quality background recovery.

When it comes to foreground recovery, `TCMF` outperforms benchmark algorithms significantly across all metrics. The inability of robust PCA to achieve high-quality foreground recovery is likely due to its inability to separate sparse noise from the foreground. JIVE and HMF yield high MSE and low PSNR, indicating noisy foreground reconstruction. Although heuristic methods, such as RJIVE and RaJIVE, exhibit slight performance improvements over JIVE and HMF, they still fall short of the performance exhibited by `TCMF`. This comparison underscores `TCMF`'s remarkable power to identify unique components from sparse noise accurately.

We also report the running time of each experiment in Table 3. Compared with heuristic methods to robustly separate the shared and unique components, `TCMF` exhibits a slightly longer running time than RJIVE but significantly outperforms RaJIVE in terms of speed. The comparison highlights `TCMF`'s superior performance with moderate computation demands. Although Robust PCA demonstrates a relatively short running time in this instance, larger-scale experiments presented in Appendix B will show that Robust PCA has larger running time scaling as the problem size increases.

7.3 Case Study: Defect Detection on Steel Surface

Hot rolling is an important process in steel manufacturing. For better product quality, a critical task is to detect and locate the defects that arise in the rolling process (Jin et al., 2000). In this study, the dataset (Jin et al., 2000; Yan et al., 2018) comes from the HotEye video of a rolling steel plate. The video captures sharp pictures of the surface of the steel plate. An example is shown in the left graph of Figure 3. The irregular dark dots in the graph indicate surface defects that require subsequent investigations (Jin et al., 2000).

As different frames of the rolling video are related, they possess similar background patterns. Meanwhile, each frame also contains unique variations that reflect frame-by-frame differences. On top of the changing patterns, there are small defects on the surface of the steel plate. The defects, as shown in the left graph of Figure 3, only occupy small spatial regions and thus can naturally be modeled by sparse outliers.

In such scenarios, the application of `TCMF` enables the identification of defects and extraction of common and unique patterns simultaneously. For this experiment, we use `TCMF` to segment 100 hot-rolling video frames. The right graph of Figure 3 illustrates two frames selected from the rolling video alongside the corresponding recovered global, local, and sparse components. We set the reduction parameter $\rho = 0.97$ and the number of epochs $T = 100$. The details of the subroutine `JIMF` are relegated to Appendix A.1. Additionally, as a comparative analysis, we employ nonconvex robust PCA (Netrapalli et al., 2014) to recover and display the low-rank and sparse components from frames. Our robust PCA

implementation alternatively applies SVD and hard thresholding. We require all entries in the sparse component to be negative in the hard-thresholding step to encode the domain knowledge that surface defects tend to have lower temperatures. The hyper-parameters for SVD and thresholding are consistent for both `TCMF` and Robust PCA.

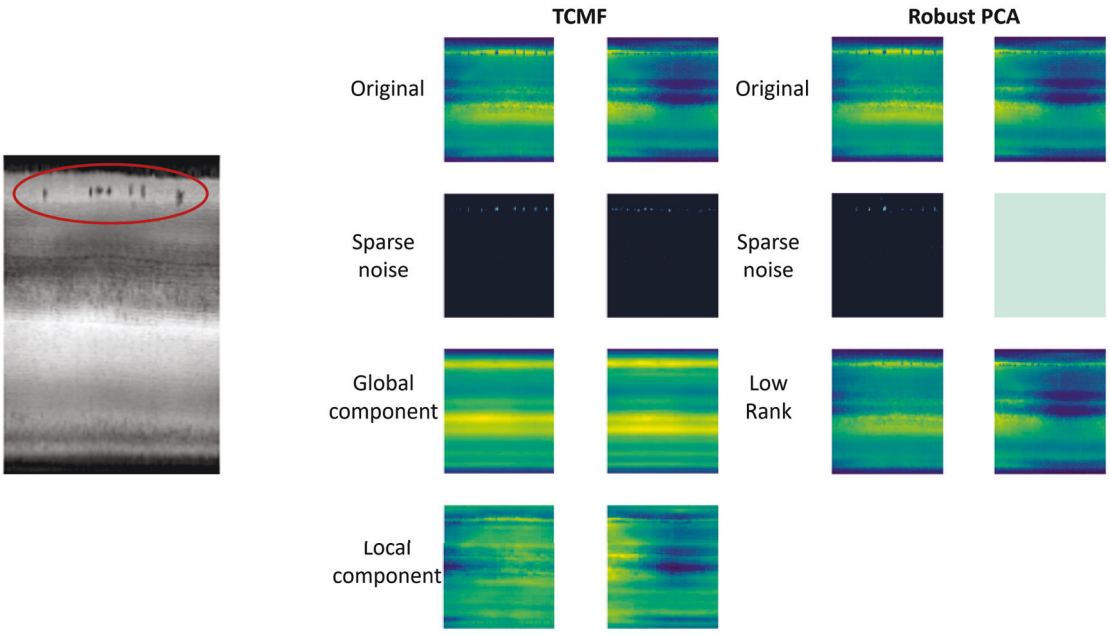


Figure 3: *Left*: An example of the surface of the steel bar. There are a few anomalies inside the red ellipse. *Right*: Recovered sparse noises, shared components, and unique components from 2 frames.

In Figure 3, we can see that `TCMF` effectively identifies the small defects on the steel plate surface. The global component reflects the general patterns in the video frames, while the local component accentuates the variations in different frames. In contrast, the sparse components recovered by robust PCA do not faithfully represent the surface defects.

We proceed to show that `TCMF`-recovered sparse components can be conveniently leveraged for frame-level anomaly detection. Our task here is to identify which frames contain surface anomalies. Inspired by the statistics-based anomaly detection (Chandola et al., 2009), we construct simple test statistics to monitor the anomalies. The test statistics is defined as the ℓ_1 norm of the recovered sparse noise on each frame $\left\|\hat{\mathbf{S}}_{(i)}\right\|_1$. Indeed, a large $\left\|\hat{\mathbf{S}}_{(i)}\right\|_1$ provides strong evidence for surface defects. The choice of ℓ_1 norm is not special as we find other norms, such as ℓ_2 norm, would yield a similar performance.

After using `TCMF` to extract the sparse components, we calculate the test statistics for each frame. Then, we can set up a simple threshold-based classification rule for anomaly detection: when the ℓ_1 norm exceeds the threshold, we report an anomaly in the corresponding frame. In the case study, the threshold is set to be the highest value in the first 50 frames, which is the in-control group that does not contain anomalies (Yan et al., 2018). We plot the test statistics and thresholds in Figure 5. The blue dots and red crosses denote the (ground truth)

normal and abnormal frame labels in Yan et al. (2018). In an ideal plot of test statistics, one would expect the abnormal samples to have higher ℓ_1 norms, while normal samples should have lower norms. This is indeed the case for Figure 5, where a simple threshold based on the sparse features can distinguish abnormal samples from normal ones with high accuracy.

In comparison, we also calculate the ℓ_1 norm of sparse noise recovered by robust PCA and plot the testing statistics in Figure 4. In Figure 4, the ℓ_1 norm is less indicative of anomaly labels, as some abnormal samples have small test statistics, while some normal samples have large statistics. It is also hard to use a threshold on the test statistics to predict anomalies.

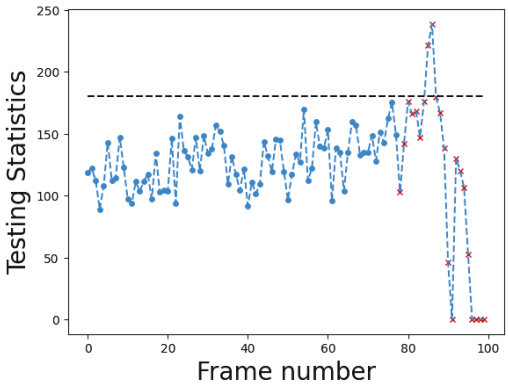


Figure 4: Test statistics of robust PCA.

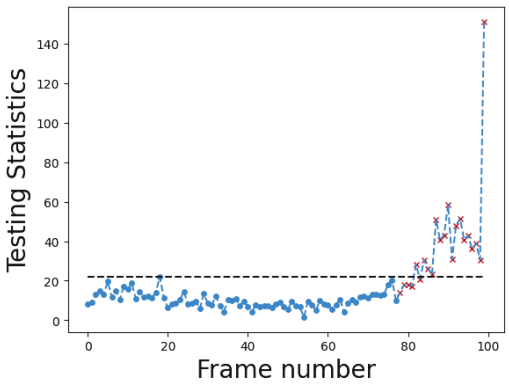


Figure 5: Test statistics of `TCMF`.

The comparison highlights `TCMF`'s ability to find surface defects. The results are understandable as `TCMF` uses a more refined model to decompose the thermal frames into three parts, thus having more representation power to fit the underlying physics in the manufacturing process. As a result, the recovered sparse components are more representative of the anomalies.

8. Conclusion

In this work, we propose a systematic method `TCMF` to separate shared, unique, and noise components from noisy observation matrices. `TCMF` is the first algorithm that is provably convergent to the ground truth under identifiability conditions that require the three components to have "small overlaps". `TCMF` outperforms previous heuristic algorithms by large margins in numerical experiments and finds interesting applications in video segmentation, anomaly detection, and time series imputation.

Our work also opens up several venues for future theoretical exploration in separating shared and unique low-rank features from noisy matrices. For example, a minimax lower bound on the μ, θ, and α can help fathom the statistical difficulty of such separation. Also, as many existing methods for `JIMF` rely on good initialization to excel, designing efficient algorithms for `JIMF` that are independent of the initialization is also an interesting topic. On the practical side, methods to integrate `TCMF` with other machine learning models, e.g., auto-encoders, to find nonlinear features in data are worth exploring.

Acknowledgments

This research is supported in part by Raed Al Kontar's NSF CAREER Award 2144147 and Salar Fattahi's NSF CAREER Award CCF-2337776, NSF Award DMS-2152776, and ONR Award N00014-22-1-2127,

Appendix A. Details of subroutine algorithm

In this section, we will elaborate on two subroutine algorithms in literature to solve the problem (8), specifically known as `HMF` and `PerPCA`. Among a plethora of existing methodologies aiming to distinguish shared and unique features, these two exhibit an exceptional significance, as they are proved to converge to the optimal resolution of problem (8) linearly under appropriate initial conditions. We underscore that the usage of `JIMF` is not restricted solely to these two methods. In essence, any algorithm with the ability to segregate common and unique components can be effectively employed as `JIMF`.

A.1 Heterogeneous matrix factorization

Heterogeneous matrix factorization (`HMF`) (Shi et al., 2023) is an algorithm proposed to solve the following problem,

$$\begin{aligned}
&\min_{\mathbf{U}_g,\{\mathbf{U}_{(i),l}\}_{i=1,\cdots,N}} \sum_{i=1}^{N} \tilde{f}_i\left(\mathbf{U}_g,\{\mathbf{V}_{(i),g},\mathbf{U}_{(i),l},\mathbf{V}_{(i),l}\}\right)\\
=&\sum_{i=1}^{N}\frac{1}{2}\left\|\hat{\mathbf{M}}_{(i)}-\mathbf{U}_g\mathbf{V}_{(i),g}^T-\mathbf{U}_{(i),l}\mathbf{V}_{(i),l}^T\right\|_F^2+\frac{\beta}{2}\left\|\mathbf{U}_g^T\mathbf{U}_g-\mathbf{I}\right\|_F^2+\frac{\beta}{2}\left\|\mathbf{U}_{(i),l}^T\mathbf{U}_{(i),l}-\mathbf{I}\right\|_F^2\\
\text{subject to }&\mathbf{U}_g^T\mathbf{U}_{(i),l}=\mathbf{0},\ \forall i.
\end{aligned}\tag{16}$$

Compared with (8), the objective in (16) contains two additional regularization terms $\frac{\beta}{2}\left\|\mathbf{U}_g^T\mathbf{U}_g-\mathbf{I}\right\|_F^2+\frac{\beta}{2}\left\|\mathbf{U}_{(i),l}^T\mathbf{U}_{(i),l}-\mathbf{I}\right\|_F^2$. The regularization terms enhance the smoothness of the optimization objective thereby facilitating convergence. Despite the regularization terms, any optimal solution to (16) is also an optimal solution to (8). We can prove the claim in the following proposition.

Proposition 9 *Let* $\hat{\mathbf{U}}_g^{HMF},\{\hat{\mathbf{V}}_{(i),g}^{HMF},\hat{\mathbf{U}}_{(i),l}^{HMF},\hat{\mathbf{V}}_{(i),l}^{HMF}\}$ *be one set of optimal solutions to* (16)*, then* $\hat{\mathbf{U}}_g^{HMF},\{\hat{\mathbf{V}}_{(i),g}^{HMF},\hat{\mathbf{U}}_{(i),l}^{HMF},\hat{\mathbf{V}}_{(i),l}^{HMF}\}$ *is also a set of optimal solution to* (8)

Proof The proof is straightforward. We first claim that $\hat{\mathbf{U}}_g^{\mathtt{HMF}T}\hat{\mathbf{U}}_g^{\mathtt{HMF}}=\mathbf{I}$ and $\hat{\mathbf{U}}_{(i),l}^{\mathtt{HMF}\,T}\hat{\mathbf{U}}_{(i),l}^{\mathtt{HMF}}=\mathbf{I}$. We prove the claim by contradiction. Suppose otherwise, we can find a QR decomposition of $\hat{\mathbf{U}}_g^{\mathtt{HMF}}$ and $\hat{\mathbf{U}}_{(i),l}^{\mathtt{HMF}}$ as $\hat{\mathbf{U}}_g^{\mathtt{HMF}}=\mathbf{Q}_g\mathbf{R}_g$ and $\hat{\mathbf{U}}_g^{\mathtt{HMF}}=\mathbf{Q}_{(i),l}\mathbf{R}_{(i),l}$, where $\mathbf{Q}_g$ and $\mathbf{Q}_{(i),l}$'s are orthonormal and $\mathbf{R}_g$ and $\mathbf{R}_{(i),l}$'s are upper-triangular. Furthermore, not both $\mathbf{R}_g$ and $\mathbf{R}_{(i),l}$ are identity matrices, thus $\left\|\mathbf{R}_g^T\mathbf{R}_g-\mathbf{I}\right\|_F^2+\left\|\mathbf{R}_{(i),l}^T\mathbf{R}_{(i),l}-\mathbf{I}\right\|_F^2>0$. Now, we construct a refined set of solutions as,

$$\begin{aligned}
\hat{\mathbf{U}}_g^{\mathtt{HMF},refined}&=\mathbf{Q}_g\\
\hat{\mathbf{V}}_{(i),g}^{\mathtt{HMF},refined}&=\hat{\mathbf{V}}_{(i),g}^{\mathtt{HMF}}\mathbf{R}_g^T\\
\hat{\mathbf{U}}_{(i),l}^{\mathtt{HMF},refined}&=\mathbf{Q}_{(i),l}\\
\hat{\mathbf{V}}_{(i),l}^{\mathtt{HMF},refined}&=\hat{\mathbf{V}}_{\mathtt{HMF}}^{(i),l}\mathbf{R}_{(i),l}^T.
\end{aligned}$$

Then it's easy to verify that

$$\begin{aligned}
&\sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF},refined}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF},refined}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF},refined}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF},refined} \right) \\
&= \sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right) - \frac{\beta}{2} \left(\left\| \mathbf{R}_g^T \mathbf{R}_g - \mathbf{I} \right\|_F^2 + \left\| \mathbf{R}_{(i),l}^T \mathbf{R}_{(i),l} - \mathbf{I} \right\|_F^2 \right) \\
&< \sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right),
\end{aligned}$$

which contradicts with the global optimality of $\hat{\mathbf{U}}_g^{\texttt{HMF}}, \{\hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}}\}$. This proves the claim.

From the orthogonality, we know $f_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right) = \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right)$.

Now suppose $\hat{\mathbf{U}}_g^{\texttt{HMF}}, \{\hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}}\}$ is not an optimal solution to (8). Then, we can find a different set of feasible solution $\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \{\hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}}\}$ such that

$$\begin{aligned}
&\sum_{i=1}^{N} f_i \left(\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}} \right) \\
&< \sum_{i=1}^{N} f_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right) \\
&= \sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right).
\end{aligned}$$

We can similarly define a set of refined solutions

$$\begin{aligned}
\hat{\mathbf{U}}_g^{\texttt{JIMF},refined} &= \mathbf{Q}_g^{\texttt{JIMF}} \\
\hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF},refined} &= \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}} \mathbf{R}_g^{\texttt{JIMF}T} \\
\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF},refined} &= \mathbf{Q}_{(i),l}^{\texttt{JIMF}} \\
\hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF},refined} &= \hat{\mathbf{V}}_{\texttt{HMF}}^{(i),l} \mathbf{R}_{(i),l}^{\texttt{JIMF}T},
\end{aligned}$$

where $\mathbf{Q}_g^{\texttt{JIMF}}$, $\mathbf{R}_g^{\texttt{JIMF}}$, $\mathbf{Q}_{(i),l}^{\texttt{JIMF}}$, $\mathbf{R}_{(i),l}^{\texttt{JIMF}}$ are QR decompositions that satisfy $\hat{\mathbf{U}}_g^{\texttt{JIMF},refined} = \mathbf{Q}_g^{\texttt{JIMF}} \mathbf{R}_g^{\texttt{JIMF}}$ and $\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF},refined} = \mathbf{Q}_{(i),l}^{\texttt{JIMF}} \mathbf{R}_{(i),l}^{\texttt{JIMF}}$. Based on the refined set of solutions, we

can prove that,

$$\begin{aligned}
&\sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{JIMF},refined}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF},refined}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF},refined}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF},refined} \right) \\
&= \sum_{i=1}^{N} f_i \left(\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}} \right) \\
&< \sum_{i=1}^{N} \tilde{f}_i \left(\hat{\mathbf{U}}_g^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}} \right),
\end{aligned}$$

which contradicts the optimality of $\hat{\mathbf{U}}_g^{\texttt{HMF}}, \{\hat{\mathbf{V}}_{(i),g}^{\texttt{HMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{HMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{HMF}}\}$.

This completes the proof. ■

`HMF` optimizes the objective by gradient descent. To ensure feasibility, `HMF` employs a special correction step to orthogonalize $\mathbf{U}_g$ and $\mathbf{U}_{(i),l}$ without changing the objective at every step. The pseudo-code is presented in Algorithm 2.

Algorithm 2 `JIMF` by heterogeneous matrix factorization

Input matrices $\{\hat{\mathbf{M}}_{(i)}\}_{i=1}^N$, stepsize η_τ, iteration budget R.
Initialize $\mathbf{U}_{g,1}, \mathbf{V}_{(i),g,\frac{1}{2}}, \mathbf{U}_{(i),l,\frac{1}{2}}, \mathbf{V}_{(i),l,1}$ to be small random matrices.
for Iteration $\tau = 1, ..., R$ **do**
  for index $i = 1, \cdots, N$ **do**
    Correct $\mathbf{U}_{(i),l,\tau} = \mathbf{U}_{(i),l,\tau-\frac{1}{2}} - \mathbf{U}_{g,\tau} \left(\mathbf{U}_{g,\tau}^T \mathbf{U}_{g,\tau}\right)^{-1} \mathbf{U}_{g,\tau}^T \mathbf{U}_{(i),l,\tau-\frac{1}{2}}$
    Correct $\mathbf{V}_{(i),g,\tau} = \mathbf{V}_{(i),g,\tau-\frac{1}{2}} + \mathbf{V}_{(i),l,\tau} \mathbf{U}_{(i),l,\tau-\frac{1}{2}}^T \mathbf{U}_{g,\tau} \left(\mathbf{U}_{g,\tau}^T \mathbf{U}_{g,\tau}\right)^{-1}$
    Update $\mathbf{U}_{(i),g,\tau+1} = \mathbf{U}_{g,\tau} - \eta_\tau \nabla_{\mathbf{U}_g} \tilde{f}_i$
    Update $\mathbf{V}_{(i),g,\tau+\frac{1}{2}} = \mathbf{V}_{(i),g,\tau} - \eta_\tau \nabla_{\mathbf{V}_{(i),g}} \tilde{f}_i$
    Update $\mathbf{U}_{(i),l,\tau+\frac{1}{2}} = \mathbf{U}_{(i),l,\tau} - \eta_\tau \nabla_{\mathbf{U}_{(i),l}} \tilde{f}_i$
    Update $\mathbf{V}_{(i),l,\tau+1} = \mathbf{V}_{(i),l,\tau} - \eta_\tau \nabla_{\mathbf{V}_{(i),l}} \tilde{f}_i$
  end for
  Calculates $\mathbf{U}_{g,\tau+1} = \frac{1}{N} \sum_{i=1}^N \mathbf{U}_{(i),g,\tau+1}$
end for
Return $\mathbf{U}_{g,R}, \{\mathbf{V}_{(i),g,R}\}, \{\mathbf{U}_{(i),l,R}\}, \{\mathbf{V}_{(i),l,R}\}$.

In Algorithm 2, we use τ to denote the iteration index, where the half-integer index denotes the update of the variable is half complete: it is updated by gradient descent but is not feasible yet. It is proven that under a group of sufficient conditions, Algorithm 2 converges to the optimal solutions of problem (16). The sufficient conditions require the stepsize η_τ to be chosen appropriately and the initialization close to the optimal solution (Shi et al., 2023).

In practice, Algorithm 2 is often efficient and accurate. Therefore, we implement `HMF` as the subroutine `JIMF` for all of our numerical simulations in Section 7. To initialize Algorithm 2, we adopt a spectral initialization approach. Specifically, we concatenate all

matrices column-wise to form $\mathbf{M}^{concat} = [\mathbf{M}_{(1)}, \mathbf{M}_{(2)}, \cdots, \mathbf{M}_{(N)}]$. Subsequently, we perform a Singular Value Decomposition (SVD) on the concatenated matrix $\mathbf{M}^{concat}$ to extract the top r_1 column singular vectors, which serve as the initialization for $\mathbf{U}_{g,1}$ in Algorithm 2. Utilizing the calculated $\mathbf{U}_{g,1}$, we deflate $\mathbf{M}_{(i)}$ by subtracting the projection of $\mathbf{M}_{(i)}$ onto $\mathbf{U}_{g,1}$, denoted as $\mathbf{M}_{(i)}^{deflate} = \mathbf{M}_{(i)} - \mathbf{U}_{g,1}\mathbf{U}_{g,1}^T\mathbf{M}_{(i)}$. We then conduct another SVD to identify the top r_2 singular vectors of $\mathbf{M}_{(i)}^{deflate}$, which are utilized as the initialization for $\mathbf{U}_{(i),l,\frac{1}{2}}$. The initializations for the coefficient matrices are established as $\mathbf{V}_{(i),g,\frac{1}{2}} = \mathbf{M}_{(i)}^T\mathbf{U}_{g,1}$ and $\mathbf{V}_{(i),l,1} = \mathbf{M}_{(i)}^T\mathbf{U}_{(i),l,\frac{1}{2}}$.

The stepsize η in Algorithm 2 is individually adjusted for each dataset to achieve the fastest convergence. We choose a large total number of iterations R to ensure a small optimization error ϵ. Specifically, in the synthetic data, we set the stepsize to 0.005 and $R = 500$. In the video segmentation task, we set the stepsize to 5×10^{-6} and $R = 200$. And on the hot rolling data, we set the stepsize to 4×10^{-5} and $R = 500$. In our experiments, we observe that the regularization parameter β exerts a negligible influence on the convergence of Algorithm 2. Consequently, we maintain β within the range of 10^{-6} to 10^{-5} in all our experiments.

A.2 Personalized PCA

Personalized PCA (Shi and Kontar, 2024) is another subroutine to solve (8). More specifically, personalized PCA seeks to find orthonormal features $\mathbf{U}_g$ and $\mathbf{U}_{(i),l}$ to minimize the residual of fitting, as shown in the following objective,

$$\begin{aligned} \min_{\mathbf{U}_g, \{\mathbf{U}_{(i),l}\}_{i=1,\cdots,N}} & \frac{1}{2}\sum_{i=1}^{N}\left\|\hat{\mathbf{M}}_{(i)} - \mathbf{U}_g\mathbf{U}_g^T\hat{\mathbf{M}}_{(i)} - \mathbf{U}_{(i),l}\mathbf{U}_{(i),l}^T\hat{\mathbf{M}}_{(i)}\right\|_F^2 \\ \text{subject to } & \mathbf{U}_g^T\mathbf{U}_g = \mathbf{I},\ \mathbf{U}_{(i),l}^T\mathbf{U}_{(i),l} = \mathbf{I},\ \mathbf{U}_g^T\mathbf{U}_{(i),l} = \mathbf{0},\ \forall i. \end{aligned} \tag{17}$$

The objective only optimizes the feature matrices $\mathbf{U}_g$ and $\mathbf{U}_{(i),l}$, but it's essentially equivalent to problem (8). The formal statement is presented in the following proposition.

Proposition 10 *Let* $\hat{\mathbf{U}}_g^{PerPCA}, \{\hat{\mathbf{U}}_{(i),l}^{PerPCA}\}$ *be one set of optimal solutions to* (17)*, then* $\hat{\mathbf{U}}_g^{PerPCA}, \{\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g^{PerPCA}, \hat{\mathbf{U}}_{(i),l}^{PerPCA}, \hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}^{PerPCA}\}$ *is also a set of optimal solution to* (8)

Proof We will also prove the proposition by contradiction. If $\hat{\mathbf{U}}_g^{\texttt{PerPCA}}, \{\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g^{\texttt{PerPCA}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{PerPCA}}, \hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}^{\texttt{PerPCA}}\}$ is not a set of optimal solution to (8), we can find a different set of feasible solutions $\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \{\hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}}\}$ such that

$$\begin{aligned} &\sum_{i=1}^{N} f_i\left(\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}}\right) \\ &< \sum_{i=1}^{N} f_i\left(\hat{\mathbf{U}}_g^{\texttt{PerPCA}}, \hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g^{\texttt{PerPCA}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{PerPCA}}, \hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}^{\texttt{PerPCA}}\right) \\ &= \sum_{i=1}^{N}\left\|\hat{\mathbf{M}}_{(i)} - \mathbf{U}_g^{\texttt{PerPCA}}\mathbf{U}_g^{\texttt{PerPCA}T}\hat{\mathbf{M}}_{(i)} - \mathbf{U}_{(i),l}^{\texttt{PerPCA}}\mathbf{U}_{(i),l}^{\texttt{PerPCA}T}\hat{\mathbf{M}}_{(i)}\right\|_F^2. \end{aligned}$$

If we fix $\mathbf{U}_g$ and $\mathbf{U}_{(i),l}$ to be $\hat{\mathbf{U}}_g^{\texttt{JIMF}}$ and $\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}$ in problem (8), then the optimal solution of $\mathbf{V}_{(i),g}$ and $\mathbf{V}_{(i),l}$ is $\mathbf{V}_{(i),g}^{\texttt{JIMF},opt} = \hat{\mathbf{M}}_{(i)}^T \hat{\mathbf{U}}_g^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_g^{\texttt{JIMF}T} \hat{\mathbf{U}}_g^{\texttt{JIMF}}\right)^{-1}$ and $\mathbf{V}_{(i),l}^{\texttt{JIMF},opt} = \hat{\mathbf{M}}_{(i)}^T \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}T} \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}\right)^{-1}$. As a result,

$$
\begin{aligned}
&\sum_{i=1}^{N} \left\| \hat{\mathbf{M}}_{(i)} - \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}T} \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}\right)^{-1} \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}T} \hat{\mathbf{M}}_{(i)} - \hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}T} \hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}}\right)^{-1} \hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}T} \hat{\mathbf{M}}_{(i)} \right\|_F^2 \\
&= \sum_{i=1}^{N} f_i \left(\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF},opt}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF},opt}\right) \\
&\leq \sum_{i=1}^{N} f_i \left(\hat{\mathbf{U}}_g^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),g}^{\texttt{JIMF}}, \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}, \hat{\mathbf{V}}_{(i),l}^{\texttt{JIMF}}\right) \\
&< \sum_{i=1}^{N} \left\| \hat{\mathbf{M}}_{(i)} - \mathbf{U}_g^{\texttt{PerPCA}} \mathbf{U}_g^{\texttt{PerPCA}T} \hat{\mathbf{M}}_{(i)} - \mathbf{U}_{(i),l}^{\texttt{PerPCA}} \mathbf{U}_{(i),l}^{\texttt{PerPCA}T} \hat{\mathbf{M}}_{(i)} \right\|_F^2 .
\end{aligned}
$$

If we define $\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF},refine} = \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}T} \hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF}}\right)^{-1/2}$ and $\hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF},refine} = \hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}} \left(\hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}T} \hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF}}\right)^{-1/2}$, then $\hat{\mathbf{U}}_{(i),l}^{\texttt{JIMF},refine}$ and $\hat{\mathbf{U}}_{(i),g}^{\texttt{JIMF},refine}$ are also feasible for (17) and achieve lower objective. This contradicts the optimality of $\mathbf{U}_g^{\texttt{PerPCA}}$ and $\mathbf{U}_{(i),l}^{\texttt{PerPCA}}$.

■

To solve the constrained optimization problem (17), personalized PCA adopts a distributed version of Stiefel gradient descent. The pseudo-code is presented in Algorithm 3.

In Algorithm 3, $\mathcal{GR}$ denotes generalized retraction. In practice, it can be implemented via polar projection $\mathcal{GR}_{\mathbf{U}}(\mathbf{V}) = (\mathbf{U}+\mathbf{V})\left(\mathbf{U}^T\mathbf{U} + \mathbf{V}^T\mathbf{U} + \mathbf{U}^T\mathbf{V} + \mathbf{V}^T\mathbf{V}\right)^{-\frac{1}{2}}$. Algorithm 3 can also be proved to converge to the optimal solutions with suitable choices of stepsize and initialization (Shi and Kontar, 2024).

Appendix B. Additional running time comparisons

In this section, we include the additional running time comparison between `TCMF` and Robust PCA. We use a set of synthetic datasets with varying numbers of sources N, then compare the per-iteration running time of the two algorithms. More specifically, we follow the setting in Section 7.1 where $n_1 = 15$ and $n_2 = 1000$, and generate synthetic datasets where the number of sources N changes from 100 to 10000. Then, we apply `TCMF` and Robust PCA on the same dataset. We do not parallelize computations for either algorithm for fair comparison. The per-iteration running time of the two algorithms is collected and plotted in Figure 6.

Algorithm 3 `JIMF` by personalized PCA

Input observation matrices $\{\hat{\mathbf{M}}_{(i)}\}_{i=1}^{N}$, stepsize η_τ, iteration budget R.
Initialize $\mathbf{U}_{g,1}$, and $\mathbf{U}_{(1),l,\frac{1}{2}}, \cdots, \mathbf{U}_{(N),l,\frac{1}{2}}$.
Calculate $\mathbf{S}_{(i)} = \hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T$ for each i.
for iteration $\tau = 1, ..., R$ **do**
  for index $i = 1, \cdots, N$ **do**
    Correct $\mathbf{U}_{(i),l,\tau} = \mathcal{GR}_{\mathbf{U}_{(i),l,\tau-\frac{1}{2}}}\left(-\mathbf{U}_{g,\tau}\mathbf{U}_{g,\tau}\mathbf{U}_{(i),l,\tau-\frac{1}{2}}\right)$
    Calculate $\mathbf{G}_{(i),\tau} = \left(\mathbf{I} - \mathbf{U}_{g,\tau}\mathbf{U}_{g,\tau}^T - \mathbf{U}_{(i),l,\tau}\mathbf{U}_{(i),l,\tau}^T\right)\left(\boldsymbol{S}_{(i)}\left[\mathbf{U}_{g,\tau}, \mathbf{U}_{(i),l,\tau}\right]\right)$
    Update $\mathbf{U}_{(i),g,\tau+1} = \mathbf{U}_{g,\tau} + \eta_\tau(\mathbf{G}_{(i),\tau})_{1:d,1:r_1}$
    Update $\mathbf{U}_{(i),l,\tau+\frac{1}{2}} = \mathcal{GR}_{\mathbf{U}_{(i),l,\tau}}\left(\eta_\tau(\mathbf{G}_{(i),\tau})_{1:d,(r_1+1):(r_1+r_{2,(i)})}\right)$
  end for
  Update $\mathbf{U}_{g,\tau+1} = \mathcal{GR}_{\mathbf{U}_{g,\tau}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{U}_{(i),g,\tau+1} - \mathbf{U}_{g,\tau}\right)$
end for
Calculate $\mathbf{V}_{(i),g,R} = \hat{\mathbf{M}}_{(i)}^T\mathbf{U}_{g,R}$ and $\mathbf{V}_{(i),l,R} = \hat{\mathbf{M}}_{(i)}^T\mathbf{U}_{(i),l,R}$.
Return $\mathbf{U}_{g,R}, \{\mathbf{V}_{(i),g,R}\}, \{\mathbf{U}_{(i),l,R}\}, \{\mathbf{V}_{(i),l,R}\}$.

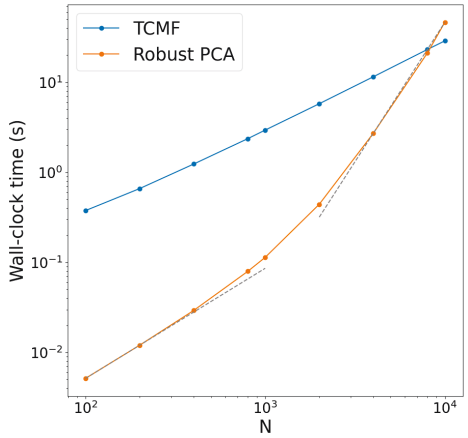


Figure 6: Running time comparison between runtime of `TCMF` and Robuct PCA.

From Figure 6, it is clear that the running time of `TCMF` scales linearly with the number of sources N, which is consistent with the complexity analysis.

In contrast, though Robuct PCA has a smaller per-iteration runtime when N is small, as N becomes larger, the runtime increases faster than `TCMF`. This is because Robuct PCA vectorizes the observation matrices from each source. The resulting vector from each source has dimension n_1n_2. Robuct PCA then concatenates these vectors into a $n_1n_2 \times N$ matrix and alternatively performs Singular Value Decomposition and hard-thresholding. For each application of SVD, the computational complexity is $\mathcal{O}(n_1n_2(N^2+N)+N^3)$ when $N \leq n_1n_2$ (Li et al., 2019). Indeed, in Figure 6, the slope of the initial part of the Robuct PCA curve is around 1.2, and the slope of the final part is around 3.0, suggesting that the running time scales cubically in the large N regime.

Such comparison highlights TCMF's computational advantage when N is large.

Appendix C. Proof of Theorem 5

In this section, we will introduce the details of the proof of Theorem 5. We will firstly introduce a few basic lemmas, then prove the KKT conditions in Lemma 7. Based on the KKT conditions, we introduce an infinite series to represent the solutions to (8). Next, we will prove Lemma 20, which is a formal version of Lemma 6. Finally, we will use induction to prove Theorem 21, which is the formal version of Theorem 5.

Remember that we use ${\mathbf{L}^\star}_{(i)}$ to denote ${\mathbf{L}^\star}_{(i)} = {\mathbf{L}^\star}_{(i),g} + {\mathbf{L}^\star}_{(i),l}$, where ${\mathbf{L}^\star}_{(i),g}$ and ${\mathbf{L}^\star}_{(i),l}$ are the global and local components for source i defined as ${\mathbf{L}^\star}_{(i),g} = {\mathbf{U}^\star}_g {\mathbf{V}^\star}^T_{(i),g}$ and ${\mathbf{L}^\star}_{(i),l} = {\mathbf{U}^\star}_{(i),l} {\mathbf{V}^\star}^T_{(i),l}$. We assume all nonzero singular values of ${\mathbf{L}^\star}_{(i)}$ are lower bounded by $\sigma_{\min} > 0$ and upper bounded by $\sigma_{\max} > 0$. As introduced in the proof sketch, we use $\mathbf{E}_{(i),t} = {\mathbf{S}^\star}_{(i)} - \hat{\mathbf{S}}_{(i),t}$ to denote the difference between our estimate of the sparse noise at epoch t and the ground truth. The following notations will be used throughout our proof:

$$\mathbf{F}_{(i)} = \mathbf{E}_{(i),t} {\mathbf{L}^\star}^T_{(i)} + {\mathbf{L}^\star}_{(i)} \mathbf{E}^T_{(i),t} + \mathbf{E}_{(i),t} \mathbf{E}^T_{(i),t}, \quad i \in [N], \quad \text{and} \quad \mathbf{F}_{(0)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{F}_{(i)} \tag{18}$$

$$\mathbf{T}_{(i)} = {\mathbf{L}^\star}_{(i)} {\mathbf{L}^\star}^T_{(i)}, \quad i \in [N], \quad \text{and} \quad \mathbf{T}_{(0)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{T}_{(i)}. \tag{19}$$

Since in the ground truth model, the SVD of ${\mathbf{L}^\star}_{(i)}$ can be written as ${\mathbf{L}^\star}_{(i)} = \left[{\mathbf{H}^\star}_g, {\mathbf{H}^\star}_{(i),l}\right] \operatorname{diag}({\mathbf{\Sigma}}_{(i),g}, {\mathbf{\Sigma}}_{(i),l}) \left[{\mathbf{W}^\star}_{(i),g}, {\mathbf{W}^\star}_{(i),l}\right]^T$, one can immediately see that $\mathbf{T}_{(i)}$'s nonzero eigenvalues are upper bounded by $\sigma^2_{\max}$ and lower bounded by $\sigma^2_{\min}$. Finally, recall that we use $\hat{\mathbf{U}}_g$, $\hat{\mathbf{U}}_{(i),l}$, $\hat{\mathbf{V}}_{(i),g}$, and $\hat{\mathbf{V}}_{(i),l}$ to denote the optimal solutions to (8) (we omit the subscript t here for brevity.) For a series of square matrices of the same shape $\mathbf{A}_1, \cdots, \mathbf{A}_k \in \mathbb{R}^{r \times r}$, we use $\prod_{m=1}^{k} \mathbf{A}_m$ to denote the product of these matrices in the ascending order of indices, and $\prod_{m=k}^{1} \mathbf{A}_m$ to denote the product of these matrices in the descending order of indices,

$$\prod_{m=1}^{k} \mathbf{A}_m = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{k-1} \mathbf{A}_k$$

$$\prod_{m=k}^{1} \mathbf{A}_m = \mathbf{A}_k \mathbf{A}_{k-1} \cdots \mathbf{A}_2 \mathbf{A}_1.$$

Our next two lemmas provide upper bound on the maximum row-norm of the errors with respect to the ℓ_∞-norms of $\mathbf{E}_{(i)}$. By building upon these two lemmas, we provide a key result in Lemma 13 connecting $\{\mathbf{F}_{(i)}\}$ and the error matrices $\{\mathbf{E}_{(i)}\}$.

Lemma 11 *Suppose that* $\mathbf{E}_{(1)}, \cdots, \mathbf{E}_{(N)}, \in \mathbb{R}^{n_1 \times n_2}$ *are* α*-sparse and* $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ *is* μ*-incoherent and* $\|U\| \leq 1$*. For any integers* $p_1, p_2, \cdots, p_k \geq 0$*, and* $i_1, i_2, \cdots, i_k \in \{0, 1, \cdots, N\}$*, we have*

$$\max_j \left\| \mathbf{e}_j^T \left(\prod_{\ell=1}^{k} (\mathbf{E}_{(i_\ell)} \mathbf{E}^T_{(i_\ell)})^{p_\ell} \right) \mathbf{U} \right\|_2 \leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \right)^{2(p_1 + p_2 + \cdots + p_k)}. \tag{20}$$

With a slight abuse of notation, in Lemma 11 and the rest of the paper, we define $\mathbf{E}_{(0)}\mathbf{E}_{(0)}^T$ to be,

$$\mathbf{E}_{(0)}\mathbf{E}_{(0)}^T = \frac{1}{N}\sum_{i=0}^{N}\mathbf{E}_{(i)}\mathbf{E}_{(i)}^T. \tag{21}$$

Proof We will prove it by induction on the exponent. From the definition of incoherence, we know that when $p_1 + \cdots + p_k = 0$, the inequality (20) holds. Now suppose that the inequality (20) holds for all $p_1, p_2, \cdots, p_k \geq 0$ such that $p_1 + \cdots + p_k \leq s-1$ and $i_1, i_2, \cdots, i_k \in \{0, 1, \cdots, N\}$. We will prove the statement for $p_1 + \cdots + p_k = s$. Without loss of generality, we assume $p_1 \geq 1$. One can write

$$\begin{aligned}
&\left\|\mathbf{e}_j^T\left(\prod_{\ell=1}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\right\|_2^2 = \sum_l\left(\mathbf{e}_j^T\left(\prod_{\ell=1}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\mathbf{e}_l\right)^2\\
&=\sum_l\left(\mathbf{e}_j^T\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\left(\prod_{\ell=2}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\mathbf{e}_l\right)^2\\
&=\sum_l\left(\sum_h\left[\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right]_{j,h}\mathbf{e}_h^T(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\left(\prod_{\ell=2}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\mathbf{e}_l\right)^2\\
&=\sum_l\sum_{h_1,h_2}\left[\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right]_{j,h_1}\left[\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right]_{j,h_2}\\
&\times\mathbf{e}_{h_1}^T(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\left(\prod_{\ell=2}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\mathbf{e}_l\mathbf{e}_l^T\mathbf{U}^T\left(\prod_{\ell=k}^{2}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\mathbf{e}_{h_2}
\end{aligned}\tag{22}$$

.

Since $\sum_l \mathbf{e}_l\mathbf{e}_l^T = \mathbf{I}$, we can simplify the summation as,

$$\begin{aligned}
&\sum_{h_1,h_2}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_1}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_2}\\
&\times\mathbf{e}_{h_1}^T(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\left(\prod_{\ell=2}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\mathbf{U}^T\left(\prod_{\ell=k}^{2}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\mathbf{e}_{h_2}\\
&\leq\left(\sum_{h_1,h_2}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_1}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_2}\right)\\
&\times\max_m\left\|\mathbf{e}_m^T(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T)^{p_1-1}\left(\prod_{\ell=2}^{k}(\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right)\mathbf{U}\right\|_2^2\\
&\leq\sum_{h_1,h_2}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_1}\left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_2}\frac{\mu^2 r}{n_1}\left(\alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{4s-4},
\end{aligned}$$

where in the last step, we used the induction hypothesis. Now, to complete the proof, we consider two cases.

If $i_1 > 0$, we have:

$$\sum_{h_1,h_2} \left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_1} \left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)_{j,h_2} = \sum_{h_1,h_2,g_1,g_1} (\mathbf{E}_{(i_1)})_{j,g_1}(\mathbf{E}_{(i_1)})_{h_1,g_1}(\mathbf{E}_{(i_1)})_{j,g_2}(\mathbf{E}_{(i_1)})_{h_2,g_2}$$
$$\leq \alpha n_1 \left\|\mathbf{E}_{(i_1)}\right\|_\infty \alpha n_2 \left\|\mathbf{E}_{(i_1)}\right\|_\infty \alpha n_1 \left\|\mathbf{E}_{(i_1)}\right\|_\infty \alpha n_2 \left\|\mathbf{E}_{(i_1)}\right\|_\infty = \left(\alpha \overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^4,$$

where the last inequality holds because at most αn_1 entries in each column of $\mathbf{E}_{(i_1)}$ are nonzero and at most αn_2 entries in each row of $\mathbf{E}_{(i_1)}$ are nonzero. On the other hand, if $i_1 = 0$, we have:

$$\sum_{h_1,h_2} \left(\mathbf{E}_{(0)}\mathbf{E}_{(0)}^T\right)_{j,h_1} \left(\mathbf{E}_{(0)}\mathbf{E}_{(0)}^T\right)_{j,h_2} \leq \frac{1}{N^2} \sum_{h_1,h_2} \left(\sum_{f_1>0} \mathbf{E}_{(f_1)}\mathbf{E}_{(f_1)}^T\right)_{jk_1} \left(\sum_{f_2>0} \mathbf{E}_{(f_2)}\mathbf{E}_{(f_2)}^T\right)_{jk_2}$$
$$= \frac{1}{N^2} N^2 \left(\alpha \overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^4.$$

Therefore, in both cases, we have,

$$\left\|\mathbf{e}_j^T \prod_{\ell=1}^{k} (\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\mathbf{U}\right\|_2^2 \leq \frac{\mu^2 r}{n_1} \left(\alpha \overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{4s},$$

for every possible i_1 and every j. This concludes our proof. ■

Next, we present a slightly different lemma.

Lemma 12 *Suppose that $\mathbf{E}_{(1)}, \cdots, \mathbf{E}_{(N)} \in \mathbb{R}^{n_1 \times n_2}$ are α-sparse and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$ is μ-incoherent. For any integers $p_1, p_2, \cdots, p_k \geq 0$, and $i_1, i_2, \cdots, i_k \in \{0, 1, \cdots, N\}$, we have,*

$$\begin{aligned} &\max_j \left\|\mathbf{e}_j^T \left(\prod_{\ell=1}^{k} (\mathbf{E}_{(i_\ell)}\mathbf{E}_{(i_\ell)}^T)^{p_\ell}\right) \mathbf{E}_{(i_{k+1})}\mathbf{V}\right\|_2 \\ &\leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2(p_1+p_2+\cdots+p_k)+1}. \end{aligned} \tag{23}$$

Proof The proof is analogous to that of Lemma 11, and hence, omitted for brevity. ■

Combining Lemma 11 and 12, we can show the following key lemma on the connection between $\{\mathbf{F}_{(i)}\}$ and the error matrices $\{\mathbf{E}_{(i)}\}$.

Lemma 13 *For every $i \in [N]$, suppose that $\mathbf{E}_{(i)} \in \mathbb{R}^{n_1 \times n_2}$ is α-sparse and $\mathbf{L}^\star{}_{(i)} = \mathbf{H}^\star{}_{(i)}\mathbf{\Sigma}^\star{}_{(i)}\mathbf{W}^\star{}_{(i)}$ is rank-r with μ-incoherent matrices $\mathbf{H}^\star{}_{(i)} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{W}^\star{}_{(i)} \in \mathbb{R}^{n_2 \times r}$.*

For any integers $p_1, p_2, \cdots, p_k \geq 0$, and $i_1, i_2, \cdots, i_k \in \{0, 1, \cdots, N\}$, the following holds for any μ-incoherent matrix $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$,

$$\max_j \left\| \mathbf{e}_j^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{U} \right\|_2 \leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty (\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty + 2\sigma_{\max}) \right)^{p_1+p_2+\cdots+p_k}, \tag{24}$$

where $\mathbf{F}_{(i)}$ is defined as (18).

Proof We firstly expand $\mathbf{F}_{(i_1)}^{p_1} \mathbf{F}_{(i_2)}^{p_2} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U}$ and rearrange the terms by the number of consecutive $\mathbf{E}_{(i)} \mathbf{E}_{(i)}^T$ terms appearing in the beginning of each factor.

$$\begin{aligned}
&\mathbf{F}_{(i_1)}^{p_1} \mathbf{F}_{(i_2)}^{p_2} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U} \\
&= \left(\mathbf{E}_{(i_1)} (\mathbf{L}^\star{}_{i_1})^T + \mathbf{L}^\star{}_{i_1} \mathbf{E}_{(i_1)}^T + \mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right) \cdots \left(\mathbf{E}_{(i_1)} (\mathbf{L}^\star{}_{i_1})^T + \mathbf{L}^\star{}_{i_1} \mathbf{E}_{(i_1)}^T + \mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right) \\
&\left(\mathbf{E}_{(i_2)} (\mathbf{L}^\star{}_{i_2})^T + \mathbf{L}^\star{}_{i_2} \mathbf{E}_{(i_2)}^T + \mathbf{E}_{(i_2)} \mathbf{E}_{(i_2)}^T \right) \cdots \left(\mathbf{E}_{(i_2)} (\mathbf{L}^\star{}_{i_2})^T + \mathbf{L}^\star{}_{i_2} \mathbf{E}_{(i_2)}^T + \mathbf{E}_{(i_2)} \mathbf{E}_{(i_2)}^T \right) \\
&\cdots \\
&\left(\mathbf{E}_{(i_k)} (\mathbf{L}^\star{}_{(i_k)})^T + \mathbf{L}^\star{}_{(i_k)} \mathbf{E}_{(i_k)}^T + \mathbf{E}_{(i_k)} \mathbf{E}_{(i_k)}^T \right) \cdots \left(\mathbf{E}_{(i_k)} (\mathbf{L}^\star{}_{(i_k)})^T + \mathbf{L}^\star{}_{(i_k)} \mathbf{E}_{(i_k)}^T + \mathbf{E}_{(i_k)} \mathbf{E}_{(i_k)}^T \right) \mathbf{U} \\
&= \left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{(i_k)} \mathbf{E}_{(i_k)}^T \right)^{p_k} \mathbf{U} \\
&+ \sum_{r=0}^{p_1+\cdots+p_k-1} \left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})} \mathbf{E}_{(i_{t-1})}^T \right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)} \mathbf{E}_{(i_t)}^T \right)^{r-(\sum_{g=1}^{t-1} p_g)} \\
&\cdot \left(\mathbf{E}_{(i_t)} \mathbf{L}^\star{}_{(i_t)}^T + \mathbf{L}^\star{}_{(i_t)} \mathbf{E}_{(i_t)}^T \right) \mathbf{F}_{(i_t)}^{(\sum_{g=1}^{t} p_g)-1-r} \mathbf{F}_{(i_{t+1})}^{p_{t+1}} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U}.
\end{aligned}$$

For the first term, by Lemma 11, we have

$$\left\| \mathbf{e}_j^T \left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{i_k} \mathbf{E}_{i_k}^T \right)^{p_k} \mathbf{U} \right\|_2 \leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \right)^{2(p_1+\cdots+p_k)}.$$

For the remaining terms, we have

$$\begin{aligned}
&\left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})} \mathbf{E}_{(i_{t-1})}^T \right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)} \mathbf{E}_{(i_t)}^T \right)^{r-(\sum_{g=1}^{t-1} p_g)} \\
&\left(\mathbf{E}_{(i_t)} (\mathbf{L}^\star{}_{(i_t)})^T + \mathbf{L}^\star{}_{(i_t)} \mathbf{E}_{(i_t)}^T \right) \mathbf{F}_{(i_t)}^{(\sum_{g=1}^{t} p_g)-1-r} \mathbf{F}_{(i_{t+1})}^{p_{t+1}} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U} \\
&= \left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})} \mathbf{E}_{(i_{t-1})}^T \right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)} \mathbf{E}_{(i_t)}^T \right)^{r-(\sum_{g=1}^{t-1} p_g)} \mathbf{E}_{(i_t)} \mathbf{W}^\star{}_{(i_t)} \\
&\times \mathbf{\Sigma}^\star{}_{(i_t)} \mathbf{H}^\star{}_{(i_t)}^T \mathbf{F}_{(i_t)}^{(\sum_{g=1}^{t} p_g)-1-r} \mathbf{F}_{(i_{t+1})}^{p_{t+1}} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U} \\
&+ \left(\mathbf{E}_{(i_1)} \mathbf{E}_{(i_1)}^T \right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})} \mathbf{E}_{(i_{t-1})}^T \right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)} \mathbf{E}_{(i_t)}^T \right)^{r-(\sum_{g=1}^{t-1} p_g)} \mathbf{H}^\star{}_{(i_t)} \\
&\times \mathbf{\Sigma}^\star{}_{(i_t)} \mathbf{W}^\star{}_{(i_t)}^T \mathbf{E}_{(i_t)}^T \mathbf{F}_{(i_t)}^{(\sum_{g=1}^{t} p_g)-1-r} \mathbf{F}_{(i_{t+1})}^{p_{t+1}} \cdots \mathbf{F}_{(i_k)}^{p_k} \mathbf{U}.
\end{aligned}$$

We can bound the two terms separately. By Lemma 12,

$$\left\|\mathbf{e}_j^T \left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})}\mathbf{E}_{(i_{t-1})}^T\right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)}\mathbf{E}_{(i_t)}^T\right)^{r-(\sum_{g=1}^{t-1} p_g)} \mathbf{E}_{(i_t)}\mathbf{W}^\star{}_{(i_t)}\right\|$$
$$\leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2r+1}.$$

And by Lemma 11,

$$\left\|\mathbf{e}_j^T \left(\mathbf{E}_{(i_1)}\mathbf{E}_{(i_1)}^T\right)^{p_1} \cdots \left(\mathbf{E}_{(i_{t-1})}\mathbf{E}_{(i_{t-1})}^T\right)^{p_{t-1}} \left(\mathbf{E}_{(i_t)}\mathbf{E}_{(i_t)}^T\right)^{r-(\sum_{g=1}^{t-1} p_g)} \mathbf{H}^\star{}_{(i_t)}\right\|$$
$$\leq \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2r}.$$

For an α-sparse matrix $\mathbf{E}_{(i)} \in \mathbb{R}^{n_1 \times n_2}$, its operator norm is bounded by

$$\begin{aligned}
\left\|\mathbf{E}_{(i)}\right\|_2 &= \max_{\|\mathbf{v}\|=1,\|\mathbf{h}\|=1} \mathbf{v}^T\mathbf{E}_{(i)}\mathbf{h} = \max_{\|\mathbf{v}\|=1,\|\mathbf{h}\|=1} \sum_{j,k} \mathbf{v}_j\mathbf{h}_k[\mathbf{E}_{(i)}]_{jk} \\
&\leq \max_{\|\mathbf{v}\|=1,\|\mathbf{h}\|=1} \sum_{j,k} \frac{1}{2}\left(\mathbf{v}_j^2\sqrt{\frac{n_1}{n_2}} + \mathbf{h}_k^2\sqrt{\frac{n_2}{n_1}}\right)[\mathbf{E}_{(i)}]_{jk} \\
&\leq \max_{\|\mathbf{v}\|=1,\|\mathbf{h}\|=1} \left\|\mathbf{E}_{(i)}\right\|_\infty \frac{1}{2}\left(\sum_j \mathbf{v}_j^2\sqrt{\frac{n_1}{n_2}}\alpha n_2 + \sum_k \mathbf{h}_k^2\sqrt{\frac{n_2}{n_1}}\alpha n_1\right) \\
&= \alpha\sqrt{n_1 n_2}\left\|\mathbf{E}_{(i)}\right\|_\infty.
\end{aligned}$$

Therefore $\left\|\mathbf{E}_{(i)}\right\|_2 \leq \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty$. As a result, we know,

$$\begin{aligned}
\left\|\mathbf{F}_{(i)}\right\|_2 &\leq 2\sigma_{\max}\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty + \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^2 \\
&= \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \left(2\sigma_{\max} + \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right).
\end{aligned}$$

We thus have:

$$\begin{aligned}
&\left\|\mathbf{e}_j^T\mathbf{F}_{(i_1)}^{p_1}\mathbf{F}_{(i_2)}^{p_2}\cdots\mathbf{F}_{(i_k)}^{p_k}\mathbf{U}^\star\right\| \\
&\leq \sqrt{\frac{\mu^2 r}{n_1}}\Bigg(\left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2(\sum_{\ell=1}^k p_\ell)} \\
&+ \sum_{r=0}^{\sum_{\ell=1}^k p_\ell - 1} \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2r+1} \\
&\times 2\sigma_{\max}\left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \left(2\sigma_{\max} + \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)\right)^{\sum_{\ell=1}^k p_\ell - 1 - r}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{\mu^2 r}{n_1}}\Big(\left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{2\sum_{\ell=1}^k p_\ell} \\
&+ \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{\sum_{\ell=1}^k p_\ell} \Big(\left(2\sigma_{\max} + \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{\sum_{\ell=1}^k p_\ell} \\
&- \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{\sum_{\ell=1}^k p_\ell} \Big)\Big) \\
&= \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty\right)^{\sum_{\ell=1}^k p_\ell} \left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty + 2\sigma_{\max}\right)^{\sum_{\ell=1}^k p_\ell}.
\end{aligned}
$$

This finishes our proof. ■

Lemma 13 is an important lemma as it provides an upper bound on the maximum row norm of the product of a group of sparse matrices and an incoherent matrix. We will use Lemma 13 extensively when we calculate the ℓ_∞ norm of error terms in the output of `JIMF`.

Next, we prove that the optimal solution indeed satisfies the KKT conditions delineated in Lemma 7.

Proof of Lemma 7 The proof is presented in three parts. In the first part, we show that the optimal solution optimal $\hat{\mathbf{U}}_g$, $\{\hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l}\}$ satisfies the linear independence constraint qualification (LICQ). This ensures that the optimal solution satisfies the KKT conditions. In the second part, we prove the validity of the equations in (12). Finally, we prove the correctness of the equations in (13).

Proof of LICQ. We begin by showing $\hat{\mathbf{U}}_g$ has full column rank. By contradiction, suppose $\hat{\mathbf{U}}_g$ has rank $r_1^{'} < r_1$. Since $\hat{\mathbf{M}}_{(i)}$ has rank at least $r_1 + r_2$, the residual $\hat{\mathbf{M}}_{(i)} - \hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T - \hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T$ has rank at least 1. Therefore we can always find another $\hat{\mathbf{U}}_g^{'}$ such that $f_i(\hat{\mathbf{U}}_g^{'}, \hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l}) < f_i(\hat{\mathbf{U}}_g, \hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l})$. This contradicts the fact that $\hat{\mathbf{U}}_g$ is optimal.

Next we will establish the LICQ of the constraints. We define h_{ijk} as the inner product between the j-th column of $\mathbf{U}_g$ and the k-th column of $\mathbf{U}_{(i),l}$, $h_{ijk}(\mathbf{x}) = [\mathbf{U}_g]_{:,j}^T[\mathbf{U}_{(i),l}]_{:,k}$. The constraints in (8) can be rewritten as $h_{ijk}(\hat{\mathbf{x}}) = 0, \forall i \in [r_1], \forall j \in [r_2], \forall k \in [N]$. LICQ requires $\nabla h_{ijk}(\hat{\mathbf{x}})$ to be linearly independent for all ijk (Bertsekas, 1997, Proposition 3.1.1).

Suppose we can find constants ψ_{ijk} such that $\sum_{i=1}^N \sum_{j=1}^{r_1} \sum_{k=1}^{r_2} \psi_{ijk}\nabla h_{ijk}(\hat{\mathbf{x}}) = 0$. We consider the partial derivative of h_{ijk} over the $k^{'}$-th column of $\mathbf{U}_{(i^{'}),l}$. It is easy to derive,

$$
\frac{\partial}{\partial[\mathbf{U}_{(i^{'}),l}]_{:,k^{'}}} h_{ijk}(\hat{\mathbf{x}}) = \delta_{ii^{'}}\delta_{kk^{'}}[\hat{\mathbf{U}}_g]_{:,j},
$$

where $\delta_{ii^{'}}$ is the Kronecker delta function. Then the constants ψ_{ijk} should satisfy,

$$
\sum_{j=1}^{r_2} \psi_{i^{'}jk^{'}}[\hat{\mathbf{U}}_g]_{:,j} = 0.
$$

As the columns of $\hat{\mathbf{U}}_g$ are linearly independent, $\psi_{i^{'}jk^{'}} = 0$ for each j. This holds for any $i^{'}$ and $k^{'}$. Therefore $\psi_{ijk} = 0$ for all i, j, k. This implies ∇h_{ijk}'s are linearly independent.

Proof of Equations (12). The Lagrangian of the optimization problem (8) can be written as

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\sum_{i=1}^{N}\left\|\mathbf{U}_g\mathbf{V}_{(i),g}^T+\mathbf{U}_{(i),l}\mathbf{V}_{(i),l}^T-\hat{\mathbf{M}}_{(i)}\right\|_F^2 \\ &+\operatorname{Tr}\left(\mathbf{\Lambda}_{8,(i)}\mathbf{U}_g^T\mathbf{U}_{(i),l}\right),\end{aligned} \tag{25}$$

where $\mathbf{\Lambda}_{8,(i)}$ is the dual variable for the constraint $\mathbf{U}_g^T\mathbf{U}_{(i),l}=0$.

Under the LICQ, we know that $\hat{\mathbf{U}}_g, \{\hat{\mathbf{V}}_{(i),g}, \hat{\mathbf{U}}_{(i),l}, \hat{\mathbf{V}}_{(i),l}\}$ satisfies KKT condition. Setting the gradient of $\mathcal{L}$ with respect to $\mathbf{V}_{(i),g}$ and $\mathbf{V}_{(i),l}$ to zero, we can prove (12d) and (12c). Considering the constraint $\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_{(i),l}=0$, we can solve them as $\hat{\mathbf{V}}_{(i),g}=\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g\left(\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g\right)^{-1}$ and $\hat{\mathbf{V}}_{(i),l}=\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}\left(\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}\right)^{-1}$. Then we examine the gradient of $\mathcal{L}$ with respect to $\mathbf{U}_{(i),l}$:

$$\frac{\partial}{\partial\mathbf{U}_{(i),l}}\mathcal{L}=\left(\mathbf{U}_g\mathbf{V}_{(i),g}^T+\mathbf{U}_{(i),l}\mathbf{V}_{(i),l}^T-\hat{\mathbf{M}}_{(i)}\right)\mathbf{V}_{(i),l}+\mathbf{U}_g\mathbf{\Lambda}_{(8),i}^T.$$

Substituting $\hat{\mathbf{V}}_{(i),g}=\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g\left(\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g\right)^{-1}$ and $\hat{\mathbf{V}}_{(i),l}=\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}\left(\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}\right)^{-1}$ in the above gradient and setting it to zero, we have

$$\begin{aligned}&\left(\hat{\mathbf{U}}_g\left(\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g\right)^{-1}\hat{\mathbf{U}}_g^T+\hat{\mathbf{U}}_{(i),l}\left(\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}\right)^{-1}\hat{\mathbf{U}}_{(i),l}^T-\mathbf{I}\right)\hat{\mathbf{M}}_{(i)}\hat{\mathbf{V}}_{(i),l}\\&+\hat{\mathbf{U}}_g\mathbf{\Lambda}_{(8),i}^T=0.\end{aligned}$$

Left multiplying both sides by $\hat{\mathbf{U}}_g^T$, we have $\mathbf{\Lambda}_{8,(i)}=0$. Left multiplying both sides by $\hat{\mathbf{U}}_{(i),l}^T$, we have $\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}-\mathbf{I}=0$. Therefore we also have $\left(\hat{\mathbf{U}}_g\left(\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g\right)^{-1}\hat{\mathbf{U}}_g^T+\hat{\mathbf{U}}_{(i),l}\left(\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}\right)^{-1}\hat{\mathbf{U}}_{(i),l}^T-\mathbf{I}\right)\hat{\mathbf{M}}_{(i)}\mathbf{V}_{(i),l}=0$. This proves equation (12b). Now, setting the derivative of $\mathcal{L}$ with respect to $\mathbf{U}_g$ to zero, we have

$$\frac{\partial}{\partial\mathbf{U}_g}\mathcal{L}=\sum_{i=1}^{N}\left(\hat{\mathbf{U}}_g\hat{\mathbf{V}}_{(i),g}^T+\hat{\mathbf{U}}_{(i),l}\hat{\mathbf{V}}_{(i),l}^T-\hat{\mathbf{M}}_{(i)}\right)\hat{\mathbf{V}}_{(i),g}=0.$$

Left multiplying both sides by $\hat{\mathbf{U}}_g^T$, we have $\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_g-\mathbf{I}=0$. We have thus proven (12a). This completes the proof for (12).

Proof of Equations (13). Equation (12b) can be rewritten as:

$$\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}=\hat{\mathbf{U}}_{(i),l}\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}+\hat{\mathbf{U}}_g\hat{\mathbf{U}}_g^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}. \tag{26}$$

Since $\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}$ is positive definite, we can use $\mathbf{W}_{(i),l}\mathbf{\Lambda}_{2,(i)}\mathbf{W}_{(i),l}^T=\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_{(i),l}$ to denote its eigen-decomposition, where $\mathbf{\Lambda}_{2,(i)}\in\mathbb{R}^{r_1\times r_1}$ is a positive definite diagonal matrix and $\mathbf{W}_{(i),l}\in\mathbb{R}^{r_1\times r_1}$ is orthonormal. Upon defining $\hat{\mathbf{H}}_{(i),l}=\hat{\mathbf{U}}_{(i),l}\mathbf{W}_{(i),l}$,

$\hat{\mathbf{H}}_{(i),l}$ is also orthonormal as $\hat{\mathbf{H}}_{(i),l}^T\hat{\mathbf{H}}_{(i),l} = \mathbf{W}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}^T\hat{\mathbf{U}}_{(i),l}\mathbf{W}_{(i),l} = \mathbf{I}$. Similarly, we rewrite the equation (12a) as:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_{(i)}\mathbf{M}_{(i)}^T\hat{\mathbf{U}}_g = \hat{\mathbf{U}}_g\hat{\mathbf{U}}_g^T\frac{1}{N}\sum_{i=1}^{N}\mathbf{M}_{(i)}\mathbf{M}_{(i)}^T\hat{\mathbf{U}}_g + \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{U}}_{(i),l}\hat{\mathbf{U}}_{(i),l}^T\mathbf{M}_{(i)}\mathbf{M}_{(i)}^T\hat{\mathbf{U}}_g. \tag{27}$$

Since $\hat{\mathbf{U}}_g^T\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g$ is positive definite, we can use $\mathbf{W}_g\mathbf{\Lambda}_1\mathbf{W}_g^T = \hat{\mathbf{U}}_g^T\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{U}}_g$ to denote its eigen decomposition, where $\mathbf{\Lambda}_1 \in \mathbb{R}^{r_1\times r_1}$ is positive diagonal, $\mathbf{W}_g \in \mathbb{R}^{r_1\times r_1}$ is orthogonal $\mathbf{W}_g\mathbf{W}_g^T = \mathbf{W}_g^T\mathbf{W}_g = \mathbf{I}$. We define $\hat{\mathbf{H}}_g$ as $\hat{\mathbf{H}}_g = \hat{\mathbf{U}}_g\mathbf{W}_g$, then $\hat{\mathbf{H}}_g$ is also orthonormal. Additionally, $\hat{\mathbf{H}}_g^T\hat{\mathbf{H}}_{(i),l} = \mathbf{W}_g^T\hat{\mathbf{U}}_g^T\hat{\mathbf{U}}_{(i),l}\mathbf{W}_{(i),l} = 0$. This completes the proof of equation (13c).

Next, we proceed with the proof of equations (13b) and (13a). By right multiplying both sides of (27) with $\mathbf{W}_g$ and replacing $\hat{\mathbf{U}}_g$ and $\hat{\mathbf{U}}_{(i),l}$ by $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_g = \hat{\mathbf{H}}_g\mathbf{\Lambda}_1 + \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_g. \tag{28}$$

Similarly, by right multiplying both sides of (26) with $\mathbf{W}_{(i),l}$, we can rewrite (26) as,

$$\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_{(i),l} = \hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)} + \hat{\mathbf{H}}_g\hat{\mathbf{H}}_g^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_{(i),l}. \tag{29}$$

We thus prove the equations (13b) and (13a), where $\mathbf{\Lambda}_{3,(i)} = \hat{\mathbf{H}}_g^T\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}_{(i)}^T\hat{\mathbf{H}}_{(i),l}$. ■

We note that the KKT conditions provide a set of conditions that must be satisfied for *all* stationary points of (8). Our next key contribution is to use these conditions to characterize a few interesting properties satisfied by *all* the optimal solutions. To this goal, we heavily rely on the spectral properties of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_{2,(i)}$, and $\mathbf{\Lambda}_{3,(i)}$.

For simplicity, we introduce three additional notations, $\mathbf{\Lambda}_{4,(i)} = -\mathbf{\Lambda}_{3,(i)}$, $\mathbf{\Lambda}_{5,(i)} = -\mathbf{\Lambda}_{3,(i)}^T/N$, and $\mathbf{\Lambda}_6 \in \mathbb{R}^{r_1\times r_1}$ defined as,

$$\mathbf{\Lambda}_6 = \mathbf{\Lambda}_1 - \frac{1}{N}\sum_{i=1}^{N}\mathbf{\Lambda}_{3,(i)}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{3,(i)}^T. \tag{30}$$

$\mathbf{\Lambda}_6$ is a symmetric matrix. It is worth noting that $\mathbf{\Lambda}_6$ is well defined as the diagonal matrix $\mathbf{\Lambda}_{2,(i)}$ is invertible throughput the proof. We also introduce short-hand notation $\Delta\mathbf{P}_g$ to denote $\mathbf{P}_{\hat{\mathbf{H}}_g} - \mathbf{P}_{\mathbf{U}^\star{}_g}$ and $\Delta\mathbf{P}_{(i),l}$ to denote $\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}} - \mathbf{P}_{\mathbf{U}^\star{}_{(i),l}}$.

Spectral properties of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_{2,(i)}$, $\mathbf{\Lambda}_{3,(i)}$, and $\mathbf{\Lambda}_6$ are critical for developing the solutions to KKT conditions. We will establish these properties in the following lemmas.

Before diving into these properties, we investigate the deviance of the estimates features $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$ to ground truth features $\mathbf{U}^\star{}_g$ and $\mathbf{U}^\star{}_{(i),l}$.

Lemma 14 *If* $\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \leq 4\sigma_{\max}\frac{\mu^2 r}{\overline{n}}$, *and* $\mathbf{E}_{(i)}$ *is* α*-sparse with* $\alpha \leq \frac{1}{60\mu^4 r^2}\frac{\sigma_{\min}^4}{\sigma_{\max}^4}\left(1+\frac{4\sigma_{\max}^2}{\sqrt{\theta}\sigma_{\min}^2}\right)^{-2}$, *we have,*

$$\left\|\mathbf{P}_{\mathbf{U}^\star{}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right\|_F \leq \sqrt{\alpha\overline{n}}\max_i\left\|\mathbf{E}_{(i)}\right\|_\infty\frac{5\sigma_{\max}}{\sqrt{\theta}\sigma_{\min}^2}, \tag{31}$$

and

$$\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}} - \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\| \leq 3\sqrt{\alpha}\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \frac{\sigma_{\max}}{\sigma^2_{\min}} \left(1 + 4\frac{\sigma^2_{\max}}{\sqrt{\theta}\sigma^2_{\min}}\right). \tag{32}$$

Proof

By Shi and Kontar (2024, Theorem 1), we know that $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$ corresponding to the global optimal solutions to the problem (8) satisfy

$$\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right\|_F^2 + \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}} - \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|_F^2 \leq \frac{4}{N}\sum_{i=1}^{N}\frac{\left\|\mathbf{F}_{(i)}\right\|_F^2}{\theta\sigma^4_{\min}}. \tag{33}$$

Note that the norm of the error term $\left\|\mathbf{F}_{(i)}\right\|_F$ is bounded by:

$$\begin{aligned}
\left\|\mathbf{F}_{(i)}\right\|_F &= \left\|\mathbf{L}^{\star}{}_{(i)}\mathbf{E}^T_{(i),t} + \mathbf{E}_{(i),t}{\mathbf{L}^{\star}{}_{(i)}}^T + \mathbf{E}_{(i),t}\mathbf{E}^T_{(i),t}\right\|_F \\
&\leq \left\|\mathbf{E}_{(i),t}\right\|_F \left(2\left\|\mathbf{L}^{\star}{}_{(i)}\right\|_2 + \left\|\mathbf{E}_{(i),t}\right\|_2\right) \\
&\leq \sqrt{\alpha}\overline{n}\left\|\mathbf{E}_{(i),t}\right\|_{\infty}\left(2\sigma_{\max} + \alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty}\right).
\end{aligned} \tag{34}$$

Therefore, we know from (33) that

$$\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right\|_F \\
&\leq 2\frac{\left\|\mathbf{F}_{(i)}\right\|_F}{\sqrt{\theta\sigma^4_{\min}}} \leq \sqrt{\alpha}\overline{n}\left\|\mathbf{E}_{(i),t}\right\|_{\infty}\left(2\sigma_{\max} + \alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty}\right)\frac{2}{\sqrt{\theta\sigma^4_{\min}}} \\
&\leq \sqrt{\alpha}\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty}\frac{5\sigma_{\max}}{\sqrt{\theta}\sigma^2_{\min}},
\end{aligned}$$

where we used the condition $\alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty} \leq \sigma_{\max}/2$ for the last inequality.

From (8), we can also deduce that that the column vectors of $\hat{\mathbf{H}}_{(i),l}$ span the top invariant subspace of $\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right)\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}^T_{(i)}\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right)$. Since column vectors of $\mathbf{U}^{\star}{}_{(i),l}$ span the top invariant subspace of $\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\mathbf{M}^{\star}{}_{(i)}\mathbf{M}^{\star T}_{(i)}\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)$, we know from Weyl's theorem (Tao, 2010) and Davis-Khan theorem (Rinaldo, 2017) that,

$$\begin{aligned}
&\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}} - \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\| \leq \\
&\frac{\left\|\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right)\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}^T_{(i)}\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right) - \left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\mathbf{M}^{\star}{}_{(i)}\mathbf{M}^{\star T}_{(i)}\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\right\|_F}{\sigma^2_{\min} - \left\|\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right)\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}^T_{(i)}\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right) - \left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\mathbf{M}^{\star}{}_{(i)}\mathbf{M}^{\star T}_{(i)}\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\right\|}.
\end{aligned}$$

Since

$$\left\|\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right)\hat{\mathbf{M}}_{(i)}\hat{\mathbf{M}}^T_{(i)}\left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right) - \left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\mathbf{M}^{\star}{}_{(i)}\mathbf{M}^{\star T}_{(i)}\left(\mathbf{I} - \mathbf{P}_{\mathbf{U}^{\star}{}_g}\right)\right\|_F$$

$$
\begin{aligned}
&= \left\| \left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g} \right) \left(\hat{\mathbf{M}}_{(i)} \hat{\mathbf{M}}_{(i)}^T - \mathbf{M}^\star{}_{(i)} \mathbf{M}^{\star T}_{(i)} \right) \left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g} \right) \right\|_F \\
&+ \left\| \mathbf{M}^\star{}_{(i)} \mathbf{M}^{\star T}_{(i)} \left(\mathbf{P}_{\mathbf{U}^\star{}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g} \right) \right\|_F + \left\| \left(\mathbf{P}_{\mathbf{U}^\star{}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g} \right) \mathbf{M}^\star{}_{(i)} \mathbf{M}^{\star T}_{(i)} \right\|_F \\
&\leq \left\| \mathbf{F}_{(i)} \right\|_F + 2\sigma_{\max}^2 \left\| \mathbf{P}_{\hat{\mathbf{H}}_g} - \mathbf{P}_{\mathbf{U}^\star{}_g} \right\|_F \leq \frac{5}{2} \sqrt{\alpha} \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \sigma_{\max} \left(1 + 4 \frac{\sigma_{\max}^2}{\sqrt{\theta} \sigma_{\min}^2} \right) \\
&\leq \frac{\sigma_{\min}^2}{6},
\end{aligned}
$$

we have,

$$
\begin{aligned}
&\left\| \mathbf{P}_{\mathbf{U}^\star{}_{(i),l}} - \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}} \right\| \\
&\leq \frac{1}{\sigma_{\min}^2 - \frac{\sigma_{\min}^2}{6}} \frac{5}{2} \sqrt{\alpha} \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \sigma_{\max} \left(1 + 4 \frac{\sigma_{\max}^2}{\sqrt{\theta} \sigma_{\min}^2} \right) \\
&\leq 3 \sqrt{\alpha} \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \frac{\sigma_{\max}}{\sigma_{\min}^2} \left(1 + 4 \frac{\sigma_{\max}^2}{\sqrt{\theta} \sigma_{\min}^2} \right).
\end{aligned}
$$

This completes the proof. ■

With Lemma 14, we first provide upper bound on the operator norm of $\mathbf{\Lambda}_{3,(i)}$.

Lemma 15 *For every $i \in [N]$, suppose that $\mathbf{U}^\star_{(i),l}$'s are θ-misaligned, $\max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \leq 4\sigma_{\max} \frac{\mu^2 r}{\overline{n}}$, and $\mathbf{E}_{(i)}$ is α-sparse with $\alpha \leq \frac{1}{10\mu^2 r}$, we have,*

$$
\left\| \mathbf{\Lambda}_{3,(i)} \right\| \leq 2\sigma_{\max}.
$$

Proof From the KKT condition (13), we know,

$$
\begin{aligned}
\left\| \mathbf{\Lambda}_{3,(i)} \right\| &= \left\| \hat{\mathbf{H}}_g^T \left(\mathbf{T}_{(i)} + \mathbf{F}_{(i)} \right) \hat{\mathbf{H}}_{(i),l} \right\| \\
&\leq \left\| \mathbf{T}_{(i)} \right\| + \left\| \mathbf{F}_{(i)} \right\| \leq 2\sigma_{\max}.
\end{aligned}
$$

This completes the proof. ■

We then estimate lower bounds on the smallest eigenvalues of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_{2,(i)}$, and $\mathbf{\Lambda}_6$. These estimates rely on more refined matrix perturbation analysis.

Lemma 16 *For every $i \in [N]$, suppose that $\mathbf{U}^\star_{(i),l}$'s are θ-misaligned, $\max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \leq 4\sigma_{\max} \frac{\mu^2 r}{\overline{n}}$, and $\mathbf{E}_{(i)}$ is α-sparse with*

$$
\alpha \leq \frac{1}{64} \frac{1}{\mu^4 r^2} \left(\frac{\sigma_{\min}}{\sigma_{\max}} \right)^4 \left(1 + 2 \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^2 + \frac{8}{\sqrt{\theta}} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^4 \right)^{-2}.
$$

The minimum eigenvalues of $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_{2,(i)}$ are lower bounded by $\frac{3}{4}\sigma_{\min}^2$.

Proof This lemma is a result of Weyl's theorem (Tao, 2010) and the perturbation bound on the eigenspaces. From the first equation in (13), we know,

$$\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\hat{\mathbf{H}}_{(i),l}=\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}.$$

Therefore, $\hat{\mathbf{H}}_{(i),l}$'s columns are the eigenvectors of the symmetric matrix $\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}$, with eigenvalues corresponding to the diagonal entries of $\mathbf{\Lambda}_{2,(i)}$. According to the definition of $\mathbf{T}_{(i)}$, we know that the eigenvalues of $\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}=\mathbf{H^\star}_{(i),l}\mathbf{\Sigma^\star}^2_{(i),l}\mathbf{H^\star}^T_{(i),l}$ are lower bounded by $\sigma^2_{\min}$. Hence, as a result of Weyl's inequality, we have

$$\begin{aligned}&\lambda_{\min}\left(\mathbf{\Lambda}_{2,(i)}\right)=\lambda_{\min}\left(\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\\&\geq\sigma^2_{\min}-\left\|\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|_2.\end{aligned}$$

On the other hand, by triangle inequalities, we have

$$\begin{aligned}&\left\|\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|_2\\&\leq\left\|\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|_2+\left\|\mathbf{F}_{(i)}\right\|_2\\&\leq\left\|\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\right\|_2\\&\quad+\left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\mathbf{U^\star}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{T}_{(i)}\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|_2+\left\|\mathbf{F}_{(i)}\right\|_2\\&\leq 2\sigma^2_{\max}\left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}-\mathbf{P}_{\mathbf{U^\star}_{(i),l}}\right\|_2+\left\|\mathbf{F}_{(i)}\right\|_2\\&\leq\sqrt{\alpha\overline{n}}\max_i\left\|\mathbf{E}_{(i)}\right\|_\infty\left(2\sigma_{\max}+\sqrt{\alpha\overline{n}}\max_i\left\|\mathbf{E}_{(i)}\right\|_\infty\right)\\&+\frac{5}{2}\sqrt{\alpha\overline{n}}\max_i\left\|\mathbf{E}_{(i)}\right\|_\infty\frac{\sigma_{\max}}{\sigma^2_{\min}}\left(1+4\frac{\sigma^2_{\max}}{\sqrt{\theta}\sigma^2_{\min}}\right)\\&\leq\frac{1}{4}\sigma^2_{\min},\end{aligned}$$

where we used the fact $\left\|\mathbf{T}_{(i)}\right\|\leq\sigma^2_{\max}$ in the third inequality, Lemma 14 in the 4th inequality, and the assumed upper bound on α in the last inequality. We thus have $\lambda_{\min}\left(\mathbf{\Lambda}_{2,(i)}\right)\geq\frac{3}{4}\sigma^2_{\min}$.

Similarly, we can solve $\mathbf{\Lambda}_{3,(i)}$ from the first equation of (13) as $\mathbf{\Lambda}_{3,(i)}=\hat{\mathbf{H}}_g^T\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\hat{\mathbf{H}}_{(i),l}$. Plugging this into the second equation of (13), we have

$$\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\hat{\mathbf{H}}_g=\hat{\mathbf{H}}_g\mathbf{\Lambda}_1.$$

Thus, the columns of $\hat{\mathbf{H}}_g$ are the eigenvectors of the matrix $\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)$, with eigenvalues corresponding to the diagonal entries of $\mathbf{\Lambda}_1$. Again, since the minimum eigenvalue of

$\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)$ is lower bounded by $\sigma^2_{\min}$, Weyl's inequality can be invoked to provide a lower bound on the minimum eigenvalue of $\mathbf{\Lambda}_1$:

$$\begin{aligned}
&\lambda_{\min}\left(\mathbf{\Lambda}_1\right)=\lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\right)\\
&\geq \sigma^2_{\min}-\\
&\left\|\frac{\sum_{i=1}^{N}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)-\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)}{N}\right\|_2.
\end{aligned}$$

The operator norm on the right hand side can be bounded by triangle inequalities. For each term in the summation, we have

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)-\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\left(\mathbf{T}_{(i)}+\mathbf{F}_{(i)}\right)\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\right\|_2\\
\leq&\left\|\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)-\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)\right\|\\
&+\left\|\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}\right)-\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\right\|\\
&+\left\|\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{F}_{(i)}\left(\mathbf{I}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\right\|\\
\leq&2\sigma^2_{\max}\left\|\mathbf{P}_{\mathbf{U}^{\star}{}_{(i),l}}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|+\left\|\mathbf{F}_{(i)}\right\|\\
\leq&\frac{1}{4}\sigma^2_{\min}.
\end{aligned}$$

where we used the assumed upper bound on α. This completes the proof. ■

We also provide a lower bound on the minimum eigenvalue of the symmetric matrix $\mathbf{\Lambda}_6$. Remember that $\mathbf{\Lambda}_6$ is defined as $\mathbf{\Lambda}_6=\mathbf{\Lambda}_1-\frac{1}{N}\sum_{i=1}^{N}\mathbf{\Lambda}_{3,(i)}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{3,(i)}^{T}$.

Lemma 17 *For every $i\in[N]$, suppose that $\mathbf{U}^{\star}_{(i),l}$'s are θ-misaligned, $\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty}\leq 4\sigma_{\max}\frac{\mu^2 r}{n}$, and $\mathbf{E}_{(i)}$ is α-sparse with $\alpha\leq\frac{1}{(\mu^2 r 640)^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^8\left(1+\frac{4\sigma^2_{\max}}{\sqrt{\theta}\sigma^2_{\min}}\right)^{-2}$, then, the minimum eigenvalue of $\mathbf{\Lambda}_6$ is lower bounded by $\frac{3}{4}\sigma^2_{\min}$.*

Proof The proof is constructive. We use two steps. In the first step, we introduce a block matrix $\mathbf{\Lambda}_7$ defined as,

$$\mathbf{\Lambda}_7=\begin{pmatrix}\sqrt{N}\mathbf{\Lambda}_1 & \mathbf{\Lambda}_{3,(1)} & \cdots & \mathbf{\Lambda}_{3,(N)}\\ \mathbf{\Lambda}_{3,(1)}^{T} & \sqrt{N}\mathbf{\Lambda}_{2,(1)} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{\Lambda}_{3,(N)} & 0 & \cdots & \sqrt{N}\mathbf{\Lambda}_{2,(N)}\end{pmatrix},\tag{35}$$

and show that the minimum eigenvalue of the minimum eigenvalue of $\mathbf{\Lambda}_7$ is lower bounded by $\sqrt{N}\frac{3}{4}\sigma^2_{\min}$. Then, in the second step, we prove that the minimum eigenvalue of $\mathbf{\Lambda}_6$ is lower bounded by $\frac{1}{\sqrt{N}}$ multiplies the minimum eigenvalue of $\mathbf{\Lambda}_7$, $\lambda_{min}(\mathbf{\Lambda}_6)\geq\lambda_{min}(\frac{1}{\sqrt{N}}\mathbf{\Lambda}_7)$.

During this proof, we further introduce

$$\begin{cases} \mathbf{\Lambda^\star}_{1,(i)} = \mathbf{H^\star}_g^T \mathbf{M^\star}_{(i)} \mathbf{M^\star}_{(i)}^T \mathbf{H^\star}_g \\ \quad \mathbf{\Lambda^\star}_1 = \dfrac{1}{N} \sum_{i=1}^N \mathbf{\Lambda^\star}_{1,(i)} \\ \mathbf{\Lambda^\star}_{2,(i)} = \mathbf{H^\star}_{(i),l}^T \mathbf{M^\star}_{(i)} \mathbf{M^\star}_{(i)}^T \mathbf{H^\star}_{(i),l} \\ \mathbf{\Lambda^\star}_{3,(i)} = \mathbf{H^\star}_g^T \mathbf{M^\star}_{(i)} \mathbf{M^\star}_{(i)}^T \mathbf{H^\star}_{(i),l}, \end{cases}$$

for notational simplicity. From the SVD (3) and the assumption on singular values of $\mathbf{L^\star}_{(i)}$, we know:

$$[\mathbf{H^\star}_g, \mathbf{H^\star}_{(i),l}]^T \mathbf{L^\star}_{(i)} \mathbf{L^\star}_{(i)}^T [\mathbf{H^\star}_g, \mathbf{H^\star}_{(i),l}] = \begin{pmatrix} \mathbf{\Lambda^\star}_{1,(i)} & \mathbf{\Lambda^\star}_{3,(i)} \\ \mathbf{\Lambda^\star}_{3,(i)}^T & \mathbf{\Lambda^\star}_{2,(i)} \end{pmatrix} \succ \sigma_{\min}^2 \mathbf{I}. \tag{36}$$

Step 1: Minimum eigenvalue of $\mathbf{\Lambda}_7$: From definitions of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_{2,(i)}$, and $\mathbf{\Lambda}_{3,(i)}$ in (13), we know,

$$\mathbf{\Lambda}_7 = \underbrace{\begin{pmatrix} \sqrt{N}\hat{\mathbf{H}}_g^T \mathbf{T}_{(0)} \hat{\mathbf{H}}_g & \hat{\mathbf{H}}_g^T \mathbf{T}_{(1)} \hat{\mathbf{H}}_{(1),l} & \cdots & \hat{\mathbf{H}}_g^T \mathbf{T}_{(N)} \hat{\mathbf{H}}_{(N),l} \\ \hat{\mathbf{H}}_{(1),l}^T \mathbf{T}_{(1)} \hat{\mathbf{H}}_g & \sqrt{N}\hat{\mathbf{H}}_{(1),l}^T \mathbf{T}_{(1)} \hat{\mathbf{H}}_{(1),l} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{H}}_{(N),l}^T \mathbf{T}_{(N)} \hat{\mathbf{H}}_g & 0 & \cdots & \sqrt{N}\hat{\mathbf{H}}_{(N),l}^T \mathbf{T}_{(N)} \hat{\mathbf{H}}_{(N),l} \end{pmatrix}}_{\mathbf{\Lambda}_{7,2}}$$
$$+ \underbrace{\begin{pmatrix} \sqrt{N}\hat{\mathbf{H}}_g^T \mathbf{F}_{(0)} \hat{\mathbf{H}}_g & \hat{\mathbf{H}}_g^T \mathbf{F}_{(1)} \hat{\mathbf{H}}_{(1),l} & \cdots & \hat{\mathbf{H}}_g^T \mathbf{F}_{(N)} \hat{\mathbf{H}}_{(N),l} \\ \hat{\mathbf{H}}_{(1),l}^T \mathbf{F}_{(1)} \hat{\mathbf{H}}_g & \sqrt{N}\hat{\mathbf{H}}_{(1),l}^T \mathbf{F}_{(1)} \hat{\mathbf{H}}_{(1),l} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{H}}_{(N),l}^T \mathbf{F}_{(N)} \hat{\mathbf{H}}_g & 0 & \cdots & \sqrt{N}\hat{\mathbf{H}}_{(N),l}^T \mathbf{F}_{(N)} \hat{\mathbf{H}}_{(N),l} \end{pmatrix}}_{\mathbf{\Lambda}_{7,1}}.$$

By Lemma 26, the operator norm of $\mathbf{\Lambda}_{7,1}$ is upper bounded by,

$$\begin{aligned} \|\mathbf{\Lambda}_{7,1}\| &\leq \max_{i=0,1,\cdots,N} \{\sqrt{N} \left\|\mathbf{F}_{(i)}\right\|\} + \sqrt{2 \sum_{i=1}^N \left\|\mathbf{F}_{(i)}\right\|^2} \\ &\leq 2\sqrt{N}\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \sigma_{\max} \frac{5}{2} \leq \frac{\sqrt{N}\sigma_{\min}^2}{16}. \end{aligned} \tag{37}$$

where we used the condition $\alpha \leq \frac{\sigma_{\min}^2}{\sigma_{\max}^2} \frac{1}{320\mu^2 r}$ in the last inequality.

We can further decompose $\mathbf{\Lambda}_{7,2}$. . Then, we can derive $\hat{\mathbf{H}}_g^T \mathbf{T}_{(0)} \hat{\mathbf{H}}_g = \hat{\mathbf{H}}_g^T \mathbf{H^\star}_g \mathbf{H^\star}_g^T \mathbf{T}_{(0)} \mathbf{H^\star}_g \mathbf{H^\star}_g^T \hat{\mathbf{H}}_g + \hat{\mathbf{H}}_g^T \Delta \mathbf{P}_g \mathbf{T}_{(0)} \hat{\mathbf{H}}_g + \hat{\mathbf{H}}_g^T \mathbf{H^\star}_g \mathbf{H^\star}_g^T \mathbf{T}_{(0)} \Delta \mathbf{P}_g \hat{\mathbf{H}}_g$, $\hat{\mathbf{H}}_{(i),l}^T \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} = \hat{\mathbf{H}}_{(i),l}^T \mathbf{H^\star}_{(i),l} \mathbf{H^\star}_{(i),l}^T \mathbf{T}_{(i)} \mathbf{H^\star}_{(i),l} \mathbf{H^\star}_{(i),l}^T \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_{(i),l}^T \Delta \mathbf{P}_{(i),l} \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} +$

$\hat{\mathbf{H}}_{(i),l}^T \mathbf{H}^\star{}_{(i),l} \mathbf{H}^{\star T}_{(i),l} \mathbf{T}_{(i)} \Delta \mathbf{P}_{(i),l} \hat{\mathbf{H}}_{(i),l}$, and $\hat{\mathbf{H}}_g^T \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} = \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(i)} \mathbf{H}^\star{}_{(i),l} \mathbf{H}^{\star T}_{(i),l} \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_g^T \Delta \mathbf{P}_g \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(i)} \Delta \mathbf{P}_{(i),l} \hat{\mathbf{H}}_{(i),l}$.

Therefore, we can rewrite $\mathbf{\Lambda}_{7,2}$ as,

$$\mathbf{\Lambda}_{7,2} = \mathbf{\Lambda}_{7,3} + \underbrace{\begin{pmatrix} \sqrt{N}\hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{\Lambda}^\star{}_1 \mathbf{H}^{\star T}_g \hat{\mathbf{H}}_g & \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{\Lambda}^\star{}_{3,(1)} \mathbf{H}^{\star T}_{(1),l} \hat{\mathbf{H}}_{(1),l} & \cdots & \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{\Lambda}^\star{}_{3,(N)} \mathbf{H}^{\star T}_{(N),l} \hat{\mathbf{H}}_{(N),l} \\ \hat{\mathbf{H}}_{(1),l}^T \mathbf{H}^\star{}_{(1),l} \mathbf{\Lambda}^{\star T}_{3,(1)} \mathbf{H}^{\star T}_g \hat{\mathbf{H}}_g & \sqrt{N}\hat{\mathbf{H}}_{(1),l}^T \mathbf{H}^\star{}_{(1),l} \mathbf{\Lambda}^\star{}_{2,(1)} \mathbf{H}^{\star T}_{(1),l} \hat{\mathbf{H}}_{(1),l} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{H}}_{(N),l}^T \mathbf{H}^\star{}_{(N),l} \mathbf{\Lambda}^{\star T}_{3,(N)} \mathbf{H}^{\star T}_g \hat{\mathbf{H}}_g & 0 & \cdots & \sqrt{N}\hat{\mathbf{H}}_{(N),l}^T \mathbf{H}^\star{}_{(N),l} \mathbf{\Lambda}^\star{}_{2,(N)} \mathbf{H}^{\star T}_{(N),l} \hat{\mathbf{H}}_{(N),l} \end{pmatrix}}_{\mathbf{\Lambda}_{7,4}},$$

where $\mathbf{\Lambda}_{7,3}$ consists of residual terms that contain $\Delta \mathbf{P}_g$ or $\Delta \mathbf{P}_{(i),l}$.

We use Lemma 26 to estimate an upper bound for the operator norm of $\mathbf{\Lambda}_{7,3}$. The maximum operator norm of the diagonal block of $\mathbf{\Lambda}_{7,3}$ is

$$\begin{aligned} &\max\{\sqrt{N}\left\|\hat{\mathbf{H}}_g^T \Delta \mathbf{P}_g \mathbf{T}_{(0)} \hat{\mathbf{H}}_g + \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(0)} \Delta \mathbf{P}_g \hat{\mathbf{H}}_g\right\|, \\ &\quad \max_i \{\sqrt{N}\left\|\hat{\mathbf{H}}_{(i),l}^T \Delta \mathbf{P}_{(i),l} \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_{(i),l}^T \mathbf{H}^\star{}_{(i),l} \mathbf{H}^{\star T}_{(i),l} \mathbf{T}_{(i)} \Delta \mathbf{P}_{(i),l} \hat{\mathbf{H}}_{(i),l}\right\|\}\} \\ &\leq \sqrt{N} 2\sigma_{\max}^2 \frac{5}{2}\sqrt{\alpha}\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \frac{\sigma_{\max}}{\sigma_{\min}^2}\left(1 + 4\frac{\sigma_{\max}^2}{\sqrt{\theta}\sigma_{\min}^2}\right). \end{aligned}$$

The summation of the operator norm of the off-diagonal blocks of $\mathbf{\Lambda}_{7,3}$ is

$$\begin{aligned} &\sqrt{2\sum_{i=1}^N \left\|\hat{\mathbf{H}}_g^T \Delta \mathbf{P}_g \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(i)} \Delta \mathbf{P}_{(i),l} \hat{\mathbf{H}}_{(i),l}\right\|^2} \\ &\leq \sqrt{2N} 2\sigma_{\max}^2 \frac{5}{2}\sqrt{\alpha}\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \frac{\sigma_{\max}}{\sigma_{\min}^2}\left(1 + 4\frac{\sigma_{\max}^2}{\sqrt{\theta}\sigma_{\min}^2}\right). \end{aligned}$$

As a result, Lemma 26 implies that,

$$\begin{aligned} \|\mathbf{\Lambda}_{7,3}\| &\leq \sqrt{N} 10\sigma_{\max}^2 \sqrt{\alpha}\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \frac{\sigma_{\max}}{\sigma_{\min}^2}\left(1 + 4\frac{\sigma_{\max}^2}{\sqrt{\theta}\sigma_{\min}^2}\right) \\ &\leq \frac{\sqrt{N}\sigma_{\min}^2}{16}, \end{aligned} \tag{38}$$

where we applied the condition $\alpha \leq \frac{1}{(\mu^2 r 640)^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^8 \left(1 + \frac{4\sigma_{\max}^2}{\sqrt{\theta}\sigma_{\min}^2}\right)^{-2}$ in the last inequality.

We proceed to estimate the eigenvalue lower bound for $\mathbf{\Lambda}_{7,4}$. We first factorize $\mathbf{\Lambda}_{7,4}$ as,

$$\begin{aligned} \mathbf{\Lambda}_{7,4} = &\operatorname{Diag}\left(\hat{\mathbf{H}}_g^T \mathbf{H}^\star{}_g, \hat{\mathbf{H}}_{(1),l}^T \mathbf{H}^\star{}_{(1),l}, \cdots, \hat{\mathbf{H}}_{(N),l}^T \mathbf{H}^\star{}_{(N),l}\right) \\ &\times \begin{pmatrix} \sqrt{N}\mathbf{\Lambda}^\star{}_1 & \mathbf{\Lambda}^\star{}_{3,(1)} & \cdots & \mathbf{\Lambda}^\star{}_{3,(N)} \\ \mathbf{\Lambda}^{\star T}_{3,(1)} & \sqrt{N}\mathbf{\Lambda}^\star{}_{2,(1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Lambda}^{\star T}_{3,(N)} & 0 & \cdots & \sqrt{N}\mathbf{\Lambda}^\star{}_{2,(N)} \end{pmatrix} \end{aligned}$$

$$\times \operatorname{Diag}\left(\hat{\mathbf{H}}^{\star T}_{g}\hat{\mathbf{H}}_{g}, \mathbf{H}^{\star T}_{(1),l}\hat{\mathbf{H}}_{(1),l}, \cdots, \mathbf{H}^{\star T}_{(N),l}\hat{\mathbf{H}}_{(N),l}\right).$$

Lemma 24 and (36) indicate that $\mathbf{\Lambda^{\star}}_{1,(i)} - \sigma^2_{\min}\mathbf{I} \succ 0$, $\mathbf{\Lambda^{\star}}_{2,(i)} - \sigma^2_{\min}\mathbf{I} \succ 0$, and

$$\left(\mathbf{\Lambda^{\star}}_{1,(i)} - \sigma^2_{\min}\mathbf{I}\right) - \mathbf{\Lambda^{\star T}}_{3,(i)}\left(\mathbf{\Lambda^{\star}}_{2,(i)} - \sigma^2_{\min}\mathbf{I}\right)^{-1}\mathbf{\Lambda^{\star}}_{3,(i)} \succeq 0. \tag{39}$$

Summing both sides of (39) for $i = 1$ to N, we know, $N\mathbf{\Lambda^{\star}}_1 - N\sigma^2_{\min}\mathbf{I} - \sum_{i=1}^{N}\mathbf{\Lambda^{\star T}}_{3,(i)}\left(\mathbf{\Lambda^{\star}}_{2,(i)} - \sigma^2_{\min}\mathbf{I}\right)^{-1}\mathbf{\Lambda^{\star}}_{3,(i)} \succeq 0$, which is equivalent to $\sqrt{N}\mathbf{\Lambda^{\star}}_1 - \sqrt{N}\sigma^2_{\min}\mathbf{I} - \sum_{i=1}^{N}\mathbf{\Lambda^{\star T}}_{3,(i)}\left(\sqrt{N}\mathbf{\Lambda^{\star}}_{2,(i)} - \sqrt{N}\sigma^2_{\min}\mathbf{I}\right)^{-1}\mathbf{\Lambda^{\star}}_{3,(i)} \succeq 0$. Again, Lemma 24 indicates,

$$\begin{pmatrix} \sqrt{N}\mathbf{\Lambda^{\star}}_1 & \mathbf{\Lambda^{\star}}_{3,(1)} & \cdots & \mathbf{\Lambda^{\star}}_{3,(N)} \\ \mathbf{\Lambda^{\star T}}_{3,(1)} & \sqrt{N}\mathbf{\Lambda^{\star}}_{2,(1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Lambda^{\star T}}_{3,(N)} & 0 & \cdots & \sqrt{N}\mathbf{\Lambda^{\star}}_{2,(N)} \end{pmatrix} \succeq \sqrt{N}\sigma^2_{\min}\mathbf{I}.$$

On the other hand, we know form Lemma 26 that,

$$\begin{aligned}
&\left\|\mathbf{I} - \operatorname{Diag}\left(\hat{\mathbf{H}}^{T}_{g}\mathbf{H}^{\star}{}_{g}\hat{\mathbf{H}}^{\star T}_{g}\hat{\mathbf{H}}_{g}, \hat{\mathbf{H}}^{T}_{(1),l}\mathbf{H}^{\star}{}_{(1),l}\mathbf{H}^{\star T}_{(1),l}\hat{\mathbf{H}}_{(1),l}, \cdots, \hat{\mathbf{H}}^{T}_{(N),l}\mathbf{H}^{\star}{}_{(N),l}\mathbf{H}^{\star T}_{(N),l}\hat{\mathbf{H}}_{(N),l}\right)\right\| \\
&\leq \max\{\|\Delta\mathbf{P}_g\|, \max_j \left\|\Delta\mathbf{P}_{(j),l}\right\|\} \\
&\leq \frac{\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty}\left(\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max}\right)}{\sigma^2_{\min}}\left(1 + \frac{8\sigma^2_{\max}}{\sqrt{\theta}\sigma^2_{\min}}\right) \leq \frac{1}{8},
\end{aligned}$$

where we applied the condition $\alpha \leq \frac{\sigma^2_{\min}}{80\mu^2 r\sigma^2_{\max}}\left(1 + \frac{8\sigma^2_{\max}}{\sqrt{\theta}\sigma^2_{\min}}\right)^{-1}$ in the last inequality.

Hence, Lemma 25 indicates,

$$\lambda_{\min}\left(\mathbf{\Lambda}_{7,4}\right) \geq \sqrt{N}\sigma^2_{\min}\frac{7}{8}.$$

By Wely's theorem, we know that

$$\begin{aligned}
\lambda_{\min}\left(\mathbf{\Lambda}_7\right) &\geq \sqrt{N}\sigma^2_{\min}\frac{7}{8} - \|\mathbf{\Lambda}_{7,1} + \mathbf{\Lambda}_{7,3}\| \\
&\geq \sqrt{N}\sigma^2_{\min}\frac{3}{4},
\end{aligned} \tag{40}$$

where we applied the inequality (37) and (38) in the last inequality.

Step 2: Minimum eigenvalue of $\mathbf{\Lambda}_6$: The inequality (40) is equivalent to $\mathbf{\Lambda}_7 - \sqrt{N}\sigma^2_{\min}\frac{3}{4}\mathbf{I} \succ 0$. Thus, Lemma 24 implies,

$$\sqrt{N}\mathbf{\Lambda}_1 - \sqrt{N}\sigma^2_{\min}\frac{3}{4}\mathbf{I} - \sum_{i=1}^{N}\mathbf{\Lambda}^{T}_{3,(i)}\left(\sqrt{N}\mathbf{\Lambda}_{2,(i)} - \sqrt{N}\sigma^2_{\min}\frac{3}{4}\mathbf{I}\right)^{-1}\mathbf{\Lambda}_{2,(i)} \succeq 0.$$

Lemma 16 already shows that $\mathbf{\Lambda}_{2,(i)} \succ \sigma^2_{\min}\frac{3}{4}\mathbf{I}$. As a result,

$$\mathbf{\Lambda}^{T}_{3,(i)}\left(\sqrt{N}\mathbf{\Lambda}_{2,(i)} - \sqrt{N}\sigma^2_{\min}\frac{3}{4}\mathbf{I}\right)^{-1}\mathbf{\Lambda}_{2,(i)}$$

$$= \frac{1}{\sqrt{N}} \sum_{p=0}^{\infty} \mathbf{\Lambda}_{3,(i)}^T \left(\mathbf{\Lambda}_{2,(i)}^{-1} + \left(\sum_{p=0}^{\infty} \sigma_{\min}^2 \frac{3}{4} \mathbf{\Lambda}_{2,(i)}^{-1} \right)^p \right) \mathbf{\Lambda}_{2,(i)}$$
$$\succeq \frac{1}{\sqrt{N}} \mathbf{\Lambda}_{3,(i)}^T \mathbf{\Lambda}_{2,(i)}^{-1} \mathbf{\Lambda}_{3,(i)}.$$

By rearranging terms, we have,

$$\sqrt{N}\mathbf{\Lambda}_1 - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{\Lambda}_{3,(i)}^T \mathbf{\Lambda}_{2,(i)}^{-1} \mathbf{\Lambda}_{2,(i)} \succeq \sqrt{N} \sigma_{\min}^2 \frac{3}{4} \mathbf{I}.$$

This completes our proof. ■

With an understanding of spectral properties of $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_{2,(i)}$, and $\mathbf{\Lambda}_6$, we are now ready to characterize the solutions to the KKT conditions and provide a proof for Lemma 8. To this goal, we first write the solutions to (13) into Taylor-like series.

Lemma 18 *For every $i \in [N]$, suppose that $\mathbf{U}_{(i),l}^{\star}$'s are θ-misaligned, $\max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \leq 4\sigma_{\max}\frac{\mu^2 r}{n}$, and $\mathbf{E}_{(i)}$ is α-sparse with $\alpha \leq \frac{1}{40\mu^2 r}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{-3}$. The solutions to (13) satisfy the following,*

$$\begin{aligned}
\hat{\mathbf{H}}_g &= \hat{\mathbf{H}}_{g,0} + \hat{\mathbf{H}}_{g,1} \\
&+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}} \right) \right] \hat{\mathbf{H}}_{g,0} \\
&\times \prod_{l=k}^{0} \left(\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1} \right) \\
&+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}} \right) \right] \hat{\mathbf{H}}_{g,1} \\
&\times \prod_{l=k}^{0} \left(\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1} \right),
\end{aligned} \tag{41}$$

and

$$\begin{aligned}
\hat{\mathbf{H}}_{(j),l} &= \hat{\mathbf{H}}_{(i),l,0} + \sum_{p=1}^{\infty} \mathbf{F}_{(j)}^{p} \mathbf{T}_{(j)} \hat{\mathbf{H}}_{(j),l} \mathbf{\Lambda}_{2,(j)}^{-p-1} + \hat{\mathbf{H}}_{g,1} \mathbf{\Lambda}_{4,(j)} \mathbf{\Lambda}_{2,(j)}^{-1} \\
&+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}} \right) \right] \hat{\mathbf{H}}_{g,0} \\
&\times \prod_{l=k}^{0} \left(\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1} \right) \mathbf{\Lambda}_{4,(j)} \mathbf{\Lambda}_{2,(j)}^{-1}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=0}^{\infty}\sum_{p_0+p_1\geq 1}^{\infty}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty}\sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N}\left[\prod_{l=0}^{k}\left(\mathbf{F}_{(0)}^{p_{2l}}\mathbf{F}_{(i_{2l+1})}^{p_{2l+1}}\right)\right]\hat{\mathbf{H}}_{g,1}\\
&\times\prod_{l=k}^{0}\left(\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1}\mathbf{\Lambda}_{5,(i_{2l+1})}\mathbf{\Lambda}_{1}^{-p_{2l}}\mathbf{\Lambda}_{6}^{-1}\right)\mathbf{\Lambda}_{4,(j)}\mathbf{\Lambda}_{2,(j)}^{-1}\\
&+\sum_{p=1}^{\infty}\mathbf{F}_{(j)}^{p}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{4,(j)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{4,(j)}\mathbf{\Lambda}_{2,(j)}^{-1},
\end{aligned}
\tag{42}
$$

where $\hat{\mathbf{H}}_{g,0}$ *is defined as*

$$
\hat{\mathbf{H}}_{g,0}=\mathbf{T}_{(0)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{6}^{-1}+\sum_{i=1}^{N}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_{6}^{-1},
\tag{43}
$$

$\hat{\mathbf{H}}_{g,1}$ *is defined as*

$$
\hat{\mathbf{H}}_{g,1}=\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{T}_{(i_1)}\hat{\mathbf{H}}_{(i_1),l}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_{1}^{-p_0}\mathbf{\Lambda}_{6}^{-1},
\tag{44}
$$

and $\hat{\mathbf{H}}_{(i),l,0}$ *is defined as*

$$
\hat{\mathbf{H}}_{(i),l,0}=\mathbf{T}_{(j)}\hat{\mathbf{H}}_{(j),1}\mathbf{\Lambda}_{2}^{-1}+\hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(j)}\mathbf{\Lambda}_{2,(j)}^{-1}
\tag{45}
$$

.

Proof Notice that as we defined $\mathbf{\Lambda}_{4,(i)}=-\mathbf{\Lambda}_{3,(i)}$ and $\mathbf{\Lambda}_{5,(i)}=-\mathbf{\Lambda}_{3,(i)}^{T}/N$, the KKT condition in (13) can be written as the following Sylvester equations

$$
\begin{cases}
\mathbf{F}_{(i)}\hat{\mathbf{H}}_{(i),l}-\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}=-\left(\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}+\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{4,(i)}\right)\\
\mathbf{F}_{(0)}\hat{\mathbf{H}}_{g}-\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{1}=-\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_{g}+\sum_{i=1}^{N}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{5,(i)}\right).
\end{cases}
\tag{46}
$$

Note that $\sigma_{min}(\mathbf{\Lambda}_{2,(i)})>\left\|\mathbf{F}_{(i)}\right\|$ and $\sigma_{min}(\mathbf{\Lambda}_{1,(i)})>\left\|\mathbf{F}_{(0)}\right\|$. Therefore, according to Bhatia (2013, Theorem VII.2.2), the solution to (46) satisfies the following equation

$$
\begin{cases}
\hat{\mathbf{H}}_{g}=\sum_{p=0}^{\infty}\mathbf{F}_{(0)}^{p}\mathbf{T}_{(0)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{1}^{-p-1}+\sum_{p=0}^{\infty}\sum_{i=1}^{N}\mathbf{F}_{(0)}^{p}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_{1}^{-p-1}\\
\hat{\mathbf{H}}_{(i),l}=\sum_{p=0}^{\infty}\mathbf{F}_{(i)}^{p}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-p-1}+\sum_{p=0}^{\infty}\mathbf{F}_{(i)}^{p}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1}
\end{cases}.
\tag{47}
$$

We can substitute $\hat{\mathbf{H}}_{(i),l}$ in the right hand side of the first equation of (47) by the second equation in (47)

$$
\hat{\mathbf{H}}_{g}=\sum_{p=0}^{\infty}\mathbf{F}_{(0)}^{p}\mathbf{T}_{(0)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{1}^{-p-1}+\sum_{p_0=0}^{\infty}\sum_{p_1=0}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{T}_{(i_1)}\hat{\mathbf{H}}_{(i_1),l}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_{1}^{-p_0-1}
$$

$$
\begin{aligned}
&+\sum_{p_0=0}^{\infty}\sum_{p_1=0}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i_1)}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0-1}\\
&=\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_1^{-1}+\sum_{i=1}^{N}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_1^{-1}+\sum_{i=1}^{N}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_1^{-1}\\
&+\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{T}_{(i_1)}\hat{\mathbf{H}}_{(i_1),l}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0-1}\\
&+\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i_1)}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0-1}.
\end{aligned}
\tag{48}
$$

We then move the third term on the right hand side of (48) to the left hand side, multiply both sides by $\mathbf{\Lambda}_1\mathbf{\Lambda}_6^{-1}$ on the right, and recall the definition of $\mathbf{\Lambda}_{6,(i)}$ in (30), we have,

$$
\begin{aligned}
\hat{\mathbf{H}}_g=&\underbrace{\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{i=1}^{N}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}}_{\hat{\mathbf{H}}_{g,0}}\\
&+\underbrace{\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{T}_{(i_1)}\hat{\mathbf{H}}_{(i_1),l}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0}\mathbf{\Lambda}_6^{-1}}_{\hat{\mathbf{H}}_{g,1}}\\
&+\underbrace{\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i_1)}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0}\mathbf{\Lambda}_6^{-1}}_{\text{residual term}}.
\end{aligned}
\tag{49}
$$

On the right hand side of (49), one can see that the $\hat{\mathbf{H}}_{g,0}$ and $\hat{\mathbf{H}}_{g,1}$ are products of sparse matrices, incoherent matrices, and remaining terms. Therefore we can use Lemma 13 to calculate an upper bound on their maximum row norm. However, the residual term does not have such specific structure as we do not know whether $\hat{\mathbf{H}}_g$ is incoherent. As a result we cannot provide precise estimate on its maximum row norm directly. To circumvent the issue, notice that (49) has a recursive form. Therefore, the residual term can be replaced by

$$
\hat{\mathbf{H}}_g\rightarrow\hat{\mathbf{H}}_{g,0}+\hat{\mathbf{H}}_{g,1}+\sum_{p_0+p_1\geq 1}^{\infty}\sum_{i_1=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i_1)}\mathbf{\Lambda}_{2,(i_1)}^{-p_1-1}\mathbf{\Lambda}_{5,(i_1)}\mathbf{\Lambda}_1^{-p_0}\mathbf{\Lambda}_6^{-1}. \tag{50}
$$

The result will have 5 terms, the first 4 of which have the structure specified in Lemma 13. The 5-th term does not as it contains $\hat{\mathbf{H}}_g$. We can apply the replacement rule (50) again for the 5-th term, generating 7 terms. After applying the replacement rule ω times, where ω

is an integer, the results become,

$$\begin{aligned}
&\hat{\mathbf{H}}_g = \hat{\mathbf{H}}_{g,0} + \hat{\mathbf{H}}_{g,1} \\
&+\sum_{k=0}^{\omega}\sum_{p_0+p_1\geq 1}^{\infty}\sum_{p_2+p_3\geq 1}^{\infty}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty}\sum_{i_1=1}^{N}\sum_{i_3=1}^{N}\cdots\sum_{i_{2k+1}=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{F}_{(0)}^{p_2}\mathbf{F}_{(i_3)}^{p_3}\cdots\mathbf{F}_{(0)}^{p_{2k}}\mathbf{F}_{(i_{2k+1})}^{p_{2k+1}} \\
&\hat{\mathbf{H}}_{g,0}\boldsymbol{\Lambda}_{4,(i_{2k+1})}\boldsymbol{\Lambda}_{2,(i_{2k+1})}^{-p_{2k-1}-1}\boldsymbol{\Lambda}_{5,(i_{2k+1})}\boldsymbol{\Lambda}_1^{-p_{2k-2}}\boldsymbol{\Lambda}_6^{-1}\cdots\boldsymbol{\Lambda}_{4,(i_1)}\boldsymbol{\Lambda}_{2,(i_1)}^{-p_1-1}\boldsymbol{\Lambda}_{5,(i_1)}\boldsymbol{\Lambda}_1^{-p_0}\boldsymbol{\Lambda}_6^{-1} \\
&+\sum_{k=0}^{\omega}\sum_{p_0+p_1\geq 1}^{\infty}\sum_{p_2+p_3\geq 1}^{\infty}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty}\sum_{i_1=1}^{N}\sum_{i_3=1}^{N}\cdots\sum_{i_{2k+1}=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{F}_{(0)}^{p_2}\mathbf{F}_{(i_3)}^{p_3}\cdots\mathbf{F}_{(0)}^{p_{2k}}\mathbf{F}_{(i_{2k+1})}^{p_{2k+1}} \\
&\hat{\mathbf{H}}_{g,1}\boldsymbol{\Lambda}_{4,(i_{2k+1})}\boldsymbol{\Lambda}_{2,(i_{2k+1})}^{-p_{2k-1}-1}\boldsymbol{\Lambda}_{5,(i_{2k+1})}\boldsymbol{\Lambda}_1^{-p_{2k-2}}\boldsymbol{\Lambda}_6^{-1}\cdots\boldsymbol{\Lambda}_{4,(i_1)}\boldsymbol{\Lambda}_{2,(i_1)}^{-p_1-1}\boldsymbol{\Lambda}_{5,(i_1)}\boldsymbol{\Lambda}_1^{-p_0}\boldsymbol{\Lambda}_6^{-1} \\
&+\sum_{p_0+p_1\geq 1}^{\infty}\sum_{p_2+p_3\geq 1}^{\infty}\cdots\sum_{p_{2\omega+2}+p_{2\omega+3}\geq 1}^{\infty}\sum_{i_1=1}^{N}\sum_{i_3=1}^{N}\cdots\sum_{i_{2\omega+3}=1}^{N}\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{F}_{(0)}^{p_2}\mathbf{F}_{(i_3)}^{p_3}\cdots\mathbf{F}_{(0)}^{p_{2\omega+2}}\mathbf{F}_{(i_{2\omega+3})}^{p_{2\omega+3}} \\
&\hat{\mathbf{H}}_g\boldsymbol{\Lambda}_{4,(i_{2\omega+3})}\boldsymbol{\Lambda}_{2,(i_{2\omega+3})}^{-p_{2\omega+3}-1}\boldsymbol{\Lambda}_{5,(i_{2\omega+3})}\boldsymbol{\Lambda}_1^{-p_{2\omega+2}}\boldsymbol{\Lambda}_6^{-1}\cdots\boldsymbol{\Lambda}_{4,(i_1)}\boldsymbol{\Lambda}_{2,(i_1)}^{-p_1-1}\boldsymbol{\Lambda}_{5,(i_1)}\boldsymbol{\Lambda}_1^{-p_0}\boldsymbol{\Lambda}_6^{-1},
\end{aligned} \tag{51}$$

which holds for any integer $\omega \geq 0$.

Recall that our goal is to write $\hat{\mathbf{H}}_g$ in a form with which we can easily determine its maximum row norm. By observing (51), we know Lemma 13 can be applied to estimate the maximum row norm of all but the last terms. The last summation term still cannot be handled by Lemma 13 directly. To resolve the issue, we take an alternative route to use ω to control the last summation term.

We claim that under the provided upper bound for α, the last term will approach zero in the limit $\omega \to \infty$. To see this, note that Lemma 16 and Lemma 17 show that $\sigma_{\min}(\boldsymbol{\Lambda}_1)$, $\sigma_{\min}(\boldsymbol{\Lambda}_{2,(i)})$, and $\sigma_{\min}(\boldsymbol{\Lambda}_6)$ are lower bounded by $\frac{3}{4}\sigma_{\min}^2$. Since $\left\|\mathbf{F}_{(i)}\right\| \leq \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty (2\sigma_{\max} + \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty) \leq \alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \frac{5}{2}\sigma_{\max}$ for each i, we have,

$$\begin{aligned}
&\sum_{p_0+p_1\geq 1}^{\infty}\cdots\sum_{p_{2\omega+2}+p_{2\omega+3}\geq 1}^{\infty}\sum_{i_1=1}^{N}\sum_{i_3=1}^{N}\cdots\sum_{i_{2\omega+3}=1}^{N}\left\|\mathbf{F}_{(0)}^{p_0}\mathbf{F}_{(i_1)}^{p_1}\mathbf{F}_{(0)}^{p_2}\mathbf{F}_{(i_3)}^{p_3}\cdots\mathbf{F}_{(i_{2\omega+3})}^{p_{2\omega+3}}\right\|_F \\
&\left\|\hat{\mathbf{H}}_g\boldsymbol{\Lambda}_{4,(i_{2\omega+3})}\boldsymbol{\Lambda}_{2,(i_{2\omega+3})}^{-p_{2\omega+3}-1}\boldsymbol{\Lambda}_{5,(i_{2\omega+3})}\boldsymbol{\Lambda}_1^{-p_{2\omega+2}}\boldsymbol{\Lambda}_6^{-1}\cdots\boldsymbol{\Lambda}_{4,(i_1)}\boldsymbol{\Lambda}_{2,(i_1)}^{-p_1-1}\boldsymbol{\Lambda}_{5,(i_1)}\boldsymbol{\Lambda}_1^{-p_0}\boldsymbol{\Lambda}_6^{-1}\right\| \\
&\leq \sum_{p_0+p_1\geq 1}^{\infty}\cdots\sum_{p_{2\omega+2}+p_{2\omega+3}\geq 1}^{\infty}\left(\frac{\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \left(\frac{5}{2}\sigma_{\max}\right)}{\frac{3}{4}\sigma_{\min}^2}\right)^{p_0+\cdots+p_{2\omega+3}} \\
&\times\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{2(\omega+2)} \\
&\leq \left(4\frac{\alpha\overline{n}\max_i \left\|\mathbf{E}_{(i)}\right\|_\infty \left(\frac{5}{2}\sigma_{\max}\right)}{\frac{3}{4}\sigma_{\min}^2}\right)^{2(\omega+2)}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{2(\omega+2)} \\
&\leq \left(\frac{1}{2}\right)^{2(\omega+2)},
\end{aligned}$$

where we used Lemma 23 in the first inequality and the condition that $\alpha \leq \frac{1}{40\mu^2 r}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{-3}$ in the last inequality.

Therefore, we can take the limit $\omega \to \infty$ in (51) and rewrite it as a series. The series is absolutely convergent when α is small. Finally, we prove (41). Though (41) is an infinite series, each term in the series is the product of sparse matrices and an incoherent matrix. Such structure will be useful later when we use Lemma 13 to calculate the maximum row norm of $\hat{\mathbf{H}}_g$.

Now we proceed to derive an expansion for $\hat{\mathbf{H}}_{(i),l}$. We can replace $\hat{\mathbf{H}}_g$ on the right hand side of the second equation of (47) with (51) to derive,

$$\begin{aligned}
\hat{\mathbf{H}}_{(i),l} &= \mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1} + \hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1} \\
&+ \sum_{p=1}^{\infty}\mathbf{F}_{(i)}^{p}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-p-1} + \sum_{p=1}^{\infty}\mathbf{F}_{(i)}^{p}\hat{\mathbf{H}}_g\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1} \\
&+ \sum_{k=0}^{\infty}\sum_{p_0+p_1\geq 1}^{\infty}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty}\sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N}\left[\prod_{l=0}^{k}\left(\mathbf{F}_{(0)}^{p_{2l}}\mathbf{F}_{(i_{2l+1})}^{p_{2l+1}}\right)\right]\hat{\mathbf{H}}_{g,0} \\
&\times\prod_{l=k}^{0}\left(\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1}\mathbf{\Lambda}_{5,(i_{2l+1})}\mathbf{\Lambda}_1^{-p_{2l}}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1} \\
&+ \sum_{k=0}^{\infty}\sum_{p_0+p_1\geq 1}^{\infty}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty}\sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N}\left[\prod_{l=0}^{k}\left(\mathbf{F}_{(0)}^{p_{2l}}\mathbf{F}_{(i_{2l+1})}^{p_{2l+1}}\right)\right]\hat{\mathbf{H}}_{g,1} \\
&\times\prod_{l=k}^{0}\left(\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{4,(i_{2l+1})}\mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1}\mathbf{\Lambda}_{5,(i_{2l+1})}\mathbf{\Lambda}_1^{-p_{2l}}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1}.
\end{aligned} \tag{52}$$

We thus prove (42). ■

In Lemma 18, although the series of $\hat{\mathbf{H}}_g$ and $\hat{\mathbf{H}}_{(i),l}$'s have infinite terms, when α is not too large, the leading term is only the first term. This is delineated in the following lemma, which is a formal version of Lemma 8. For simplicity, we introduce a notation

$$\zeta = \frac{\alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty}\left(\alpha\overline{n}\max_i\left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max}\right)}{\frac{3}{4}\sigma_{\min}^2}. \tag{53}$$

Lemma 19 *Suppose that the conditions of Lemma 18 are satisfied. Additionally, suppose that* $\alpha \leq \frac{6-3\sqrt{2}}{80}\frac{1}{\mu^2 r}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^2$, *we have*

$$\begin{aligned}
\hat{\mathbf{H}}_g &= \hat{\mathbf{H}}_{g,0} + \delta\mathbf{H}_g \\
\hat{\mathbf{H}}_{(i),l} &= \hat{\mathbf{H}}_{(i),l,0} + \delta\mathbf{H}_{(i),l}
\end{aligned}$$

where $\delta\mathbf{H}_g$ *and* $\delta\mathbf{H}_{(i),l}$ *satisfy*

$$\|\delta\mathbf{H}_g\| \leq \zeta \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2 \left(1 + 2\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right) + 2\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right) \tag{54}$$

$$\max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_g\right\| \leq \zeta\sqrt{\frac{\mu^2 r}{n_1}} 4 \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2 \left(1 + \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right) + \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right) \tag{55}$$

$$\left\|\delta\mathbf{H}_{(i),l}\right\| \leq \zeta \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \left(2 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} + 3\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2 + 4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^3 + 4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^4\right) \tag{56}$$

$$\begin{aligned}\max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_{(i),l}\right\| \leq{}& \zeta\sqrt{\frac{\mu^2 r}{n_1}} 2 \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \\ &\times \left(1 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} + 5\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2 + 4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^3 + 4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^4\right),\end{aligned} \tag{57}$$

with ζ *is defined in* (53).

Proof We need to provide upper bounds on the series in Lemma 18. From Lemma 18, $\hat{\mathbf{H}}_g$ can be written as a series. We can define $\delta\mathbf{H}_g$ as the summation of all but the first term in the series, as in

$$\begin{aligned}\delta\mathbf{H}_g = \hat{\mathbf{H}}_{g,1} &+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}}\right)\right] \hat{\mathbf{H}}_{g,0} \\ &\times \prod_{l=k}^{0} \left(\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1}\right) \\ &+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}}\right)\right] \hat{\mathbf{H}}_{g,1} \\ &\times \prod_{l=k}^{0} \left(\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1}\right).\end{aligned}$$

Hence, by applying Lemma 22, we have

$$\begin{aligned}\|\delta\mathbf{H}_g\| \leq{}& \left\|\hat{\mathbf{H}}_{g,1}\right\| + \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \\ &\left[\prod_{l=0}^{k} \left(\left\|\mathbf{F}_{(0)}\right\|^{p_{2l}} \left\|\mathbf{F}_{(i_{2l+1})}\right\|^{p_{2l+1}}\right)\right] \left\|\hat{\mathbf{H}}_{g,0}\right\| \prod_{l=k}^{0} \left(\left\|\mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1}\right\|\right) \\ &+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \left[\prod_{l=0}^{k} \left(\left\|\mathbf{F}_{(0)}\right\|^{p_{2l}} \left\|\mathbf{F}_{(i_{2l+1})}\right\|^{p_{2l+1}}\right)\right] \left\|\hat{\mathbf{H}}_{g,1}\right\|\end{aligned}$$

$$\times \prod_{l=k}^{0} \left\| \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_{1}^{-p_{2l}} \mathbf{\Lambda}_{6}^{-1} \right\|.$$

We first estimate an upper bound for $\left\|\hat{\mathbf{H}}_{g,1}\right\|$,

$$\begin{aligned}
\left\|\hat{\mathbf{H}}_{g,1}\right\| &\leq \sum_{p_0+p_1\geq 1}^{\infty} \sum_{i_1=1}^{N} \left\|\mathbf{F}_{(0)}\right\|^{p_0} \left\|\mathbf{F}_{(i_1)}\right\|^{p_1} \left\| \mathbf{T}_{(i_1)} \hat{\mathbf{H}}_{(i_1),l} \mathbf{\Lambda}_{2,(i_1)}^{-p_1-1} \mathbf{\Lambda}_{5,(i_1)} \mathbf{\Lambda}_{1}^{-p_0} \mathbf{\Lambda}_{6}^{-1} \right\| \\
&\leq \sum_{p_0+p_1\geq 1}^{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max} \right) \right)^{p_0+p_1} \\
&\times 2 \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \left(\frac{1}{\frac{3}{4}\sigma_{\min}^2} \right)^{p_0+p_1} \\
&\leq \frac{\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max} \right)}{\frac{3}{4}\sigma_{\min}^2} 2 \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \\
&2 \left(1 - \frac{\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max} \right)}{\frac{3}{4}\sigma_{\min}^2} \right)^{-1} \\
&\leq \frac{\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max} \right)}{\frac{3}{4}\sigma_{\min}^2} 4\sqrt{2} \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2,
\end{aligned}$$

where we used Lemma 22 in the first inequality, the upper bound on $\left\|\mathbf{F}_{(i)}\right\|$ in the second inequality. Because of the upper bound on α, we can use auxiliary Lemma 23 to derive an upper bound on the series. The last inequality comes from the fact that $(1 - \frac{\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} \left(\alpha\overline{n} \max_i \left\|\mathbf{E}_{(i)}\right\|_{\infty} + 2\sigma_{\max} \right)}{\frac{3}{4}\sigma_{\min}^2})^{-1} \leq \sqrt{2}$.

Therefore, we can proceed to estimate,

$$\begin{aligned}
&\|\delta\mathbf{H}_g\| \\
&\leq \zeta 4\sqrt{2} \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 + \sum_{k=0}^{\infty} \left(\zeta \frac{2}{1-\zeta} \right)^{k+1} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^{2(k+1)} \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 2\zeta \frac{2}{1-\zeta} \\
&+ \sum_{k=0}^{\infty} \left(\zeta \frac{2}{1-\zeta} \right)^{k+1} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^{2(k+1)} \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right) \left(1 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right) \\
&\leq \zeta 4\sqrt{2} \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \left(1 - \frac{2\zeta}{1-\zeta} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \right)^{-1} \\
&+ \zeta \frac{2}{1-\zeta} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \left(\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right) \left(1 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right) \left(1 - \frac{2\zeta}{1-\zeta} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2 \right)^{-1}.
\end{aligned}$$

We can estimate an upper bound on $\max_k \left\|\mathbf{e}_k^T \delta\mathbf{H}_g\right\|$ in a similar fashion.

We first show that, for any $j = 1, 2, \cdots, N$,

$$\begin{aligned}
&\max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{T}_{(j)} \right\| \\
&= \max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \left(\mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g + \mathbf{H}^\star{}_{(j),l} \mathbf{H}^{\star T}_{(j),l} \right) \mathbf{T}_{(j)} \right\| \\
&\leq \max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(j)} \right\| + \max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{H}^\star{}_g \mathbf{H}^{\star T}_g \mathbf{T}_{(j)} \right\| \\
&\leq 2\sigma_{\max}^2 \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty + 2\sigma_{\max} \right) \right)^{\sum_{m=1}^k p_m},
\end{aligned} \tag{58}$$

where we used the triangle inequality in the first inequality, and Lemma 13 together with $r = r_1 + r_2$ in the second inequality.

A similar equality also holds for $\max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{T}_{(0)} \right\|$:

$$\begin{aligned}
&\max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{T}_{(0)} \right\| = \max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \frac{1}{N} \sum_{j=1}^{N} \mathbf{T}_{(j)} \right\| \leq \frac{1}{N} \sum_{j=1}^{N} \max_k \left\| \mathbf{e}_k^T \prod_{\ell=1}^{k} \mathbf{F}_{(i_\ell)}^{p_\ell} \mathbf{T}_{(j)} \right\| \\
&\leq 2\sigma_{\max}^2 \sqrt{\frac{\mu^2 r}{n_1}} \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty \left(\alpha \overline{n} \max_i \left\| \mathbf{E}_{(i)} \right\|_\infty + 2\sigma_{\max} \right) \right)^{\sum_{m=1}^k p_m}.
\end{aligned} \tag{59}$$

Combining the above two inequalities, we have

$$\begin{aligned}
&\max_j \left\| \mathbf{e}_j^T \delta\mathbf{H}_g \right\| \\
&\leq \max_j \left\| \mathbf{e}_j^T \hat{\mathbf{H}}_{g,1} \right\| + \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \\
&\max_j \left\| \mathbf{e}_j^T \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}} \right) \right] \hat{\mathbf{H}}_{g,0} \right\| \prod_{l=k}^{0} \left\| \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1} \right\| \\
&+ \sum_{k=0}^{\infty} \sum_{p_0+p_1\geq 1}^{\infty} \cdots \sum_{p_{2k}+p_{2k+1}\geq 1}^{\infty} \sum_{i_1,i_3,\cdots,i_{2k+1}=1}^{N} \max_j \left\| \mathbf{e}_j^T \left[\prod_{l=0}^{k} \left(\mathbf{F}_{(0)}^{p_{2l}} \mathbf{F}_{(i_{2l+1})}^{p_{2l+1}} \right) \right] \hat{\mathbf{H}}_{g,1} \right\| \\
&\times \prod_{l=k}^{0} \left\| \mathbf{\Lambda}_{4,(i_{2l+1})} \mathbf{\Lambda}_{2,(i_{2l+1})}^{-p_{2l+1}-1} \mathbf{\Lambda}_{5,(i_{2l+1})} \mathbf{\Lambda}_1^{-p_{2l}} \mathbf{\Lambda}_6^{-1} \right\| \\
&\leq \sum_{p_0+p_1\geq 1} \zeta^{p_0+p_1} \sqrt{\frac{\mu^2 r}{n_1}} \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \right)^2
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=0}^{\infty}\sum_{p_0+p_1\geq 1}\cdots\sum_{p_{2k+2}+p_{2k+3}\geq 1}\zeta^{p_0+\cdots+p_{2k+3}}\sqrt{\frac{\mu^2 r}{n_1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{2(k+1)}\\
&+\sum_{k=0}^{\infty}\sum_{p_0+p_1\geq 1}\cdots\sum_{p_{2k}+p_{2k+1}\geq 1}\zeta^{p_0+\cdots+p_{2k+1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^{2(k+1)}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\left(1+\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right)\\
&\leq \zeta\frac{2}{1-\zeta}\sqrt{\frac{\mu^2 r}{n_1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2+\sum_{k=0}^{\infty}\left(\frac{2\zeta}{1-\zeta}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right)^{k+1}\sqrt{\frac{\mu^2 r}{n_1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\\
&+\sum_{k=0}^{\infty}\left(\frac{2\zeta}{1-\zeta}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right)^{k+1}\sqrt{\frac{\mu^2 r}{n_1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\\
&\leq \zeta\frac{2}{1-\zeta}\sqrt{\frac{\mu^2 r}{n_1}}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\left(1-\frac{2\zeta}{1-\zeta}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right)^{-1}\left(1+\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\right),
\end{aligned}
$$

where we applied (58), (59) in the second inequality, and Lemma 23 in the third inequality.

Similarly, we define $\delta\mathbf{H}_{(i),l}$ as the summation,

$$
\begin{aligned}
\delta\mathbf{H}_{(i),l}=&\sum_{p=1}^{\infty}\mathbf{F}_{(i)}^p\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-p-1}+\sum_{p=1}^{\infty}\mathbf{F}_{(i)}^p\hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1}\\
&+\sum_{p=0}^{\infty}\mathbf{F}_{(i)}^p\delta\mathbf{H}_g\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1}.
\end{aligned}
$$

We first calculate the ℓ_2 norm of $\delta\mathbf{H}_{(i),l}$ as,

$$
\begin{aligned}
\left\|\delta\mathbf{H}_{(i),l}\right\|\leq&\sum_{p=1}^{\infty}\left\|\mathbf{F}_{(i)}^p\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-p-1}\right\|+\sum_{p=1}^{\infty}\left\|\mathbf{F}_{(i)}^p\hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1}\right\|\\
&+\sum_{p=0}^{\infty}\left\|\mathbf{F}_{(i)}^p\delta\mathbf{H}_g\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-p-1}\right\|\\
&\leq\sum_{p=1}^{\infty}\zeta^p\sigma_{\max}^2\frac{1}{\frac{3}{4}\sigma_{\min}^2}+\sum_{p=1}^{\infty}\zeta^p\sigma_{\max}^2\frac{1}{\frac{3}{4}\sigma_{\min}^2}\frac{1}{2}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\\
&+\|\delta\mathbf{H}_g\|\sum_{p=0}^{\infty}\zeta^p\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\\
&\leq\zeta\frac{1}{1-\zeta}\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(2+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}+3\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2+4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^3+4\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^4\right),
\end{aligned}
$$

where we applied the upper bound on $\|\delta\mathbf{H}\|$ in the last inequality.

Finally, we have,

$$\begin{aligned}
&\max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_{(i),l}\right\| \\
&\leq \sum_{p_0=1}^{\infty} \max_j \left\|\mathbf{e}_j^T \mathbf{F}_{(0)}^{p_0} \mathbf{T}_{(i)}\right\| \left\|\hat{\mathbf{H}}_{(i),l} \mathbf{\Lambda}_{2,(i)}^{-p_0-1}\right\| + \sum_{p=1}^{\infty} \max_j \left\|\mathbf{e}_j^T \mathbf{F}_{(i)}^{p} \hat{\mathbf{H}}_{g,0} \mathbf{\Lambda}_{4,(i)} \mathbf{\Lambda}_{2,(i)}^{-p-1}\right\| \\
&+ \sum_{p=0}^{\infty} \max_j \left\|\mathbf{e}_j^T \mathbf{F}_{(i)}^{p} \delta\mathbf{H}_g\right\| \left\|\mathbf{\Lambda}_{4,(i)} \mathbf{\Lambda}_{2,(i)}^{-p-1}\right\|.
\end{aligned}$$

The first two summations can be upper bounded by Lemma 22, and the last summation can be estimated in a similar way we calculate $\max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_g\right\|$. We omit the details and present the estimated upper bound for brevity.

This completes our proof. ■

Equipped with the aforementioned perturbation analysis on $\hat{\mathbf{H}}_g$ and $\mathbf{H}_{(i),l}$, we are ready to provide the formal version of Lemma 6.

Lemma 20 *Under the same conditions as Lemma 19, we have:*

$$\left\|\mathbf{L^\star}_{(i)} - \hat{\mathbf{L}}_{(i)}\right\|_\infty \leq \sqrt{\alpha}\mu^2 r \max_j \left\|\mathbf{E}_{(j)}\right\|_\infty C_4,$$

where C_4 is a constant satisfying,

$$C_4 = \mathcal{O}\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{10} \frac{1}{\sqrt{\theta}} + \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{18}\right). \tag{60}$$

Proof Notice that $\mathbf{L^\star}_{(i)} = \mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l}$, and $\hat{\mathbf{L}}_{(i)} = (\mathbf{P}_{\hat{\mathbf{H}}_g} + \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}})\hat{\mathbf{M}}_{(i)} = (\mathbf{P}_{\hat{\mathbf{H}}_g} + \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}})(\mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l} + \mathbf{E}_{(i),t})$. Therefore, we have

$$\begin{aligned}
&\left\|\mathbf{L^\star}_{(i)} - \hat{\mathbf{L}}_{(i)}\right\|_\infty \\
&\leq \left\|(\mathbf{P}_{\hat{\mathbf{H}}_g} + \mathbf{P}_{\hat{\mathbf{H}}_{(i),l}})(\mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l} + \mathbf{E}_{(i),t}) - \mathbf{L^\star}_{(i),g} - \mathbf{L^\star}_{(i),l}\right\|_\infty \\
&\leq \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}(\mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l} + \mathbf{E}_{(i),t}) - \mathbf{L^\star}_{(i),g}\right\|_\infty \\
&+ \left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}(\mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l} + \mathbf{E}_{(i),t}) - \mathbf{L^\star}_{(i),l}\right\|_\infty \\
&\leq \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{L^\star}_{(i),l}\right\|_\infty + \left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{L^\star}_{(i),g}\right\|_\infty \\
&+ \left\|\left(\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^T + \hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_g^T + \delta\mathbf{H}_g\hat{\mathbf{H}}_{g,0}^T + \delta\mathbf{H}_g\delta\mathbf{H}_g^T\right)(\mathbf{L^\star}_{(i),g} + \mathbf{E}_{(i),t}) - \mathbf{L^\star}_{(i),g}\right\|_\infty \\
&+ \left\|\left(\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T + \hat{\mathbf{H}}_{(i),l,0}\delta\mathbf{H}_{(i),l}^T + \delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}_{(i),l,0}^T + \delta\mathbf{H}_{(i),l}\delta\mathbf{H}_{(i),l}^T\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\left(\mathbf{L}^{\star}{}_{(i),l}+\mathbf{E}_{(i),t}\right)-\mathbf{L}^{\star}{}_{(i),l}\Big|\Big|_{\infty}\\
&\leq\left\|\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}+\left\|\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^{T}\mathbf{L}^{\star}{}_{(i),l}-\mathbf{L}^{\star}{}_{(i),l}\right\|_{\infty}\\
&+\left\|\hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_{g}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}+\left\|\delta\mathbf{H}_{g}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}+\left\|\delta\mathbf{H}_{g}\delta\mathbf{H}_{g}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}\\
&+\left\|\hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_{g}^{T}\mathbf{E}_{(i),t}\right\|_{\infty}+\left\|\delta\mathbf{H}_{g}\hat{\mathbf{H}}_{g,0}\mathbf{E}_{(i),t}\right\|_{\infty}+\left\|\delta\mathbf{H}_{g}\delta\mathbf{H}_{g}^{T}\mathbf{E}_{(i),t}\right\|_{\infty}\\
&+\left\|\hat{\mathbf{H}}_{(i),l,0}\delta\mathbf{H}_{(i),l}^{T}\mathbf{L}^{\star}{}_{(i),l}\right\|_{\infty}+\left\|\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}_{(i),l,0}\mathbf{L}^{\star}{}_{(i),l}\right\|_{\infty}+\left\|\delta\mathbf{H}_{(i),l}\delta\mathbf{H}_{(i),l}^{T}\mathbf{L}^{\star}{}_{(i),l}\right\|_{\infty}\\
&+\left\|\hat{\mathbf{H}}_{(i),l,0}\delta\mathbf{H}_{(i),l}^{T}\mathbf{E}_{(i),t}\right\|_{\infty}+\left\|\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}_{(i),l,0}^{T}\mathbf{E}_{(i),t}\right\|_{\infty}+\left\|\delta\mathbf{H}_{(i),l}\delta\mathbf{H}_{(i),l}^{T}\mathbf{E}_{(i),t}\right\|_{\infty}\\
&+\left\|\mathbf{P}_{\hat{\mathbf{H}}_{g}}\mathbf{L}^{\star}{}_{(i),l}\right\|_{\infty}+\left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}.
\end{aligned}
\tag{61}
$$

There are 16 terms in (61), we will bound each of them respectively.

Bounding the first term of (61)*:*

$$
\begin{aligned}
&\left\|\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}\\
&\leq\left\|\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}+\left\|\left(\mathbf{I}-\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}\\
&\leq\frac{\mu^{2}r}{\overline{n}}\left\|\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|+\left\|\left(\mathbf{I}-\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}\\
&\leq\frac{\mu^{2}r}{\overline{n}}\left\|\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|+\frac{\mu^{2}r}{\overline{n}}\left\|\left(\mathbf{I}-\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|\\
&+\left\|\left(\mathbf{I}-\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}\right\|_{\infty}.
\end{aligned}
\tag{TM1}
$$

Recall the definition of $\hat{\mathbf{H}}_{g,0}$ as,

$$
\begin{aligned}
\hat{\mathbf{H}}_{g,0}&=\mathbf{T}_{(0)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{6}^{-1}+\sum_{i=1}^{N}\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_{6}^{-1}\\
&=\hat{\mathbf{H}}_{g}\underbrace{-\mathbf{F}_{(0)}\hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{6}^{-1}-\sum_{i=1}^{N}\mathbf{F}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_{6}^{-1}}_{\delta\mathbf{H}_{g,0}},
\end{aligned}
$$

where we used the KKT condition (13a) and the definition of $\mathbf{\Lambda}_{6}=\mathbf{\Lambda}_{1}+\sum_{i=1}^{N}\mathbf{\Lambda}_{3,(i)}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}$.

The first term in (TM1) is thus bounded by,

$$
\begin{aligned}
&\frac{\mu^{2}r}{\overline{n}}\left\|\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{L}^{\star}{}_{(i),g}\right\|\\
&\leq\frac{\mu^{2}r}{\overline{n}}\left\|\hat{\mathbf{H}}_{g}\hat{\mathbf{H}}_{g}^{T}\mathbf{L}^{\star}{}_{(i),g}-\mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\mathbf{L}^{\star}{}_{(i),g}\right\|
\end{aligned}
$$

$$+\frac{\mu^2 r}{\overline{n}}\left\|\hat{\mathbf{H}}_g\delta\mathbf{H}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|+\frac{\mu^2 r}{\overline{n}}\left\|\delta\mathbf{H}_{g,0}\hat{\mathbf{H}}_g^T\mathbf{L}^\star{}_{(i),g}\right\|+\frac{\mu^2 r}{\overline{n}}\left\|\delta\mathbf{H}_{g,0}\delta\mathbf{H}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|.$$

The second term in (TM1) is bounded by

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|\\
&\leq\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\hat{\mathbf{H}}_g\hat{\mathbf{H}}_g^T\mathbf{L}^\star{}_{(i),g}\right\|+\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\delta\mathbf{H}_{g,0}\hat{\mathbf{H}}_g^T\mathbf{L}^\star{}_{(i),g}\right\|\\
&+\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\hat{\mathbf{H}}_g\delta\mathbf{H}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|+\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\delta\mathbf{H}_{g,0}\delta\mathbf{H}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|\\
&\leq\left\|\hat{\mathbf{H}}_g\hat{\mathbf{H}}_g^T-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right\|\sigma_{\max}+2\left\|\delta\mathbf{H}_{g,0}\right\|\sigma_{\max}+\left\|\delta\mathbf{H}_{g,0}\right\|^2\sigma_{\max}.
\end{aligned}$$

The third term in (TM1) is bounded by

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|_\infty\\
&=\left\|\sum_{i=1}^N\left(\hat{\mathbf{H}}^\star{}_{(i)}\hat{\mathbf{H}}^\star{}_{(i)}^T\frac{\mathbf{T}_{(i)}}{N}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\hat{\mathbf{H}}^\star{}_{(i)}\hat{\mathbf{H}}^\star{}_{(i)}^T\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{L}^\star{}_{(i),g}\right\|_\infty\\
&\leq\frac{\mu^2 r}{\overline{n}}\sum_{i=1}^N\left\|\left(\hat{\mathbf{H}}^\star{}_{(i)}\hat{\mathbf{H}}^\star{}_{(i)}^T\frac{\mathbf{T}_{(i)}}{N}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\hat{\mathbf{H}}^\star{}_{(i)}\hat{\mathbf{H}}^\star{}_{(i)}^T\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{L}^\star{}_{(i),g}\right\|.
\end{aligned}\tag{62}$$

We know that,

$$\begin{aligned}
\mathbf{H}^\star{}_{(i),l}\mathbf{H}^\star{}_{(i),l}^T\mathbf{T}_{(i)}&=\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{T}_{(i)}+\left(\mathbf{H}^\star{}_{(i),l}\mathbf{H}^\star{}_{(i),l}^T-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)}\\
&=\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{S}_{(i)}-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{F}_{(i)}+\left(\mathbf{H}^\star{}_{(i),l}\mathbf{H}^\star{}_{(i),l}^T-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)},
\end{aligned}$$

and that,

$$\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\frac{\mathbf{S}_{(i)}}{N}\hat{\mathbf{H}}_g+\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{S}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}=0.$$

As a result, we have

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^\star{}_g^T\right)\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^T\mathbf{L}^\star{}_{(i),g}\right\|_\infty\\
&\leq\frac{\mu^2 r}{\overline{n}}\sum_{i=1}^N\left\|\hat{\mathbf{H}}^\star{}_{(i)}\hat{\mathbf{H}}^\star{}_{(i)}^T\frac{-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{F}_{(i)}+\left(\mathbf{H}^\star{}_{(i),l}\mathbf{H}^\star{}_{(i),l}^T-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right)\mathbf{T}_{(i)}}{N}\right.\\
&\left.\times\left(\hat{\mathbf{H}}_g+\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}N\mathbf{\Lambda}_{5,(i)}\right)\mathbf{\Lambda}_6^{-1}\mathbf{L}^\star{}_{(i),g}\right\|\\
&\leq\frac{\mu^2 r}{\overline{n}}\sigma_{\max}\frac{1}{N}\sum_{i=1}^N\frac{\left(\left\|\mathbf{F}_{(i)}\right\|+\sigma_{\max}^2\left\|\mathbf{H}^\star{}_{(i),l}\mathbf{H}^\star{}_{(i),l}^T-\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\right\|\right)}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right).
\end{aligned}$$

Combing them all, we have,

$$\begin{aligned}
&\left\|\hat{\mathbf{H}}_{g,0}\hat{\mathbf{H}}_{g,0}^T\mathbf{L}^{\star}{}_{(i),g} - \mathbf{L}^{\star}{}_{(i),g}\right\|_\infty \\
&\leq \frac{\mu^2 r}{\overline{n}}\sigma_{\max}\left(2\left\|\Delta\mathbf{P}_g\right\| + 6\frac{\left\|\mathbf{F}_{(0)}\right\|}{\frac{3}{4}\sigma_{\min}^2}\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right) + \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\mathbf{F}_{(i)}\right\| + \sigma_{\max}^2\left\|\Delta\mathbf{P}_{(i),l}\right\|}{\frac{3}{4}\sigma_{\min}^2}\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right).
\end{aligned} \tag{63}$$

Bounding the second term of (61)*:*

$$\begin{aligned}
&\left\|\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l} - \mathbf{L}^{\star}{}_{(i),l}\right\|_\infty \\
&\leq \left\|\mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l} - \mathbf{L}^{\star}{}_{(i),l}\right\|_\infty \\
&+ \left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\|_\infty \\
&\leq \frac{\mu^2 r}{\overline{n}}\left\|\mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l} - \mathbf{L}^{\star}{}_{(i),l}\right\| \\
&+ \left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\|_\infty \\
&\leq \frac{\mu^2 r}{\overline{n}}\left\|\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l} - \mathbf{L}^{\star}{}_{(i),l}\right\| + \frac{\mu^2 r}{\overline{n}}\left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\|_\infty \\
&+ \left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\|_\infty.
\end{aligned} \tag{TM2}$$

We will estimate upper bounds of three terms in (TM2) respectively.
Recall that the definition of $\hat{\mathbf{H}}_{(i),l,0}$ as,

$$\begin{aligned}
\hat{\mathbf{H}}_{(i),l,0} &= \mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1} + \hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1} \\
&= \mathbf{S}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1} + \hat{\mathbf{H}}_{g}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1} - \mathbf{F}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1} + \delta\mathbf{H}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}_{2,(i)}^{-1} \\
&= \hat{\mathbf{H}}_{(i),l} - \delta\mathbf{H}_{(i),l}.
\end{aligned}$$

The first term in (TM2) is thus bounded by,

$$\begin{aligned}
&\frac{\mu^2 r}{\overline{n}}\left\|\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l} - \mathbf{L}^{\star}{}_{(i),l}\right\| \\
&\leq \frac{\mu^2 r}{\overline{n}}\left\|\hat{\mathbf{H}}_{(i),l}\hat{\mathbf{H}}_{(i),l}\mathbf{L}^{\star}{}_{(i),g} - \mathbf{H}^{\star}{}_{g}\mathbf{H}^{\star T}_{g}\mathbf{L}^{\star}{}_{(i),g}\right\| \\
&+ \frac{\mu^2 r}{\overline{n}}\left\|\hat{\mathbf{H}}_{(i),l}\delta\mathbf{H}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\| + \frac{\mu^2 r}{\overline{n}}\left\|\delta\mathbf{H}_{(i),l,0}\hat{\mathbf{H}}_{(i),l}^T\mathbf{L}^{\star}{}_{(i),l}\right\| + \frac{\mu^2 r}{\overline{n}}\left\|\delta\mathbf{H}_{(i),l,0}\delta\mathbf{H}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\|.
\end{aligned}$$

The second term in (TM2) is upper bounded by

$$\begin{aligned}
&\left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0}\hat{\mathbf{H}}_{(i),l,0}^T\mathbf{L}^{\star}{}_{(i),l}\right\| \\
&\leq \left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{L}^{\star}{}_{(i),l}\right\| + \left\|\left(\mathbf{I} - \mathbf{H}^{\star}{}_{(i),l}\mathbf{H}^{\star T}_{(i),l}\right)\delta\mathbf{H}_{(i),l,0}\hat{\mathbf{H}}_{(i),l}^T\mathbf{L}^{\star}{}_{(i),l}\right\|
\end{aligned}$$

$$
\begin{aligned}
&+\left\|\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right)\hat{\mathbf{H}}_{(i),l}\delta\mathbf{H}^T_{(i),l,0}\mathbf{L^\star}_{(i),l}\right\|+\left\|\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right)\delta\mathbf{H}_{(i),l,0}\delta\mathbf{H}^T_{(i),l,0}\mathbf{L^\star}_{(i),l}\right\| \\
&\leq\left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right\|\sigma_{\max}+2\left\|\delta\mathbf{H}_{(i),l,0}\right\|\sigma_{\max}+\left\|\delta\mathbf{H}_{(i),l,0}\right\|^2\sigma_{\max}.
\end{aligned}
$$

Then we bound the third term of (TM2). From the definition of $\hat{\mathbf{H}}_{(i),l,0}$ and $\mathbf{T}_{(i)}$, we know,

$$
\begin{aligned}
&\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right)\hat{\mathbf{H}}_{(i),l,0} \\
&=\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)}+\mathbf{H^\star}_g\mathbf{H^\star}_g^T\hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&+\left(\mathbf{I}-\mathbf{H^\star}_g\mathbf{H^\star}_g^T-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right)\hat{\mathbf{H}}_{g,0}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&=\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&+\mathbf{H^\star}_g\mathbf{H^\star}_g^T\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{j=1}^N\mathbf{T}_{(j)}\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)}+\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right) \\
&\times\left(\sum_{j=1}^N\mathbf{H^\star}_{(j),l}\mathbf{H^\star}^T_{(j),l}\frac{\mathbf{T}_{(j)}}{N}\hat{\mathbf{H}}_g+\sum_{j=1}^N\mathbf{H^\star}_{(j),l}\mathbf{H^\star}^T_{(j),l}\mathbf{T}_{(j)}\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\right)\mathbf{\Lambda}_6^{-1}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&=\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{S}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&+\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{P}_{\hat{\mathbf{H}}_g}\left(\mathbf{S}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{j=1}^N\mathbf{S}_{(j)}\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&-\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{F}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)}-\mathbf{H^\star}_g\mathbf{H^\star}_g^T\Delta\mathbf{P}_g\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&-\mathbf{H^\star}_g\mathbf{H^\star}_g^T\mathbf{P}_{\hat{\mathbf{H}}_g}\left(\mathbf{F}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{j=1}^N\mathbf{F}_{(j)}\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&-\mathbf{H^\star}_g\mathbf{H^\star}_g^T\Delta\mathbf{P}_g\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{j=1}^N\mathbf{T}_{(j)}\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&+\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right) \\
&\times\left(\sum_{j=1}^N\mathbf{H^\star}_{(j),l}\mathbf{H^\star}^T_{(j),l}\hat{\mathbf{H}}_{(j),l}\hat{\mathbf{H}}^T_{(j),l}\mathbf{S}_{(j)}\left(\frac{\hat{\mathbf{H}}_g}{N}+\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\right)\right)\mathbf{\Lambda}_6^{-1}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&-\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right) \\
&\times\left(\sum_{j=1}^N\mathbf{H^\star}_{(j),l}\mathbf{H^\star}^T_{(j),l}\hat{\mathbf{H}}_{(j),l}\hat{\mathbf{H}}^T_{(j),l}\mathbf{F}_{(j)}\left(\frac{\hat{\mathbf{H}}_g}{N}+\hat{\mathbf{H}}_{(j),l}\mathbf{\Lambda}^{-1}_{2,(j)}\mathbf{\Lambda}_{5,(j)}\right)\right)\mathbf{\Lambda}_6^{-1}\mathbf{\Lambda}_{4,(i)}\mathbf{\Lambda}^{-1}_{2,(i)} \\
&-\left(\mathbf{I}-\mathbf{H^\star}_{(i),l}\mathbf{H^\star}^T_{(i),l}\right)
\end{aligned}
$$

$$\times \left(\sum_{j=1}^{N} \mathbf{H^\star}_{(j),l} \mathbf{H^\star}^T_{(j),l} \Delta \mathbf{P}_{(j),l} \mathbf{T}_{(j)} \left(\frac{\hat{\mathbf{H}}_g}{N} + \hat{\mathbf{H}}_{(j),l} \mathbf{\Lambda}^{-1}_{2,(j)} \mathbf{\Lambda}_{5,(j)} \right) \right) \mathbf{\Lambda}_6^{-1} \mathbf{\Lambda}_{4,(i)} \mathbf{\Lambda}^{-1}_{2,(i)}$$

.

From the KKT conditions, we know $\hat{\mathbf{H}}_g^T \mathbf{S}_i \hat{\mathbf{H}}_{(i),l} + \hat{\mathbf{H}}_g^T \left(\mathbf{S}_{(0)} \hat{\mathbf{H}}_g \mathbf{\Lambda}_6^{-1} + \sum_{j=1}^{N} \mathbf{S}_{(j)} \hat{\mathbf{H}}_{(j),l} \mathbf{\Lambda}^{-1}_{2,(j)} \mathbf{\Lambda}_{5,(j)} \mathbf{\Lambda}_6^{-1} \right) \mathbf{\Lambda}_{4,(i)} = 0$ and $\hat{\mathbf{H}}^T_{(j),l} \mathbf{S}_{(j)} \left(\frac{\hat{\mathbf{H}}_g}{N} + \hat{\mathbf{H}}_{(j),l} \mathbf{\Lambda}^{-1}_{2,(j)} \mathbf{\Lambda}_{5,(j)} \right) = 0$.

Therefore, we have,

$$\begin{aligned}
&\max_k \left| \mathbf{e}_k^T \left(\mathbf{I} - \mathbf{H^\star}_{(i),l} \mathbf{H^\star}^T_{(i),l} \right) \hat{\mathbf{H}}_{(i),l,0} \right| \\
&\leq \sqrt{\frac{\mu^2 r}{n_1}} \Big(\frac{\left\| \mathbf{F}_{(i)} \right\|}{\frac{3}{4}\sigma^2_{\min}} + \frac{\left\| \mathbf{F}_{(0)} \right\|}{\frac{3}{4}\sigma^2_{\min}} \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} + \left\| \Delta \mathbf{P}_g \right\| \frac{\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(1 + \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} + \left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right)^2 \right) \\
&+ \frac{1}{N} \sum_{j=1}^{N} \frac{\left\| \mathbf{F}_{(j)} \right\|}{\frac{3}{4}\sigma^2_{\min}} \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(2 + 3\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right) \Big) + \frac{1}{N} \sum_{j=1}^{N} \left\| \Delta \mathbf{P}_{(j)} \right\| \frac{\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} 2 \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(1 + \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right) \Big).
\end{aligned} \tag{64}$$

Combing these results, we have,

$$\begin{aligned}
&\left\| \hat{\mathbf{H}}_{(i),l,0} \hat{\mathbf{H}}^T_{(i),l,0} \mathbf{L^\star}_{(i),l} - \mathbf{L^\star}_{(i),l} \right\|_\infty \\
&\leq \frac{\mu^2 r}{\overline{n}} \sigma_{\max} \times \Big(\frac{\left\| \mathbf{F}_{(i)} \right\|}{\frac{3}{4}\sigma^2_{\min}} + \frac{\left\| \mathbf{F}_{(0)} \right\|}{\frac{3}{4}\sigma^2_{\min}} \left(6 + 7\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right) \\
&+ \left\| \Delta \mathbf{P}_g \right\| \frac{\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(1 + \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} + \left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right)^2 \right) + 2 \left\| \Delta \mathbf{P}_{(i),l} \right\| \\
&+ \frac{1}{N} \sum_{j=1}^{N} \frac{\left\| \mathbf{F}_{(j)} \right\|}{\frac{3}{4}\sigma^2_{\min}} \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(2 + 3\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right) \Big) + \frac{1}{N} \sum_{j=1}^{N} \left\| \Delta \mathbf{P}_{(j)} \right\| \frac{\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} 2 \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \left(1 + \frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right) \Big).
\end{aligned} \tag{65}$$

Bounding the third term of (61)*:*

$$\begin{aligned}
&\left\| \hat{\mathbf{H}}_{g,0} \delta \mathbf{H}_g^T \mathbf{L^\star}_{(i),g} \right\|_\infty \\
&= \max_{j,k} \left| \mathbf{e}_j^T \left(\mathbf{T}_{(0)} \hat{\mathbf{H}}_g \mathbf{\Lambda}_6^{-1} + \sum_{i=1}^{N} \mathbf{T}_{(i)} \hat{\mathbf{H}}_{(i),l} \mathbf{\Lambda}^{-1}_{2,(i)} \mathbf{\Lambda}_{5,(i)} \mathbf{\Lambda}_6^{-1} \right) \delta \mathbf{H}_g \mathbf{L^\star}_{(i),g} \mathbf{e}_k \right| \\
&\leq \frac{\mu^2 r}{\overline{n}} \sigma_{\max} \left\| \delta \mathbf{H}_g \right\| \left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} + \left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}} \right)^2 \right).
\end{aligned}$$

(66)

Bounding the fourth term of (61)*:*

$$\begin{aligned}
&\left\|\hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_g^T\mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k}\left|\sum_l \mathbf{e}_j^T\hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_g^T\mathbf{e}_l\mathbf{e}_l^T\mathbf{E}_{(i),t}\mathbf{e}_k\right| \\
&\le \max_{j,l}\left|\mathbf{e}_j^T\hat{\mathbf{H}}_{g,0}\delta\mathbf{H}_g^T\mathbf{e}_l\right|\alpha n_1\left\|\mathbf{E}_{(i)}\right\|_\infty \\
&= \max_{j,l}\left|\mathbf{e}_j^T\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{i=1}^N\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right)\delta\mathbf{H}_g^T\mathbf{e}_l\right|\alpha n_1\left\|\mathbf{E}_{(i)}\right\|_\infty \\
&\le \frac{\mu^2 r}{n_1}\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\alpha n_1\left\|\mathbf{E}_{(i)}\right\|_\infty,
\end{aligned} \tag{67}$$

where we used the definition of α-sparsity in the second inequality, and applied Lemma 19 in the last inequality.

Bounding the fifth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_g\hat{\mathbf{H}}_{g,0}^T\mathbf{L^\star}_{(i),g}\right\|_\infty \\
&= \max_{j,k}\left|\mathbf{e}_j^T\delta\mathbf{H}_g\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{i=1}^N\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right)\mathbf{L^\star}_{(i),g}\mathbf{e}_k\right| \\
&\le \max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_g\right\|\sqrt{\frac{\mu^2 r}{n_2}}\left\|\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{i=1}^N\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right\| \\
&\le \sigma_{\max}\max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_g\right\|\sqrt{\frac{\mu^2 r}{n_2}}\frac{\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right).
\end{aligned}$$

Bounding the sixth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_g\hat{\mathbf{H}}_{g,0}^T\mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k}\left|\sum_l\mathbf{e}_j^T\delta\mathbf{H}_g\left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_6^{-1}+\sum_{i=1}^N\mathbf{T}_{(i)}\hat{\mathbf{H}}_{(i),l}\mathbf{\Lambda}_{2,(i)}^{-1}\mathbf{\Lambda}_{5,(i)}\mathbf{\Lambda}_6^{-1}\right)^T\mathbf{E}_{(i),t}\mathbf{e}_k\right| \\
&\le \max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_g\right\|\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\sqrt{\frac{\mu^2 r}{n_1}}\alpha n_1\left\|\mathbf{E}\right\|_\infty,
\end{aligned}$$

where we applied the incoherence condition on $\mathbf{T}_{(0)}$ and $\mathbf{T}_{(i)}$, and the relation $\sum_l \left|\mathbf{e}_l^T \mathbf{E}_{(i),t} \mathbf{e}_k\right| \leq \alpha n_1 \left\|\mathbf{E}\right\|_\infty$.
Bounding the seventh term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_g \delta\mathbf{H}_g^T \mathbf{L}^\star{}_{(i),g}\right\|_\infty \\
&= \max_{j,k} \left|\mathbf{e}_j^T \delta\mathbf{H}_g \delta\mathbf{H}_g^T \mathbf{L}^\star{}_{(i),g} \mathbf{e}_k\right| \\
&\leq \max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_g\right\| \sqrt{\frac{\mu^2 r}{n_2}} \sigma_{\max},
\end{aligned} \tag{68}$$

where we applied the incoherence on $\mathbf{L}^\star{}_{(i),g}$ and $\|\delta\mathbf{H}_g\| \leq 1$ in the first inequality.
Bounding the eighth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_g \delta\mathbf{H}_g^T \mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k} \left|\sum_l \mathbf{e}_j^T \delta\mathbf{H}_g \delta\mathbf{H}_g^T \mathbf{e}_l \mathbf{e}_l^T \mathbf{E}_{(i),t} \mathbf{e}_k\right| \\
&\leq \max_{j,l} \left\|\mathbf{e}_j^T \delta\mathbf{H}_g\right\| \left\|\mathbf{e}_l^T \delta\mathbf{H}_g\right\| \alpha n_1 \left\|\mathbf{E}\right\|_\infty \\
&= \left(\max_j \left\|\mathbf{e}_j^T \delta\mathbf{H}_g\right\|\right)^2 \alpha n_1 \left\|\mathbf{E}\right\|_\infty,
\end{aligned}$$

where we applied $\sum_l \left|\mathbf{e}_l^T \mathbf{E}_{(i),t} \mathbf{e}_k\right| \leq \alpha n_1 \left\|\mathbf{E}\right\|_\infty$ in the first inequality.
Bounding the ninth term of (61)*:*

$$\begin{aligned}
&\left\|\hat{\mathbf{H}}_{(i),l,0} \delta\mathbf{H}_{(i),l}^T \mathbf{L}^\star{}_{(i),l}\right\|_\infty \\
&= \max_{j,k} \left|\mathbf{e}_j^T \left(\mathbf{T}_{(i)} \mathbf{H}_{(i),l} \mathbf{\Lambda}_{2,(i)}^{-1} + \hat{\mathbf{H}}_{g,0} \mathbf{\Lambda}_{4,(i)} \mathbf{\Lambda}_{2,(i)}^{-1}\right) \delta\mathbf{H}_{(i),l}^T \mathbf{L}^\star{}_{(i),l} \mathbf{e}_k\right| \\
&= \frac{\mu^2 r}{\overline{n}} \sigma_{\max} \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \left(1 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} + \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right) \left\|\delta\mathbf{H}_{(i),l}\right\|,
\end{aligned}$$

where we applied the incoherence condition of $\mathbf{T}_{(0)}$ and $\mathbf{T}_{(i)}$.
Bounding the tenth term of (61)*:*

$$\begin{aligned}
&\left\|\hat{\mathbf{H}}_{(i),l,0} \delta\mathbf{H}_{(i),l}^T \mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k} \left|\sum_l \mathbf{e}_j^T \hat{\mathbf{H}}_{(i),l,0} \delta\mathbf{H}_{(i),l}^T \mathbf{e}_l \mathbf{e}_l^T \mathbf{E}_{(i),t} \mathbf{e}_k\right| \\
&= \max_{j,l} \left|\mathbf{e}_j^T \hat{\mathbf{H}}_{(i),l,0} \delta\mathbf{H}_{(i),l}^T \mathbf{e}_l\right| \alpha n_1 \left\|\mathbf{E}\right\|_\infty \\
&\leq \sqrt{\frac{\mu^2 r}{n_1}} \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} \left(1 + \frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2} + \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right) \left\|\mathbf{e}_l^T \delta\mathbf{H}_{(i),l}\right\| \alpha n_1 \left\|\mathbf{E}\right\|_\infty,
\end{aligned}$$

where we applied the incoherence condition in the first inequality, (57) in the second inequality.

Bounding the eleventh term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}^T_{(i),l,0}\mathbf{L}^\star{}_{(i),l}\right\|_\infty \\
&= \max_{j,k}\left|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}^T_{(i),l,0}\mathbf{L}^\star{}_{(i),l}\mathbf{e}_k\right| \\
&\le \max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\frac{\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}\left(1+\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}+\left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}\right)^2\right)\sigma_{\max}\sqrt{\frac{\mu^2 r}{n_2}},
\end{aligned}$$

where we applied the incoherence condition in the first inequality, and (57) in the second inequality.

Bounding the twelfth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}^T_{(i),l,0}\mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k}\left|\sum_l \mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}^T_{(i),l,0}\mathbf{e}_l\mathbf{e}_l^T\mathbf{E}_{(i),l}\mathbf{e}_k\right| \\
&\le \max_{j,l}\left|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\hat{\mathbf{H}}^T_{(i),l,0}\mathbf{e}_l\right|\alpha n_1\left\|\mathbf{E}\right\|_\infty \\
&\le \max_j\left|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right|\sqrt{\frac{\mu^2 r}{n_1}}\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}\left(1+\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}+\left(\frac{2\sigma^2_{\max}}{\frac{3}{4}\sigma^2_{\min}}\right)^2\right)\alpha n_1\left\|\mathbf{E}\right\|_\infty,
\end{aligned} \tag{69}$$

where we applied the condition $\sum_l\left|\mathbf{e}_l^T\mathbf{E}_{(i),t}\mathbf{e}_k\right| \le \alpha n_1\left\|\mathbf{E}\right\|_\infty$ in the first inequality, the incoherence condition in the second inequality.

Bounding the thirteenth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_{(i),l}\delta\mathbf{H}^T_{(i),l}\mathbf{L}^\star{}_{(i),l}\right\|_\infty \\
&= \max_{j,k}\left|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\delta\mathbf{H}^T_{(i),l}\mathbf{L}^\star{}_{(i),l}\mathbf{e}_k\right| \\
&\le \max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\left\|\delta\mathbf{H}_{(i),l}\right\|\sigma_{\max}\sqrt{\frac{\mu^2 r}{n_2}} \\
&\le \max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\sigma_{\max}\sqrt{\frac{\mu^2 r}{n_2}},
\end{aligned} \tag{70}$$

where we apply the incoherence condition in the first inequality, and $\left\|\delta\mathbf{H}_{(i),l}\right\| \le 1$ in the second inequality.

Bounding the fourteenth term of (61)*:*

$$\begin{aligned}
&\left\|\delta\mathbf{H}_{(i),l}\delta\mathbf{H}_{(i),l}^T\mathbf{E}_{(i),t}\right\|_\infty \\
&= \max_{j,k}\left|\sum_m \mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\delta\mathbf{H}_{(i),l}^T\mathbf{e}_m\mathbf{e}_m^T\mathbf{E}_{(i),t}\mathbf{e}_k\right| \\
&\le \max_{j.m}\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\left\|\mathbf{e}_m^T\delta\mathbf{H}_{(i),l}\right\|\alpha n_1\left\|\mathbf{E}\right\|_\infty \\
&\le \left(\max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\right)^2\alpha n_1\left\|\mathbf{E}\right\|_\infty,
\end{aligned}$$

where we applied the condition $\sum_l\left|\mathbf{e}_l^T\mathbf{E}_{(i),t}\mathbf{e}_k\right|\le\alpha n_1\left\|\mathbf{E}\right\|_\infty$ in the first inequality.
Bounding the fifteenth term of (61)*:*

$$\begin{aligned}
&\left\|\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{L}^\star{}_{(i),l}\right\|_\infty \\
&= \max_{j,k}\left|\mathbf{e}_j^T\left(\hat{\mathbf{H}}_{g,0}+\delta\mathbf{H}_g\right)\hat{\mathbf{H}}_g^T\mathbf{H}^\star{}_{(i),l}\mathbf{\Sigma}_{(i),l}\mathbf{W}^{\star T}_{(i),l}\mathbf{e}_k\right| \\
&\le \left(\max_j\left|\mathbf{e}_j^T\hat{\mathbf{H}}_{g,0}\right|+\max_j\left|\mathbf{e}_j^T\delta\mathbf{H}_g\right|\right)\left\|\hat{\mathbf{H}}_g^T\mathbf{H}^\star{}_{(i),l}\right\|\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}} \\
&\le \left(\max_j\left|\mathbf{e}_j^T\hat{\mathbf{H}}_{g,0}\right|+\max_j\left|\mathbf{e}_j^T\delta\mathbf{H}_g\right|\right)\left\|\Delta\mathbf{P}_g\right\|\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}} \\
&= \frac{\mu^2r}{\overline{n}}\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)\left\|\Delta\mathbf{P}_g\right\|\sigma_{\max}+\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}}\left\|\Delta\mathbf{P}_g\right\|\max_j\left|\mathbf{e}_j^T\delta\mathbf{H}_g\right|.
\end{aligned}$$

The second inequality comes from the relation $\left\|\hat{\mathbf{H}}_g^T\mathbf{H}^\star{}_{(i),l}\right\| = \left\|\hat{\mathbf{H}}_g^T\left(\mathbf{I}-\mathbf{H}^\star{}_g\mathbf{H}^{\star T}_g\right)\mathbf{H}^\star{}_{(i),l}\right\| = \left\|\hat{\mathbf{H}}_g^T\left(\mathbf{P}_{\hat{\mathbf{H}}_g}-\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{H}^\star{}_g\mathbf{H}^{\star T}_g\right)\mathbf{H}^\star{}_{(i),l}\right\| \le \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}-\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{H}^\star{}_g\mathbf{H}^{\star T}_g\right\| = \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}(\mathbf{P}_{\hat{\mathbf{H}}_g}-\mathbf{H}^\star{}_g\mathbf{H}^{\star T}_g)\right\| \le \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}-\mathbf{H}^\star{}_g\mathbf{H}^{\star T}_g\right\|$.
Bounding the sixteenth term of (61)*:*

$$\begin{aligned}
&\left\|\mathbf{P}_{\hat{\mathbf{H}}_{(i),l}}\mathbf{L}^\star{}_{(i),g}\right\|_\infty = \max_{j,k}\left|\mathbf{e}_j^T\left(\hat{\mathbf{H}}_{(i),0}+\delta\mathbf{H}_{(i),l}\right)\hat{\mathbf{H}}_{(i),l}^T\mathbf{H}^\star{}_g\mathbf{\Sigma}^\star{}_{(i),g}\mathbf{W}^{\star T}_{(i),g}\mathbf{e}_k\right| \\
&\le \left(\max_j\left\|\mathbf{e}_j^T\hat{\mathbf{H}}_{(i),l,0}\right\|+\max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\right)\left\|\hat{\mathbf{H}}_{(i),l}^T\mathbf{H}^\star{}_g\right\|\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}} \\
&\le \left(\max_j\left\|\mathbf{e}_j^T\hat{\mathbf{H}}_{(i),l,0}\right\|+\max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\|\right)\left\|\Delta\mathbf{P}_{(i),l}\right\|\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}} \\
&\le \frac{\mu^2r}{\overline{n}}\sigma_{\max}\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\left(1+\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}+\left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^2\right)\left\|\Delta\mathbf{P}_{(i),l}\right\| \qquad (71) \\
&+\left\|\Delta\mathbf{P}_{(i),l}\right\|\sigma_{\max}\sqrt{\frac{\mu^2r}{n_2}}\max_j\left\|\mathbf{e}_j^T\delta\mathbf{H}_{(i),l}\right\| \qquad (TM16)
\end{aligned}$$

, where we again applied Lemma 19 in the first inequality. The second inequality comes from the relation $\left\|\hat{\mathbf{H}}_{(i),l}^T \mathbf{H^\star}_g\right\| = \left\|\hat{\mathbf{H}}_{(i),l}^T \left(\mathbf{I} - \mathbf{P}_{\hat{\mathbf{H}}_g}\right) \mathbf{H^\star}_g\right\| = \left\|\hat{\mathbf{H}}_{(i),l}^T \left(\mathbf{P}_{\mathbf{H^\star}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g} \mathbf{P}_{\mathbf{H^\star}_g}\right) \mathbf{H^\star}_g\right\| \leq \left\|\mathbf{P}_{\mathbf{H^\star}_g} - \mathbf{P}_{\hat{\mathbf{H}}_g} \mathbf{P}_{\mathbf{H^\star}_g}\right\| \leq \left\|\mathbf{P}_{\mathbf{H^\star}_g} - \mathbf{P}_{\mathbf{H^\star}_g}\right\|$.

Combining these sixteen terms (TM1)-(TM16) and considering the fact that $\alpha \leq 1$, we have,

$$\left\|\mathbf{L^\star}_{(i)} - \hat{\mathbf{L}}_{(i)}\right\|_\infty \leq \sqrt{\alpha}\mu^2 r \left\|\mathbf{E}\right\|_\infty C_4,$$

where

$$C_4 = 34327 \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^9 + 534 \left(\frac{2\sigma_{\max}^2}{\frac{3}{4}\sigma_{\min}^2}\right)^5 \theta^{-\frac{1}{2}} = \mathcal{O}\left(\kappa^9 + \kappa^5 \theta^{-\frac{1}{2}}\right). \tag{72}$$

This completes our proof. ■

Finally we will prove Theorem 5. We will first state its formal version below.

Theorem 21 *Suppose that the conditions of Lemma 19 are satisfied. Additionally, suppose that there exists a constant* $0 < \rho_{\min} < 1$ *such that* $\alpha \leq \frac{\rho_{\min}^2}{4\mu^4 r^2 C_4^2}$. *Then, the following statements hold at iteration* $t \geq 1$ *of Algorithm 1 with* $\lambda_1 = \frac{\sigma_{\max}\mu^2 r}{\sqrt{n_1 n_2}}$, $\epsilon \leq \lambda_1 (1 - \rho_{\min})$, *and* $1 - \frac{\epsilon}{\lambda_1} > \rho \geq \rho_{\min}$*:*

1. $supp\left(\hat{\mathbf{S}}_{(i),t}\right) \subset supp\left(\mathbf{S^\star}_{(i)}\right)$ *for every* $i \in [N]$.
2. $\left\|\hat{\mathbf{S}}_{(i),t} - \mathbf{S^\star}_{(i)}\right\|_\infty \leq 2\lambda_t \leq 4\sigma_{\max}\frac{\mu^2 r}{\overline{n}}$ *for every* $i \in [N]$.
3. $\left\|\hat{\mathbf{L}}^\epsilon_{(i),t} - \mathbf{L^\star}\right\|_\infty \leq \epsilon + \rho\lambda_t$ *for every* $i \in [N]$.

Moreover, we have

$$\left\|\hat{\mathbf{U}}^\epsilon{}_{g,t} \hat{\mathbf{V}}^{\epsilon T}_{(i),g,t} - \mathbf{U^\star}_g \mathbf{V^\star}^T_{(i),g}\right\|_\infty = \mathcal{O}\left(\rho^t + \frac{\epsilon}{1-\rho}\right), \quad \textit{for every } i \in [N]. \tag{73}$$

and

$$\left\|\hat{\mathbf{U}}^\epsilon{}_{(i),l,t} \hat{\mathbf{V}}^{\epsilon T}_{(i),l,t} - \mathbf{U^\star}_{(i),l} \mathbf{V^\star}^T_{(i),l}\right\|_\infty = \mathcal{O}\left(\rho^t + \frac{\epsilon}{1-\rho}\right), \quad \textit{for every } i \in [N]. \tag{74}$$

Remark The definition of the term $\rho_{\min}$ in the statement of the above theorem is kept intentionally implicit to streamline the presentation. In what follows, we will give an estimate of the requirements on α purely in terms of the parameters of the problem. Lemma 14 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^8\right)$. Lemma 15 requires $\alpha = \mathcal{O}\left(\frac{1}{\mu^2 r}\right)$. Lemma 16 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^{12}\right)$. Lemma 17 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^{12}\right)$. Lemma 18 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^2 r}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^6\right)$. And Lemma 19 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^2 r}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^2\right)$.

As $C_4 = \mathcal{O}\left(\frac{1}{\sqrt{\theta}}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{10}\right)$, the additional requirement in Theorem 21 requires $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^{20}\right)$. Taking the intersections of all these requirements, we can derive the upper bound on α as $\alpha = \mathcal{O}\left(\frac{\theta}{\mu^4 r^2}\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^{20}\right)$.

Proof We will prove this theorem by induction.

Base case: At $t = 1$, $\hat{\mathbf{L}}_{(i),0} = 0$. As $\lambda_1 = \frac{\mu^2 r}{\overline{n}}\sigma_{\max}$, we have $\hat{\mathbf{S}}_{(i),1} = \text{Hard}_{\frac{\mu^2 r}{\overline{n}}\sigma_{\max}}\left[\mathbf{M}_{(i)}\right]$. By definition of hard-thresholding, if the jk-th entry of $\hat{\mathbf{S}}_{(i),1}$ is nonzero, we know $\left|[\mathbf{M}_{(i)}]_{jk}\right| > \frac{\mu^2 r}{\overline{n}}\sigma_{\max}$. Since $\left|[\mathbf{L}^\star{}_{(i)}]_{jk}\right| \leq \frac{\mu^2 r}{\overline{n}}\sigma_{\max}$ for each j and k, we must have $\left|[\mathbf{S}^\star{}_{(i)}]_{jk}\right| > 0$. This proves Claim 1 for $t = 1$.

Now we will prove Claim 2 holds when $t = 1$. If $[\hat{\mathbf{S}}_{(i),1}]_{jk} = 0$, we know $\left|[\mathbf{S}^\star{}_{(i)}]_{jk} + [\mathbf{L}^\star{}_{(i)}]_{jk}\right| \leq \mu^2 r/\overline{n}\sigma_{\max}$, thus $\left|[\mathbf{S}^\star{}_{(i)}]_{jk}\right| \leq 2\mu^2 r/\overline{n}\sigma_{\max}$. If $[\hat{\mathbf{S}}_{(i),1}]_{jk} \neq 0$, by the definition of hard-thresholding, we know $[\hat{\mathbf{S}}_{(i),1}]_{jk} = [\mathbf{M}_{(i)}]_{jk} = [\mathbf{S}^\star{}_{(i)}]_{jk} + [\mathbf{L}^\star{}_{(i)}]_{jk}$. By rearranging terms, we have $\left|[\mathbf{S}^\star{}_{(i)}]_{jk} - [\hat{\mathbf{S}}_{(i),1}]_{jk}\right| = \left|[\mathbf{L}^\star{}_{(i)}]_{jk}\right| \leq \mu^2 r/\overline{n}\sigma_{\max}$. We hence proved Claim 2 for $t = 1$.

Since $\mathbf{E}_{(i),1} = \mathbf{S}^\star{}_{(i)} - \hat{\mathbf{S}}_{(i),1}$, we have $\left\|\mathbf{E}_{(i),1}\right\|_\infty \leq 2\frac{\mu^2 r}{\overline{n}}\sigma_{\max}$ for each i as well. Also, by Claim 1, $\mathbf{E}_{(i),1}$'s are α-sparse. Therefore by Lemma 20, when $\alpha \leq \frac{\rho^2_{\min}}{4\mu^4 r^2 C_4^2} \leq \frac{\rho^2}{4\mu^4 r^2 C_4^2}$, $\left\|\hat{\mathbf{L}}_{(i),1} - \mathbf{L}^\star{}_{(i)}\right\|_\infty \leq 2\sqrt{\alpha}\frac{\mu^2 r}{\overline{n}}\sigma_{\max}C_4 \leq \rho\lambda_1$. From the definition of ϵ-optimality and triangle inequality, we know $\left\|\hat{\mathbf{L}}^\epsilon_{(i),1} - \mathbf{L}^\star{}_{(i)}\right\|_\infty \leq \rho\lambda_1 + \epsilon$. We thus proved Claim 3 for $t = 1$.

Induction step: Now supposing that Claims 1, 2, and 3 hold for iterations $1, \cdots, t$, we will show their correctness for the iteration $t + 1$. Since Claim 3 holds for iteration t, we know $\left\|\hat{\mathbf{L}}^\epsilon_{(i),t} - \mathbf{L}^\star{}_{(i)}\right\|_\infty \leq \rho\lambda_t + \epsilon$ under the condition $\alpha \leq \frac{\rho^2}{4\mu^4 r^2 C_4^2}$. With the choice of $\lambda_{t+1} = \rho\lambda_t + \epsilon$, if the jk-th entry of $\hat{\mathbf{S}}_{(i),t+1}$ is nonzero, we have $\left|[\mathbf{S}^\star{}_{(i)}]_{jk} + [\mathbf{L}^\star{}_{(i)}]_{jk} - [\hat{\mathbf{L}}^\epsilon_{(i),t}]_{jk}\right| > \lambda_{t+1}$. Since $\left|[\mathbf{L}^\star{}_{(i)}]_{jk} - [\hat{\mathbf{L}}_{(i),t}]_{jk}\right| \leq \lambda_{t+1}$, we must have $\left|[\mathbf{S}^\star{}_{(i)}]_{jk}\right| > 0$. This proves Claim 1 for iteration $t + 1$.

We will now proceed to prove Claim 2. We consider each entry of $\hat{\mathbf{S}}_{(i),t+1} = \text{Hard}_{\lambda_{t+1}}\left[\mathbf{S}^\star{}_{(i)} + \mathbf{L}^\star{}_{(i)} - \hat{\mathbf{L}}^\epsilon_{(i),t}\right]$. From the definition of hard-thresholding, we know $\left|[\hat{\mathbf{S}}_{(i),t+1}]_{jk} - \left([\mathbf{S}^\star{}_{(i)}]_{jk} + [\mathbf{L}^\star{}_{(i)}]_{jk} - [\hat{\mathbf{L}}^\epsilon_{(i),t}]_{jk}\right)\right| \leq \lambda_{t+1}$. Remember that we know $\left|[\mathbf{L}^\star{}_{(i)}]_{jk} - [\hat{\mathbf{L}}^\epsilon_{(i),t}]_{jk}\right| \leq \lambda_{t+1}$ from the correctness of Claim 3 at iteration t and the upper bound on α, we can derive $\left|[\hat{\mathbf{S}}_{(i),t+1}]_{jk} - [\mathbf{S}^\star{}_{(i)}]_{jk}\right| \leq 2\lambda_{t+1}$ by triangle inequality. We hence prove Claim 2.

For Claim 3, since $\mathbf{E}_{(i),t+1} = \mathbf{S}^\star{}_{(i)} - \hat{\mathbf{S}}_{(i),t+1}$, we have $\left\|\mathbf{E}_{(i),t+1}\right\|_\infty \leq 2\lambda_{t+1}$ for each i as well. Also, by Claim 1 at iteration t, $\mathbf{E}_{(i),t}$'s are α-sparse at iteration t. Therefore by Lemma 20, $\left\|\hat{\mathbf{L}}_{(i),t+1} - \mathbf{L}^\star\right\|_\infty \leq 2\sqrt{\alpha}\mu^2 r C_4\lambda_{t+1}$. Under the constraint that $\alpha \leq \frac{\rho^2}{4\mu^4 r^2 C_4^2}$, we know

$\left\|\hat{\mathbf{L}}_{(i),t+1} - \mathbf{L}^\star\right\|_\infty \leq \rho\lambda_{t+1}$. From the definition of ϵ-optimality and triangle inequality, we have $\left\|\hat{\mathbf{L}}^\epsilon_{(i),t+1} - \mathbf{L}^\star\right\|_\infty \leq \rho\lambda_{t+1} + \epsilon$. We thus proved Claim 3 at iteration $t+1$.

Combining them, we can conclude that 1, 2, and 3 hold for every $t = 1, 2, \cdots$.

Finally, we will prove (73) and (74). We have known that $\hat{\mathbf{U}}_{g,t}\hat{\mathbf{V}}^T_{(i),g,t} = \mathbf{P}_{\hat{\mathbf{H}}_g}\hat{\mathbf{M}}_{(i)}$, then from similar analysis of (61), we have,

$$\begin{aligned}
&\left\|\hat{\mathbf{U}}_{g,t}\hat{\mathbf{V}}^T_{(i),g,t} - \mathbf{U^\star}_g\mathbf{V^\star}^T_{(i),g}\right\|_\infty \\
&= \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}\left(\mathbf{L^\star}_{(i),g} + \mathbf{L^\star}_{(i),l} + \mathbf{E}_{(i)}\right) - \mathbf{L^\star}_{(i),g}\right\|_\infty \\
&\leq \left\|\mathbf{P}_{\hat{\mathbf{H}}_g}\mathbf{L^\star}_{(i),l}\right\|_\infty \\
&+ \Big|\Big| \left(\mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_1^{-2}\hat{\mathbf{H}}_g^T\mathbf{T}_{(0)} + \mathbf{T}_{(0)}\hat{\mathbf{H}}_g\mathbf{\Lambda}_1^{-1}\delta\mathbf{H}_g^T + \delta\mathbf{H}_g\mathbf{\Lambda}_1^{-1}\hat{\mathbf{H}}_g^T\mathbf{T}_{(0)} + \delta\mathbf{H}_g\delta\mathbf{H}_g^T\right) \\
&(\mathbf{L^\star}_{(i),g} + \mathbf{E}_{(i),t}) - \mathbf{L^\star}_{(i),g}\Big|\Big|_\infty.
\end{aligned}$$

In Lemma 20, we have shown that each term above is upper bounded by $\mathcal{O}(\max_i \left\|\mathbf{E}_{(i),t}\right\|_\infty)$. Therefore by Claim 2, we have $\left\|\hat{\mathbf{U}}_{g,t}\hat{\mathbf{V}}^T_{(i),g,t} - \mathbf{U^\star}_g\mathbf{V^\star}^T_{(i),g}\right\|_\infty = \mathcal{O}(\max_i \left\|\mathbf{E}_{(i),t}\right\|_\infty) = \mathcal{O}(\lambda_t) = \mathcal{O}(\rho^t + \frac{\epsilon}{1-\rho})$. (73) follows accordingly by triangle inequality.

We can prove (74) in a similar way. This completes our proof of Theorem 21. ■

Appendix D. Auxiliary Lemma

This section discusses some helper lemmas useful for our main proofs. These lemmas are mostly derived from basic linear algebra and series.

The following lemma is a well-known result and provides an upper bound on the norm of product matrices.

Lemma 22 *Form two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, we have,*

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F$$

and

$$\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$$

Proof The proof is straightforward and can be found in Sun and Luo (2016). ■

Lemma 23 *For $x, y \in [0, 1)$ such that $x + y < 1$, the following relation holds:*

$$\sum_{p_1+p_2\geq 1} x^{p_1+p_2} \leq \frac{2x}{1-x}. \tag{75a}$$

Proof This proof follows from the direct calculation.

$$\begin{aligned}
&\sum_{p_1+p_2\geq 1} x^{p_1+p_2} = \sum_{p_1=0}^{\infty}\sum_{p_2=0}^{\infty} x^{p_1+p_2} - 1 \\
&= \left(\sum_{p_1=0}^{\infty} x^{p_1}\right)\left(\sum_{p_2=0}^{\infty} x^{p_2}\right) - 1 = \frac{1}{(1-x)^2} - 1 \\
&= \frac{2x-x^2}{(1-x)^2} \leq \frac{2x}{1-x}.
\end{aligned}$$

■

We also present a lemma related to the Schur complement of block matrices.

Lemma 24 *For symmetric matrices* $\mathbf{A}_0 \in \mathbb{R}^{r_0\times r_0}, \mathbf{A}_1 \in \mathbb{R}^{r_1\times r_1}, \cdots, \mathbf{A}_N \in \mathbb{R}^{r_1\times r_1}$*, and* $\mathbf{B}_i \in \mathbb{R}^{r_0\times r_N}$ *for* $i \in \{1, \cdots, N\}$*, we can construct a symmetric block matrix* $\mathbf{C}$ *as,*

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_N \\ \mathbf{B}_1^T & \mathbf{A}_1 & 0 & \cdots & 0 \\ \mathbf{B}_2^T & 0 & \mathbf{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & 0 \\ \mathbf{B}_N^T & 0 & 0 & \cdots & \mathbf{A}_N \end{pmatrix}. \tag{76}$$

Then, $\mathbf{C}$ *is positive definite if and only if* $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_N$ *are positive definite and* $\mathbf{A}_0 - \sum_{i=1}^N \mathbf{B}_i\mathbf{A}_i^{-1}\mathbf{B}_i^T$ *is positive definite.*

Proof Sicne $\mathbf{A}_i$'s are positive definite, they are invertible. Thus we can decompose $\mathbf{C}$ as

$$\mathbf{C} = \underbrace{\begin{pmatrix} \mathbf{I} & \mathbf{B}_1\mathbf{A}_1^{-1} & \mathbf{B}_2\mathbf{A}_2^{-1} & \cdots & \mathbf{B}_N\mathbf{A}_N^{-1} \\ 0 & \mathbf{I} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{I} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \mathbf{I} \end{pmatrix}}_{\mathbf{C}_1} \underbrace{\begin{pmatrix} \mathbf{A}_0 - \sum_i \mathbf{B}_i\mathbf{A}_i^{-1}\mathbf{B}_i^T & 0 & \cdots & 0 \\ 0 & \mathbf{A}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{A}_N \end{pmatrix}}_{\mathbf{C}_2}$$
$$\cdot \underbrace{\begin{pmatrix} \mathbf{I} & 0 & 0 & \cdots & 0 \\ \mathbf{A}_1^{-1}\mathbf{B}_1^T & \mathbf{I} & 0 & \cdots & 0 \\ \mathbf{A}_2^{-1}\mathbf{B}_2^T & 0 & \mathbf{I} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{A}_N^T\mathbf{B}_N^T & 0 & 0 & \cdots & \mathbf{I} \end{pmatrix}}_{\mathbf{C}_1^T}.$$

On the right hand side, $\mathbf{C}_1$ and $\mathbf{C}_1^T$ are both invertible. Thus, $\mathbf{C}$ is positive definite if and only if $\mathbf{C}_2$ is positive definite. Since $\mathbf{C}_2$ is a block diagonal matrix, we prove the statement in the lemma. ■

The following lemma provides an eigenvalue lower bound on the product of three matrices.

Lemma 25 *For matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, and symmetric positive semidefinite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, we know that,*

$$\lambda_{\min}\left(\mathbf{A}^T\mathbf{B}\mathbf{A}\right) \geq \lambda_{\min}\left(\mathbf{B}\right)\lambda_{\min}\left(\mathbf{A}^T\mathbf{A}\right).$$

Proof The proof follows from the Courant–Fischer–Weyl variational principle.

$$\begin{aligned}
&\lambda_{\min}\left(\mathbf{A}^T\mathbf{B}\mathbf{A}\right) = \min_{\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{A}^T\mathbf{B}\mathbf{A}\mathbf{v} \\
&\geq \lambda_{\min}\left(\mathbf{B}\right) \min_{\|\mathbf{v}\|=1} \|\mathbf{A}\mathbf{v}\|^2 \\
&= \lambda_{\min}\left(\mathbf{B}\right)\lambda_{\min}\left(\mathbf{A}^T\mathbf{A}\right).
\end{aligned}$$

■

We finally present the lemma that provides an upper bound of the operator norm of block matrices.

Lemma 26 *For a symmetric block matrix $\mathbf{C}$ defined as*

$$\mathbf{C} = \begin{pmatrix}
\mathbf{A}_1 & \mathbf{B}_{12} & \mathbf{B}_{13} & \cdots & \mathbf{B}_{1N} \\
\mathbf{B}_{12}^T & \mathbf{A}_2 & \mathbf{B}_{23} & \cdots & \mathbf{B}_{2N} \\
\mathbf{B}_{13}^T & \mathbf{B}_{23}^T & \mathbf{A}_3 & \cdots & \mathbf{B}_{3N} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{B}_{1N}^T & \mathbf{B}_{2N}^T & \mathbf{B}_{3N}^T & \cdots & \mathbf{A}_N
\end{pmatrix}, \tag{77}$$

where $\mathbf{A}_i$'s are symmetric, we have,

$$\|\mathbf{C}\| \leq \max_{i=1,\cdots,N}\{\|\mathbf{A}_i\|\} + \sqrt{2\sum_{i<j}\|\mathbf{B}_{ij}\|^2}. \tag{78}$$

Note: in (78), the diagonal blocks and off-diagonal blocks are treated differently.

Proof We first prove for the special case where $\mathbf{B}_{ij} = 0$. In this case,

$$\begin{aligned}
&\|\mathbf{C}\|^2 = \max_{\|\mathbf{v}\|=1} \|\mathbf{C}\mathbf{v}\|^2 \\
&= \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \|\mathbf{A}_i\mathbf{v}_i\|^2 \leq \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \|\mathbf{A}_i\|^2 \|\mathbf{v}_i\|^2
\end{aligned}$$

$$\leq \max_{i=1,\cdots,N}\{\|\mathbf{A}_i\|^2\} \times \sum_{i=1}^{N} \|\mathbf{v}_i\|^2 = \max_{i=1,\cdots,N}\{\|\mathbf{A}_i\|^2\}.$$

We then prove for the special case where $\mathbf{A}_i = 0$. We have

$$\begin{aligned}
\|\mathbf{C}\|^2 &= \max_{\|\mathbf{v}\|=1} \|\mathbf{C}\mathbf{v}\|^2 = \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \left\| \sum_{j\neq i} \mathbf{B}_{ij}\mathbf{v}_j \right\|^2 \\
&= \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \sum_{j,k\neq i} \mathbf{v}_k^T \mathbf{B}_{ik}^T \mathbf{B}_{ij}\mathbf{v}_j \leq \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \sum_{j,k\neq i} \|\mathbf{v}_k\| \, \|\mathbf{B}_{ik}\| \, \|\mathbf{B}_{ij}\| \, \|\mathbf{v}_j\| \\
&= \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \left(\sum_{j\neq i} \|\mathbf{B}_{ij}\| \, \|\mathbf{v}_j\| \right)^2 \leq \max_{\|\mathbf{v}\|=1} \sum_{i=1}^{N} \left(\sum_{j\neq i} \|\mathbf{B}_{ij}\|^2 \right) \left(\sum_{j} \|\mathbf{v}_j\|^2 \right) \\
&= 2 \sum_{i<j} \|\mathbf{B}_{ij}\|^2,
\end{aligned}$$

where we used Cauchy-Schwarz inequality in the second inequality.

By applying triangle inequality of the matrix operator norm, we can combine the upper bounds derived from two special cases and obtain (78). ■

References

M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11): 4311–4322, 2006.

D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48 (3):334–334, 1997.

R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

T. Bouwmans and E. H. Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.11.009. URL `https://www.sciencedirect.com/science/article/pii/S1077314213002294`.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

D. Chai, L. Wang, K. Chen, and Q. Yang. Secure federated matrix factorization. *IEEE Intelligent Systems*, 36(5):11–20, 2021. doi: 10.1109/MIS.2020.3014880.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

X. Chen, B. Zhang, T. Wang, A. Bonni, and G. Zhao. Robust principal component analysis for accurate outlier sample detection in rna-seq data. *Bmc Bioinformatics*, 21(1):1–20, 2020.

Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *Annals of statistics*, 49(5):2948, 2021.

J. Fan, W. Wang, and Y. Zhong. An l-infinity eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18 (207):1–42, 2018.

S. Fattahi and S. Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of machine learning research*, 2020.

Q. Feng, M. Jiang, J. Hannig, and J. Marron. Angle-based joint and individual variation explained. *Journal of multivariate analysis*, 166:241–265, 2018.

I. Gaynanova and G. Li. Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132, 2019.

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77, 2011.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933. doi: http://dx.doi.org/10.1037/h0071325.

D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

N. Jin, S. Zhou, and T.-S. Chang. *Identification of impacting factors of surface defects in hot rolling processes using multi-level regression analysis*. Society of Manufacturing Engineers Southfield, MI, USA, 2000.

R. Kashyap, R. Kong, S. Bhattacharjee, J. Li, J. Zhou, and B. T. Yeo. Individual-specific fmri-subspaces improve functional connectivity prediction of behavior. *NeuroImage*, 189: 804–812, 2019.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

H. Lee and S. Choi. Group nonnegative matrix factorization for eeg classification. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 320–327, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/lee09a.html`.

X. Li, J. Haupt, J. Lu, Z. Wang, R. Arora, H. Liu, and T. Zhao. Symmetry. saddle points, and global optimization landscape of nonconvex matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9, 2018. doi: 10.1109/ITA.2018.8503215.

X. Li, S. Wang, and Y. Cai. Tutorial: Complexity analysis of singular value decomposition and its variants. *arXiv preprint arXiv:1906.12085*, 2019.

G. Liang, N. Shi, R. A. Kontar, and S. Fattahi. Personalized dictionary learning for heterogeneous datasets. In *Advances in Neural Information Processing Systems*, 2023.

E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

D. Meng and F. De La Torre. Robust matrix factorization with unknown noise. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1337–1344, 2013.

P. Netrapalli, N. UN, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust pca. *Advances in neural information processing systems*, 27, 2014.

Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Robust correlated and individual component analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38 (8):1665–1678, 2015.

D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/park17a.html`.

J. Y. Park and E. F. Lock. Integrative factorization of bidimensionally linked matrices. *Biometrics*, 76(1):61–74, 2020.

E. Ponzi, M. Thoresen, and A. Ghosh. Rajive: Robust angle based jive for integrating noisy multi-source data. *arXiv preprint arXiv:2101.09110*, 2021.

A. Rinaldo. Davis-kahan theorem. *Advanced Statistical Theory I*, 2017.

C. Sagonas, Y. Panagakis, A. Leidinger, and S. Zafeiriou. Robust joint and individual variance explained. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2017.

B. Shen, W. Xie, and Z. J. Kong. Smooth robust tensor completion for background/foreground separation with missing pixels: Novel algorithm with convergence guarantee. *Journal of Machine Learning Research*, 23(217):1–40, 2022.

N. Shi and R. A. Kontar. Personalized pca: Decoupling shared and unique features. *Journal of Machine Learning Research*, 25(41):1–82, 2024. URL `http://jmlr.org/papers/v25/22-0810.html`.

N. Shi, R. A. Kontar, and S. Fattahi. Heterogeneous matrix factorization: When features differ by datasets. *arXiv preprint arXiv:2305.17744*, 2023.

R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Y. L. Tan, V. Sehgal, and H. H. Shahri. Sensoclean: Handling noisy and incomplete data in sensor networks using modeling. *Main*, pages 1–18, 2005.

T. Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. `https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/`, 2010. Accessed: 2022-03-01.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 964–973, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/tu16.html`.

A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre. A benchmark dataset for outdoor foreground/background extraction. In *Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I 11*, pages 291–300. Springer, 2013.

N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy. Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. *IEEE Signal Processing Magazine*, 35(4):32–55, 2018. doi: 10.1109/MSP.2018.2826566.

R. K. Wong and T. C. Lee. Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research*, 18(1):5404–5428, 2017.

J. Wright and Y. Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

W. Xiao, X. Huang, F. He, J. Silva, S. Emrani, and A. Chaudhuri. Online robust principal component analysis with change point detection. *IEEE Transactions on Multimedia*, 22 (1):59–68, 2020. doi: 10.1109/TMM.2019.2923097.

H. Yan, K. Paynabar, and J. Shi. Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics*, 60(2):181–197, 2018.

Z. Yang and G. Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2016.

T. Ye and S. S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.

L. Zhang, H. Shen, and J. Z. Huang. Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics*, pages 1540–1561, 2013.

G. Zhou, A. Cichocki, Y. Zhang, and D. P. Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE transactions on neural networks and learning systems*, 27(11):2426–2439, 2015.