Optimal Clustering by Lloyd's Algorithm for Low-Rank Mixture Model

Zhongyuan Lyu[1] **and Dong Xia**[1]

[1]*Department of Mathematics, Hong Kong University of Science and Technology , e-mail:* zlyuab@connect.ust.hk*;* madxia@ust.hk

Abstract: This paper investigates the computational and statistical limits in clustering matrix-valued observations. We propose a low-rank mixture model (LrMM), adapted from the classical Gaussian mixture model (GMM) to treat matrix-valued observations, which assumes low-rankness for population center matrices. A computationally efficient clustering method is designed by integrating Lloyd's algorithm and low-rank approximation. Once well-initialized, the algorithm converges fast and achieves an exponential-type clustering error rate that is minimax optimal. Meanwhile, we show that a tensor-based spectral method delivers a good initial clustering. Comparable to GMM, the minimax optimal clustering error rate is decided by the *separation strength*, i.e, the minimal distance between population center matrices. By exploiting low-rankness, the proposed algorithm is blessed with a weaker requirement on separation strength. Unlike GMM, however, the statistical and computational difficulty of LrMM is characterized by the *signal strength*, i.e, the smallest non-zero singular values of population center matrices. Evidences are provided showing that no polynomial-time algorithm is consistent if the signal strength is not strong enough, even though the separation strength is strong. The performance of our low-rank Lloyd's algorithm is further demonstrated under sub-Gaussian noise. Intriguing differences between estimation and clustering under LrMM are discussed. The merits of low-rank Lloyd's algorithm are confirmed by comprehensive simulation experiments. Finally, our method outperforms others in the literature on real-world datasets.

MSC2020 subject classifications: Primary 62C20; secondary 62F30.
Keywords and phrases: Mixtures models, low-rank matrix, clustering, Lloyd's algorithm.

1. Introduction

Nowadays, clustering *matrix-valued* observations becomes a ubiquitous task in diverse fields. For instance, each highly variable region (HVR) in the var genes of human malaria parasite [37, 31] is representable by an adjacency matrix and a key scientific question is to identify structurally-similar HVRs by, say, clustering the associated adjacency matrices. The international trade flow of a commodity across different countries can be viewed as a weighted adjacency matrix [47, 6]. Finding the similarity between trading patterns of different commodities is of great value in understanding the global economic structure. This can also be achieved by clustering the weighted adjacency matrices. Other notable examples include clustering multi-layer social networks [15, 23] and multi-view data

Dataset	n	(d_1, d_2)	K	Ranks
BHL [49]	27	(1124,4)	3	$\sim \{1, 1, 1\}$
EEG [66]	122	(256,64)	2	$\sim \{2, 1\}$
Malaria gene networks [37]	9	(212,212)	6	≤ 15
UN trade flow networks [47]	97	(48,48)	2	$\sim \{3, 2\}$

TABLE 1

Summary of datasets. Here, n is the sample size, (d_1, d_2) is the dimension of each matrix observation, and K is number of clusters. The underlying rank $(r_k's)$ of population center matrices from different clusters can be unequal.

[34, 49], modeling the connectivity of brain networks [2, 53], clustering the correlation networks between bacterial species [52], EEG data analysis [19], etc.

Since matrix-valued observations can always be vectorized, a naive approach is to ignore the matrix structure so that numerous classical clustering algorithms, e.g. K-means or spectral clustering, are readily applicable. However, matrix observations are usually blessed with hidden low-dimensional structures, among which low-rankness is perhaps the most common and explored. Network models such as *stochastic block model* [26, 31], *random dot product graph* [3] and *latent space model* [25] often assume a low-rank expectation of adjacency matrix. Low-rank structures have also been successfully explored in brain image clustering [53], EEG data analysis [19], and international trade flow data [47], to name but a few. Table 1 presents a summary of datasets analyzed in our paper, where the matrix ranks r_k's (suggested by the numerical performance of our algorithm) are much smaller than the ambient dimensions (d_1, d_2). Without loss of generality, we assume $d_1 \geq d_2$. For these applications, the naive clustering approach becomes statistically sub-optimal since the planted low-dimensional structure is overlooked.

Motivated by the aforementioned applications, throughout this paper, we assume that *each matrix-valued observation has a low-rank expectation* and the *expectations are equal for observations from the same cluster*. It is the essence of *low-rank mixture model* (LrMM), which shall be formally defined in Section 2. Several clustering methods exploiting low-rankness have emerged in the literature. [53] introduces a tensor Gaussian mixture model and recasts the clustering task as estimating the factors in low-rank tensor decomposition. K-means clustering is then applied to the estimated factors. While sharp estimation error rate is derived under a suitable signal-to-noise ratio (SNR) condition, the accuracy of clustering is not provided. A tensor normal mixture model is proposed by [49], where the authors have designed an enhanced EM algorithm for estimating the distributional parameters. Under appropriate conditions, sharp estimation error rates are established showing that minimax optimal *test* clustering error rate is attainable. However, the *training* clustering error is missing, and it is even unclear whether the proposed EM algorithm can consistently recover the true cluster memberships. Aimed at analyzing multi-layer networks, [31] propose a mixture multi-layer SBM where a spectral clustering method based on tensor decomposition is investigated. Clustering error rate is established under a fairly weak network sparsity condition, although the rate is likely sub-optimal.

More recently, [47] extends the mixture framework to latent space model and sub-optimal clustering error rate is also derived. Note that [31] and [47] both require a rather restrictive condition in that $n = O(d_1)$ rendering their theories unattractive in many scenarios. Other representative works include [9], [5], [19] and [52], but clustering error rates are not studied.

Note that LrMM reduces to the famous *Gaussian mixture model* (GMM) in the dimension $d^* := d_1 d_2$ if each matrix-valued observation has a full-rank expectation, and the noise matrix has i.i.d. standard normal entries. Under GMM, [43] proves that a spectral method attains, with high probability, an average mis-clustering error rate $\exp(-\Delta^2/8)$ that is optimal in the minimax sense. Here Δ is the minimal Euclidean distance between the expected centers of distinct clusters (i.e., population center matrices), referred to as the *separation strength*. This exponential rate is established by [43] under a separation strength[1] condition $\Delta \gg 1 + d^* n^{-1}$. [18] investigates a more general iterative algorithm that achieves the same exponential rate under a weaker separation strength condition $\Delta \gg 1 + (d^*/n)^{1/2}$. More recently, [65] applies the leave-one-out method and proves the optimality of spectral clustering under a relaxed separation strength condition. Besides deriving the optimal clustering error rate, prior works also made efforts to establish the phase transitions in exact recovery, i.e., when clustering error is zero. [50] investigates a power iteration algorithm for a two-component GMM and proves that exact recovery is attained if Δ^2 is greater than $\left(1 + (1 + 2d^* n^{-1} \log^{-1} n)^{1/2}\right) \cdot \log n$. In addition, the author shows that exact recovery is impossible if Δ^2 is smaller than the aforesaid threshold. Later, [10] establishes a similar phase transition for general K-component GMM based on a semidefinite programming (SDP) relaxation. These foregoing works suggest an intriguing gap in the regime $n = O(d^*)$: [50] and [10] reveal that exact recovery is achievable beyond the separation strength threshold $(2d^* n^{-1} \log n)^{1/4}$, whereas the exponential-type clustering error rate [18, 65] is derived only beyond the threshold $(d^*/n)^{1/2}$. To our best knowledge, the gap still exists at the moment. [30] proposes a two-component symmetric *sparse* GMM and investigates the phase transition in consistent clustering. Specifically, they show that, ignoring log factors, $\Delta \gg 1 + s/n$ is necessary for consistent clustering without restricting the computational complexity. Here s is the sparsity of the expected observation. A recent work [42] designs an SDP-based spectral method and establishes an exponential clustering error rate when Δ is greater than $1 + s^{1/2} \log^{1/4}(d^*) n^{-1/4}$. Moreover, they provide evidence supporting the claim that no polynomial-time algorithm can consistent recover the clusters if Δ is smaller than the aforesaid threshold, i.e., there exists a statistical-to-computational gap for clustering in sparse GMM. Both [30] and [42] imply that the necessary separation strength primarily depends on the intrinsic dimension s rather than the ambient dimension d^*. We remark that there is a vast literature studying the clustering problem for GMM. A representative but incomplete list includes [44, 4, 12, 16, 22, 56, 58, 1] and references therein.

In contrast, the understanding of the limit of clustering for LrMM is still at

[1] For narration simplicity, we set the number of clusters $K = O(1)$ here.

its infant stage. In this paper, we fill the void in optimal clustering error rate for LrMM and demonstrate that the rate is achievable by a computationally fast algorithm. Challenges are posed from multiple fronts. First of all, designing a computationally fast clustering procedure that sufficiently exploits low-rank structure is non-trivial. Unlike (sparse) GMM [10, 42], convex relaxation for LrMM seems not immediately accessible, especially when there are more than two clusters. Non-convex approaches based on tensor decomposition and spectral clustering [31, 45, 63] usually cannot distinguish the sample size dimension (i.e., n) and data point dimension (i.e., d_1, d_2). Their theoretical results become sub-optimal when sample size is much larger than d_1. On the technical front, low-rankness makes deriving an exact exponential-type clustering error rate even more difficult. Under GMM [18, 43], the exponential-type clustering error rate is established by carefully studying the concentration phenomenon of a Gaussian linear form that usually admits an explicit representation. Estimating procedures under LrMM, however, often require multiple iterations of low-rank approximation, say, by singular value decomposition (SVD). Consequently, deriving the concentration property of respective linear forms under LrMM is much more involved than that under GMM. Moreover, prior related works [42, 30, 64, 46] provide evidences that imply the existence of a statistical-to-computational gap. It is unclear which model parameter characterizes such a gap and how the gap depends on the sample size and dimensions. For instance, how the low-rankness benefits the separation strength requirement? Interestingly, we discover that the gap is not determined by the separation strength Δ but rather by the signal strength (to be defined in Section 2) of the population center matrices.

Our main contributions are summarized as follows. First, we propose a computationally fast clustering algorithm for LrMM. At its essence is the combination of Lloyd's algorithm [41, 44] and low-rank approximation. Basically, given the updated cluster memberships of each observation, the cluster centers are obtained by the SVD of sample average within each cluster. The whole algorithm involves only K-means clustering and matrix SVDs. Secondly, we prove that, equipped with a good initial clustering, the low-rank Lloyd's algorithm converges fast and achieves the minimax optimal clustering error rate $\exp(-\Delta^2/8)$ with high probability as long as the separation strength satisfies $\Delta^2 \gg r_{\max} + r_{\max}^2 d_1/n$ and the signal strength is strong enough. Here $r_{\max}$ is the maximum rank among all the population center matrices. This dictates that a weaker separation strength is sufficient for clustering under LrMM if the rank $r_{\max} = O(1)$. Our key technical tool to develop the exponential-type error rate is a spectral representation formula from [60], which has helped push forward the understanding of statistical inference for low-rank models [61, 62]. Thirdly, we propose a novel tensor-based spectral method for obtaining an initial clustering. Under similar separation strength and signal strength conditions, this method delivers an initial clustering that is sufficiently good for ensuring the convergence of low-rank Lloyd's algorithm. Lastly, compared with GMM that only requires a separation strength condition [42, 18], an additional signal strength condition seems necessary under LrMM. We provide evidences, based on the

low-degree framework [36], showing that if the signal strength condition fails, all polynomial-time algorithms cannot consistently recover the true clusters, even when the separation strength is much stronger than the aforesaid one. It is worth pointing out that, unlike tensor-based approaches [31, 45, 63], our theoretical results impose no constraints on the relation between n and (d_1, d_2).

The rest of the paper is organized as follows. Low-rank mixture model is formalized in Section 2, and we introduce the low-rank Lloyd's algorithm and a tensor-based method for spectral initialization. The convergence performance of Lloyds' algorithm, minimax optimal exponential-type clustering error rate, and guarantees of a tensor-based spectral initialization are established in Section 3. We discuss the computational barriers of LrMM in Section 4. In Section 5, we slightly modify the low-rank Lloyd's algorithm and derive the same minimax optimal clustering error rate requiring a slightly weaker signal strength condition. Our theoretical results are extended to the case of sub-Gaussian noise in Section 6. We discuss the difference between estimation and clustering under LrMM in Section 7. Further discussions are provided in Section 8. Numerical simulations and real data examples are presented in Section 9. All proofs and technical lemmas are relegated to the appendix.

2. Methodology

2.1. Background and notations

For nonnegative D_1, D_2 , the notation $D_1 \lesssim D_2$ (equivalently, $D_2 \gtrsim D_1$) means that there exists an absolute constant $C > 0$ such that $D_1 \leq CD_2$; $D_1 \asymp D_2$ is equivalent to $D_1 \lesssim D_2$ and $D_2 \lesssim D_1$, simultaneously. Let $\|\cdot\|$ denote the ℓ_2 norm for vectors and operator norm for matrices, and $\|\cdot\|_{\mathrm{F}}$ denotes the matrix Frobenius norm. Denote $\sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_r(\mathbf{M}) > 0$ the non-increasing singular values of $\mathbf{M}$ where $r = \mathrm{rank}(\mathbf{M})$. We also define $\sigma_{\min}(\mathbf{M}) := \sigma_r(\mathbf{M})$. A third order tensor is a three-dimensional array. Throughout the paper, a tensor is written in the calligraphic bold font, e.g. $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{d_1 \times d_2 \times n}$. We use $\mathscr{M}_1(\boldsymbol{\mathcal{M}})$ to denote the mode-1 matricization of $\boldsymbol{\mathcal{M}}$ such that $\mathscr{M}_1(\boldsymbol{\mathcal{M}}) \in \mathbb{R}^{d_1 \times (d_2 n)}$ and $\mathscr{M}_1(\boldsymbol{\mathcal{M}})(i_1, (i_2 - 1)n + i_3) = \boldsymbol{\mathcal{M}}(i_1, i_2, i_3), \forall i_1 \in [d_1], i_2 \in [d_2], i_3 \in [n]$. The mode-2 and mode-3 matricizations are defined in a similar fashion. Then $\big\{\mathrm{rank}\big(\mathscr{M}_k(\boldsymbol{\mathcal{M}})\big) : k = 1, 2, 3\big\}$ are called *Tucker rank* or *multilinear rank*. The mode-1 marginal multiplication between $\boldsymbol{\mathcal{M}}$ and a matrix $\mathbf{U}^\top \in \mathbb{R}^{r \times d_1}$ results into a tensor of size $r_1 \times d_2 \times n$, whose elements are

$$\big(\boldsymbol{\mathcal{M}} \times_1 \mathbf{U}^\top\big)(j_1, i_2, i_3) := \sum_{i_1=1}^{d_1} \boldsymbol{\mathcal{M}}(i_1, i_2, i_3)\mathbf{U}(i_1, j_1), \quad \forall j_1 \in [r], i_2 \in [d_2], i_3 \in [n]$$

Similarly, we can define the mode-2 and mode-3 marginal multiplication. Given $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \mathbf{V} \in \mathbb{R}^{d_2 \times r_2}, \mathbf{W} \in \mathbb{R}^{n \times r_3}$, the multi-linear product $\mathcal{M} := \boldsymbol{\mathcal{S}} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}$ outputs a $d_1 \times d_2 \times n$ tensor defined by,

$$\boldsymbol{\mathcal{M}}(i_1, i_2, i_3) := \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} \sum_{j_3=1}^{r_3} \boldsymbol{\mathcal{S}}(j_1, j_2, j_3)\mathbf{U}(i_1, j_1)\mathbf{V}(i_2, j_2)\mathbf{W}(i_3, j_3) \quad (1)$$

More details can be found in [32]. Denote $\mathbb{O}_{d,r}$ the set of all $d \times r$ matrices $\mathbf{U}$ such that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix. Eq. (1) is known as the *Tucker decomposition* if $r_k = \mathrm{rank}\big(\mathscr{M}_k(\boldsymbol{\mathcal{M}})\big)$, $\mathbf{U} \in \mathbb{O}_{d_1,r_1}, \mathbf{V} \in \mathbb{O}_{d_2,r_2}$, and $\mathbf{W} \in \mathbb{O}_{n,r_3}$.

2.2. Low-rank Gaussian mixture

Suppose that the $d_1 \times d_2$ matrix-valued observations $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are i.i.d., and each of them has a latent label $s_i^* \in [K]$. Here K denotes the number of underlying clusters, and without loss of generality, assume $d_1 \geq d_2$. We assume that there exist K deterministic but *unknown* matrices $\mathbf{M}_1, \cdots, \mathbf{M}_K$ such that, conditioned on $s_i^* = k$, $\mathbf{X}_i$ follows a matrix normal distribution $\mathcal{N}(\mathbf{M}_k, \mathbf{I}_{d_1} \otimes \mathbf{I}_{d_2})$ in that $\mathrm{vec}(\mathbf{X}_i) \sim \mathcal{N}\big(\mathrm{vec}(\mathbf{M}_k), \mathbf{I}_{d_1 d_2}\big)$. This implies that $\mathbf{X}_i | s_i^* = k$ is equal to $\mathbf{M}_k + \mathbf{E}_i$ *in distribution* where $\mathbf{E}_i$ has i.i.d. standard normal entries. Moreover, we assume that the latent labels $s_1^*, \cdots, s_n^*$ are i.i.d. and

$$\mathbb{P}(s_i^* = k) = \pi_k, \quad \forall k \in [K]; \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1. \tag{2}$$

Here the unknown $\pi_k > 0$ stands for the mass of k-th cluster. Put it differently, the matrix-valued observations have a marginal distribution

$$\mathbf{X}_1, \cdots, \mathbf{X}_n \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{k} \pi_k \cdot \mathcal{N}(\mathbf{M}_k, \mathbf{I}_{d_1} \otimes \mathbf{I}_{d_2}) \tag{3}$$

Let $r_k = \mathrm{rank}(\mathbf{M}_k)$ and assume $r_k \ll d_2$ for all k, i.e., all the population center matrices are low-rank. Model (3) is referred to as the low-rank mixture model (LrMM). For simplicity, we treat the ranks r_k's as known and will briefly discuss how to estimate them in Section 8. We denote the compact SVD of population center matrices by $\mathbf{M}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\top$ with $\mathbf{U}_k \in \mathbb{O}_{d_1,r_k}$ and $\mathbf{V}_k \in \mathbb{O}_{d_2,r_k}$. The *signal strength* of $\mathbf{M}_k$ is characterized by $\sigma_{\min}(\mathbf{M}_k) := \sigma_{r_k}(\mathbf{M}_k)$. We remark that estimating K is a challenging question even under GMM. Hence, throughout this paper, it is assumed that K is provided beforehand.

[53] introduced a tensor Gaussian mixture model without specifically imposing low-rank structures on the center matrices. A similar tensor normal mixture model without low-rank assumptions is proposed by [49]. Our LrMM can be viewed as a generalization of mixture multi-layer SBM proposed by [31] and as an extension of the symmetric two-component case introduced by [47]. Mixture of low-rank matrix normal models have also appeared in [19] for image analysis.

Since our goal of current paper is to investigate the fundamental limits of clustering matrix-valued observations, hereafter, we view the latent labels $s_i^*, i \in [n]$ as a *fixed* realization sampled from the mixture distribution (2). Then the matrix-valued observations can be written in the following form:

$$\mathbf{X}_i = \mathbf{M}_{s_i^*} + \mathbf{E}_i, \quad i \in [n] \tag{4}$$

Denote $\mathbf{s}^* = (s_1^*, \cdots, s_n^*)$ the collection of true latent labels, known as the *cluster membership vector*. The size of each cluster is given by $n_k^* := \sum_{i=1}^n \mathbb{I}(s_i^* = k), \forall k \in [K]$. With mild conditions under LrMM, Chernoff bound [11] guarantees $n_k^* \asymp n\pi_k$ with high probability.

Given an estimated cluster membership vector $\widehat{\mathbf{s}} := (\widehat{s}_1, \cdots, \widehat{s}_n) \in [K]^n$, its clustering error is measured by the *Hamming distance* defined by

$$h_c(\widehat{\mathbf{s}}, \mathbf{s}^*) = \min_{\pi\text{: permutation of } [K]} \sum_{i=1}^n \mathbb{I}(\pi(\widehat{s}_i) \neq s_i^*) \tag{5}$$

2.3. Low-rank Lloyd's algorithm

Lloyd's algorithm [41] or K-means algorithm is perhaps, conceptually and implementation-wise, the most simple yet effective method for clustering. It is an iterative algorithm, which consists of two main routines at each iteration: 1). provided with an estimated cluster membership vector, the cluster centers are updated by taking the sample average within every estimated cluster; 2). provided with the updated cluster centers, every data point is assigned an updated cluster label according to its distances from the cluster centers. The iterations are terminated once converged. The success of Lloyd's algorithm is highly reliant on a good initial clustering or initial cluster centers. It is proved by [44] and [18] that, if well initialized, Lloyd's algorithm converges fast and achieves minimax optimal clustering error for GMM and community detections under stochastic block model.

The original Lloyd's algorithm updates the cluster centers by taking the vanilla sample average. This approach is sub-optimal under LrMM because the underlying low-rank structure is overlooked. It is well-known that exploiting the low-rankness can further de-noise the estimates. Towards that end, we propose the low-rank Lloyd's algorithm whose details are enumerated in Algorithm 1. Compared with the original Lloyd's algorithm, the low-rank version only modifies the procedure of updating the cluster centers. At the $(t+1)$-th iteration, given the current cluster labels $\widehat{\mathbf{s}}^{(t)}$ and for each k, we calculate the sample average $\bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t)})$ defined as in Algorithm 1, and then update the cluster center by

$$\widehat{\mathbf{M}}_k^{(t+1)} := \widehat{\mathbf{U}}_k^{(t)} \widehat{\mathbf{U}}_k^{(t)\top} \bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t)}) \widehat{\mathbf{V}}_k^{(t)} \widehat{\mathbf{V}}_k^{(t)\top}$$

where $\widehat{\mathbf{U}}_k^{(t)}$ and $\widehat{\mathbf{V}}_k^{(t)}$ are the top-r_k left and right singular vectors of $\bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t)})$, respectively. The update of cluster labels is unchanged compared with the original Lloyd's algorithm.

Conceptually, our low-rank Lloyd's algorithm is a direct adaptation of Lloyd's algorithm to accommodate low-rankness. However, the low-rank update of cluster centers poses fresh and highly non-trivial challenges in studying the convergence behavior of Algorithm 1. The original Lloyd's algorithm simply takes the sample average and thus admits a clean and explicit representation form for the updated centers, which plays a critical role in technical analysis, as in

Algorithm 1 Low-rank Lloyd's Algorithm (lr-Lloyd)

Input: Observations $\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^{d_1 \times d_2}$,initial estimate $\widehat{\mathbf{s}}^{(0)}$.
for $t = 1, \ldots, T$ **do**
for each $k = 1, \cdots, K$: (*update cluster centers*)

$$\widehat{\mathbf{M}}_k^{(t)} \leftarrow \text{best rank-}r_k \text{ approximation of } \bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t-1)}) := \frac{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = k\right) \mathbf{X}_i}{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = k\right)} \tag{6}$$

for each $i = 1, \cdots, n$: (*update cluster labels*)

$$\widehat{s}_i^{(t)} \leftarrow \underset{k \in [K]}{\arg\min} \|\mathbf{X}_i - \widehat{\mathbf{M}}_k^{(t)}\|_{\mathrm{F}}^2$$

end for
Output: $\widehat{\mathbf{s}} := \widehat{\mathbf{s}}^{(T)}$

[18]. In sharp contrast, the required SVD in Algorithm 1 involves intricate and non-linear operations on the matrix-valued observations, and there is surely no clean and explicit representation form for $\widehat{\mathbf{M}}_k^{(t)}$. More advanced tools are in need for our purpose, as shall be explained in Section 3.

2.4. Tensor-based spectral initialization

The success of Algorithm 1 crucially depends on a reliable initial clustering. A naive approach is to vectorize the matrix observations, concatenate them into a new matrix of size $n \times (d_1 d_2)$, then borrow the classic spectral clustering method as in [43] and [65]. Unfortunately, the naive approach turns out to be sub-optimal for ignoring the planted low-dimensional structure in the row space.

Our proposed initial clustering is based on tensor decomposition. Towards that end, we construct a third-order data tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times n}$ by stacking the matrix-valued observations slice by slice, i.e., its i-th slice[2] $\boldsymbol{\mathcal{X}}(:,:,i) = \mathbf{X}_i$. The noise tensor $\boldsymbol{\mathcal{E}}$ is defined in the same fashion. The *signal tensor* $\boldsymbol{\mathcal{M}}$ is constructed such that $\boldsymbol{\mathcal{M}}(:,:,i) = \mathbf{M}_{s_i^*}$. The tensor form of LrMM (4) is

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{M}} + \boldsymbol{\mathcal{E}} \tag{7}$$

Interestingly, eq. (7) coincides with the famous tensor SVD or PCA model [64, 63, 40]. Let $\mathring{r} := \sum_{k=1}^K r_k$. Indeed, the signal tensor $\boldsymbol{\mathcal{M}}$ admits the following low-rank decomposition

$$\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \tag{8}$$

where the $\mathring{r} \times \mathring{r} \times K$ core tensor $\boldsymbol{\mathcal{S}}$ is constructed as $\boldsymbol{\mathcal{S}}(:,:,k) := \mathrm{diag}(\mathbf{0}_{r_1}, \cdots, \mathbf{0}_{r_{k-1}}, \boldsymbol{\Sigma}_{r_k}, \mathbf{0}_{r_{k+1}}, \cdots, \mathbf{0}_{r_K})$, $\mathbf{U} = (\mathbf{U}_1, \cdots, \mathbf{U}_K) \in \mathbb{R}^{d_1 \times \mathring{r}}$, $\mathbf{V} = (\mathbf{V}_1, \cdots, \mathbf{V}_K) \in \mathbb{R}^{d_2 \times \mathring{r}}$, and $\mathbf{W} = (\mathbf{e}_{s_1^*}, \cdots, \mathbf{e}_{s_n^*})^\top \in \{0,1\}^{n \times K}$. Here $\mathbf{e}_k$ denotes the k-th canonical basis vector in Euclidean space whose dimension might vary at different appearances. Clearly, the rows of $\mathbf{W}$

[2]We follow Matlab syntax tradition and denote $\mathcal{X}(:,:,i)$ the sub-tensor by fixing one index.

provide the cluster information and is referred to as the cluster membership matrix. Note that (8) is not necessarily the Tucker decomposition since $\mathbf{U}, \mathbf{V}$ might be rank-deficient, in which case the decomposition in the form (8) is not unique and $\mathbf{U}, \mathbf{V}$ become unrecoverable.

The singular space of $\boldsymbol{\mathcal{M}}$ is uniquely characterized by its Tucker decomposition. To this end, denote $\mathbf{U}^* \in \mathbb{O}_{d_1, r_{\mathbf{U}}}$ and $\mathbf{V}^* \in \mathbb{O}_{d_2, r_{\mathbf{V}}}$ the left singular vectors of $\mathbf{U}$ and $\mathbf{V}$, respectively. Here, $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$ are the ranks of $\mathscr{M}_1(\boldsymbol{\mathcal{M}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{M}})$, respectively. Define $\mathbf{W}^* \in \mathbb{O}_{n,K}$ by normalizing the columns of $\mathbf{W}$. Re-compute the core tensor $\boldsymbol{\mathcal{S}}^* := \boldsymbol{\mathcal{M}} \times_1 \mathbf{U}^{*\top} \times_2 \mathbf{V}^{*\top} \times_3 \mathbf{W}^{*\top}$ that is of size $r_{\mathbf{U}} \times r_{\mathbf{V}} \times K$. Finally, we re-parameterize the signal tensor via its Tucker decomposition

$$\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{S}}^* \times_1 \mathbf{U}^* \times_2 \mathbf{V}^* \times_3 \mathbf{W}^* \tag{9}$$

Here $\mathbf{U}^*, \mathbf{V}^*, \mathbf{W}^*$ are usually called the singular vectors of $\boldsymbol{\mathcal{M}}$. Still, the rows of $\mathbf{W}^*$ tell the cluster information in that $\mathbf{W}^*(i,:) = \mathbf{W}^*(j,:)$ iff $s_i^* = s_j^*$, i.e, i, j belongs to the same cluster. We note that there are interesting special cases concerning the values of $r_{\mathbf{U}}, r_{\mathbf{V}}$. For instance, if $r_{\mathbf{U}} = r_{\mathbf{V}} = r_1$, it implies that all the population center matrices share the same low-dimensional singular space with $\mathbf{M}_1$, which simplifies theoretical investigate of our proposed initialization method. Another special case is $r_{\mathbf{U}} = r_{\mathbf{V}} = \mathring{r}$, namely the singular spaces of all population center matrices are separated to a certain degree. Intuitively, the clustering problem becomes easier. See Section 3.2 for discussions of both cases.

We now present our tensor-based spectral method for initial clustering. Unlike the aforementioned naive spectral method, ours is specifically designed to exploit the low-rank structure of $\boldsymbol{\mathcal{M}}$ in the 1st and 2nd dimension. Without loss of generality, we treat $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$ as known here and shall discuss ways to estimate them in Section 8. Our method consists of three crucial steps with details in Algorithm 2. Step 1 aims to estimate the singular vectors $\mathbf{U}^*$ and $\mathbf{V}^*$. Here, higher order SVD (HOSVD) is obtained by applying SVD to the matricizations $\mathscr{M}_1(\boldsymbol{\mathcal{M}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{M}})$. See, for instance, [13] and [63]. The estimated singular vectors are used for denoising in Step 2 by projecting the noise into a low-dimensional space. Step 3 applies the classical K-means clustering [43, 65] to the denoised observations. Note that solving K-means is generally NP-hard [48], but there exist fast algorithms [35] achieving an approximate solution.

Algorithm 2 improves the naive spectral clustering whenever $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are reliable estimates of their population counterparts. This suggests that a certain signal strength condition on $\mathscr{M}_1(\boldsymbol{\mathcal{S}}^*)$ and $\mathscr{M}_2(\boldsymbol{\mathcal{S}}^*)$ is necessary. We remark that the higher order orthogonal iteration (HOOI, [64]) algorithm for tensor decomposition is not suitable for our purpose since it requires a lower bound on $\sigma_{\min}\big(\mathscr{M}_3(\boldsymbol{\mathcal{S}}^*)\big)$, which is too restrictive under LrMM. See Section 3.2 for more explanations.

3. Minimax Optimal Clustering Error Rate of LrMM

In this section, we establish the convergence performance of low-rank Lloyd's algorithm, validate our tensor-based spectral initialization, and derive the min-

Algorithm 2 Tensor-based Spectral Initialization (TS-Init)

Input: Observations $\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^{d_1 \times d_2}$ or a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times n}$ by concatenating the matrix observations slice by slice.

1. Obtain the estimated factor matrices $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ by applying HOSVD to the tensor $\boldsymbol{\mathcal{X}}$ in mode-1 and mode-2 with rank $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$, respectively.
2. Project $\boldsymbol{\mathcal{X}}$ onto the column space of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ by
$$\widehat{\boldsymbol{\mathcal{G}}} := \boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \in \mathbb{R}^{d_1 \times d_2 \times n}$$
3. Apply k-means on rows of $\widehat{\mathbf{G}} := \mathscr{M}_3(\widehat{\boldsymbol{\mathcal{G}}}) \in \mathbb{R}^{n \times d_1 d_2}$ to obtain initializer for $\mathbf{s}^*$, i.e.
$$(\widehat{\mathbf{s}}^{(0)}, \{\widehat{\mathbf{M}}_k^{(0)}\}_{k=1}^K) := \mathop{\arg\min}_{\mathbf{s}\in[K]^n, \{\mathbf{M}_k\}_{k=1}^K, \mathbf{M}_k \in \mathbb{R}^{d_1 \times d_2}, \forall k} \sum_{i=1}^n \left\| [\widehat{\mathbf{G}}]_{i\cdot} - vec(\mathbf{M}_{s_i}) \right\|^2$$

Output: $\widehat{\mathbf{s}}^{(0)}$

imax optimal clustering error rate for LrMM. The hardness of clustering under LrMM is determined primary by two quantities:

$$\textit{Separation strength } \Delta := \min_{a\neq b, a,b\in[K]} \|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}}$$
$$\textit{Signal strength } \lambda := \min_{k\in[K]} \sigma_{\min}(\mathbf{M}_k)$$

The separation strength is a generalization of the minimum ℓ_2 distance between different population centers under GMM [44, 10, 18], which characterizes the intrinsic difficult in clustering the observations. In fact, the minimax optimal error rate, i.e, the best achievable clustering accuracy, is exclusively decided by Δ. On the other hand, the signal strength determines whether the population centers or their singular spaces are estimable, only in which case the low-rank structure can be beneficial. Actually, λ determines the computational and statistical limit under LrMM.

3.1. Iterative convergence of low-rank Lloyd's algorithm

The performance of Lloyd's algorithm also relies on the minimal cluster size [44]. To this end, define $\alpha := \min_{k\in[K]} n_k^* \cdot (n/K)^{-1}$, where recall that $n_k^* := |\{i \in [n] : s_i^* = k\}|$ is the size of k-th cluster. The cluster sizes are said to be *balanced* if $\alpha \asymp 1$. The hamming distance $h_c(\widehat{\mathbf{s}}, \mathbf{s}^*)$ is defined as in eq. (5). Without loss of generality, we assume $r := r_1$ is the largest amongst $\{r_k : k \in [K]\}$ and $d := d_1 \geq d_2$.

Due to technical reasons, the following quantities are involved:

$$\gamma := \frac{\max_{a\neq b, a,b\in[K]} \|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}}}{\Delta} \quad \text{and} \quad \kappa_0 := \frac{\max_{k\in[K]} \|\mathbf{M}_k\|}{\lambda}$$

Here κ_0 can be viewed as the maximum condition number of all population center matrices. They usually do not appear in the literature of GMM, but are

of unique importance under LrMM. These two quantities play a critical role in connecting the accuracy of updated center matrix $\widehat{\mathbf{M}}_k^{(t)}$ to the current clustering accuracy $h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$. Since $\widehat{\mathbf{M}}_k^{(t)}$ stems from the SVD of $\bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t-1)})$, whose accuracy is characterized by the strength of signal $\mathbf{M}_k$ and size of perturbation $\bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t-1)}) - \mathbf{M}_k$. Besides random noise, the latter term, roughly, consists of $(n_a^*)^{-1}h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)\big(\mathbf{M}_{k'\neq k} - \mathbf{M}_k\big)$, whose operator norm can be controlled by $O\big((n_a^*)^{-1}h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \cdot \min\{\kappa_0\lambda, \gamma\Delta\}\big)$. Quantities like γ and κ_0 are, *perhaps*, the unavoidable price to be paid for taking advantage of low-rankness.

The following theorem presents the convergence performance of low-rank Lloyd's algorithm (Algorithm 1). Due to the local nature of Lloyd's algorithm, its success highly relies on a good initialization. Theorem 3.1 assumes the initial clustering is consistent, i.e., initial clustering error approaches zero asymptotically as $n \to \infty$. Under suitable conditions of separation strength and signal strength, the output of Algorithm 1 attains an exponential-type error rate. The constant factor $1/8$ in the exponential rate exactly matches the minimax lower bound in Theorem 3.4. Notice that our result is non-asymptotic, and all asymptotic conditions in Theorem 3.1 are to guarantee the sharp constant $1/8$ in eq. (13).

Theorem 3.1. *Suppose $d \geq C_0 \log K$ for some absolute constant $C_0 > 0$. Assume that $\alpha n(\kappa_0^2 r^2 K)^{-1} \to \infty$ as $n \to \infty$ and*

(i) initial clustering error:

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = o\left(\frac{\alpha}{(\kappa_0 \vee \gamma^2)\gamma^2 K}\right) \tag{10}$$

(ii) separation strength:

$$\frac{\Delta^2}{\alpha^{-1}K^2 r\left(\frac{dr}{n} + 1\right)} \to \infty \tag{11}$$

(iii) signal strength:

$$\lambda \geq C_1\left[\alpha^{-1/2}K^{1/2}\sqrt{\frac{d}{n}} + \alpha^{-1}K\sqrt{\frac{h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*)}{n}} + \alpha^{-1/4}\frac{d^{1/2}}{n^{1/4}}\right] \tag{12}$$

for some absolute constant $C_1 > 0$.

Let $\widehat{\mathbf{s}}^{(t)}$ be the cluster labels at t-th iteration generated by Algorithm 1. Then, for all $t \geq 1$, we have

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8}\right) + \frac{1}{2^t} \tag{13}$$

with probability at least $1 - \exp(-\Delta) - \exp(-c_1(d \wedge n))$ with some absolute constant $c_1 > 0$.

By Theorem 3.1, after at most $O\big(\min\{\Delta^2, \log n\}\big)$ iterations, our low-rank Lloyd's algorithm achieves the minimax optimal clustering error rate $\exp(-\Delta^2/8)$.

Akin to GMM [44, 42, 18, 65], the best achievable clustering accuracy is exclusively decided by the separation strength Δ. We note that the second term $\big(h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*)/n\big)^{1/2}$ on RHS of (12) can be simply replaced by 1. The present form is to underscore the affect of initialization to the required signal strength.

Blessing of low-rankness and comparison with GMM. If low-rankness is ignored so that LrMM is treated as GMM, the exponential-type error rate is established only in the regime of separation strength $\Delta^2 \gg 1 + (d_1 d_2/n)^{1/2}$ [18, 65]. In contrast, our condition (11) only requires $\Delta^2 \gg 1 + (rd_1/n)^{1/2}$. However, we need an additional signal strength condition that might be necessary, as discussed below.

Discussions on signal strength. The signal strength condition is typically required in low-rank models [64, 51, 38, 60, 46]. The three terms on RHS of (12) essentially depict the statistical and computational difficult of clustering under LrMM. Without loss of generality, consider the case $\alpha \asymp 1$ and $K = 2$. The denoising step for updating cluster centers in Algorithm 1 requires a sufficiently accurate estimate of $\mathbf{M}_1$. Even if the true labels are revealed, , i.e., $\widehat{\mathbf{s}}^{(t-1)} = \mathbf{s}^*$, the sample average of i.i.d. noise matrices has an operator norm, with high probability, at the order $(d/n)^{1/2}$. It suggests that λ is at least greater than $(d/n)^{1/2}$. On the other hand, if there are $n \cdot h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ mis-clustered observations, bounding the sample average of noise matrices needs more delicate treatments since these matrices are no longer independent. Through a careful investigation, we derive an operator norm bound at the order $\big(h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)/n\big)^{1/2}$ where the second term on RHS of (12) emerges. This condition weakens if a better initial clustering is provided. Hence the first two terms on RHS of (12) reflects the statistical difficulty under LrMM. On the other hand, the third on RHS of (12) that can be much stronger than the first two terms when $n \le d^2$, reflects the computational difficulty under LrMM. Previously, [46] provides evidence showing that no polynomial time can consistently *estimate* the population centers even in the symmetric two-component LrMM if the signal strength $\lambda = o(d^{1/2} n^{-1/4})$. In Section 4, evidences are provided showing that the same phenomenon exists for clustering, that is, if $\lambda = o(d^{1/2} n^{-1/4})$, consistent clustering is impossible by any polynomial time algorithms even when the separation strength Δ is much stronger than the minimal condition (11).

Discussions on separation strength. The separation strength condition is also typical in the literature of clustering problems [54, 43]. To see why our condition (11) is minimal, without loss of generality, consider the case $\alpha \asymp 1$ and $K = 2$. Moreover, assume the singular vectors $\mathbf{U}_1 = \mathbf{U}_2$ and $\mathbf{V}_1 = \mathbf{V}_2$, and they are already known. One can multiply each observation by $\mathbf{U}_1^\top$ from left and by $\mathbf{V}_1$ from right, which reduces LrMM to GMM in the dimension r^2. Literature of GMM [18, 43, 65] all impose a separation strength condition $\Delta \gg 1$. This certifies the constant 1 in eq. (11). To understand the term $(rd/n)^{1/2}$, consider that the true labels of first $n-1$ observations are revealed to us and our goal is to estimate the label of the n-th sample $\mathbf{X}_n$. A natural way is to first estimate the population centers utilizing the given labels $\mathbf{s}_1^*, \cdots, \mathbf{s}_{n-1}^*$, denoted by $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$, respectively. Literature of matrix denoising [7, 60, 20] tells that the minimax

optimal estimation error is at the order $\|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_{\mathrm{F}} \asymp \|\widehat{\mathbf{M}}_2-\mathbf{M}_2\| \asymp (rd/n)^{1/2}$. Thus $\Delta \gg (rd/n)^{1/2}$ is necessary for consistently distinguishing the two clusters. The above rationale suggests that our separation strength condition (11) might be minimal up to the order of n, if only the exponential-type error rate is sought.

We explained a gap concerning the separation strength in existing literature of GMM. Under GMM with dimension $d^* = d_1 d_2$ and $n \le d^*$, the exponential-type rate [18, 65] is established in the regime $\Delta \gg (d^*/n)^{1/2}$, whereas exact clustering results [50, 10] are attained in the regime $\Delta \gtrsim (d^* n^{-1} \log n)^{1/4}$. This leaves a natural question under LrMM: is the separation strength condition (11) is relaxable to the scale $n^{-1/4}$? Unfortunately, answering this question is perhaps more challenging than that under GMM. We note that [50] and [10] achieve the $O(n^{-1/4})$ barrier by focusing entirely on clustering and by circumventing the estimation of population centers. Nonetheless, under LrMM, exploiting the low-rank structure demands estimating the population center matrices. We suspect, together with the aforementioned special examples, that condition (11) might not be improvable in terms of the order of n. Anyhow, It's unclear whether one can obtain a sharper characterization of Δ under LrMM using other methods like SDP. Further investigation in this respect is out of the scope of current paper.

3.2. Guaranteed initialization

Besides the signal strength and separation strength conditions, Theorem 3.1 requires a consistent initial clustering. We now demonstrate the validity of tensor-based Algorithm 2. Observe that denoising by spectral projection (Step 2 of Algorithm 2) is only beneficial if $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are properly aligned with $\mathbf{U}^*$ and $\mathbf{V}^*$, respectively. For that purpose, the signal strengths of $\mathscr{M}_1(\boldsymbol{\mathcal{M}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{M}})$ need to be sufficiently strong, which can be characterized by their condition numbers defined by

$$\kappa_1 := \frac{\|\mathscr{M}_1(\boldsymbol{\mathcal{M}})\|}{\sigma_{\min}\big(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\big)} \quad \text{and} \quad \kappa_2 := \frac{\|\mathscr{M}_2(\boldsymbol{\mathcal{M}})\|}{\sigma_{\min}\big(\mathscr{M}_2(\boldsymbol{\mathcal{M}})\big)}$$

Recall that κ_0 tells whether *individual* population center matrices are well-conditioned. Here κ_1 (κ_2, resp.) measures the goodness of alignment among the column (row, resp.) spaces of *all* population center matrices. However, the exact relation between κ_1 and the column spaces $\{\mathrm{ColSpan}(\mathbf{U}_k^*)\}_{k=1}^K$ can be intricate. The following lemma unfolds two special cases. Recall that $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$ are the ranks of $\mathbf{U} = (\mathbf{U}_1, \cdots, \mathbf{U}_K)$ and $\mathbf{V} = (\mathbf{V}_1, \cdots, \mathbf{V}_K)$, respectively, and $\mathring{r} = \sum_{k=1}^K r_k$. Denote $\kappa(\mathbf{U})$ and $\kappa(\mathbf{V})$ the condition numbers of $\mathbf{U}$ and $\mathbf{V}$, respectively.

Lemma 3.2. *Let $\boldsymbol{\mathcal{M}}$ admits low-rank decomposition (8). We have*

$$\begin{aligned}\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}) &= \mathbf{U} \cdot \mathrm{diag}\big(\{n_k^* \boldsymbol{\Sigma}_k^2\}_{k=1}^K\big) \cdot \mathbf{U}^\top \\ \mathscr{M}_2(\boldsymbol{\mathcal{M}})\mathscr{M}_2^\top(\boldsymbol{\mathcal{M}}) &= \mathbf{V} \cdot \mathrm{diag}\big(\{n_k^* \boldsymbol{\Sigma}_k^2\}_{k=1}^K\big) \cdot \mathbf{V}^\top\end{aligned}$$

and $\kappa_1 \le \kappa_0\kappa(\mathbf{U}) \cdot (n^*_{max}/n^*_{min})^{1/2}$ *and* $\kappa_2 \le \kappa_0\kappa(\mathbf{V}) \cdot (n^*_{max}/n^*_{min})^{1/2}$ *where* $n^*_{min} := \min_k n^*_k$ *and* $n^*_{max} := \max_k n^*_k$. *If* $r_{\mathbf{U}} = r_{\mathbf{V}} = r_1$, *i.e., all the population center matrices share the same singular space with* $\mathbf{M}_1$, *we have* $\max\{\kappa_1, \kappa_2\} \le \kappa_0 \cdot (K^2/\alpha)^{1/2}$; *if* $r_{\mathbf{U}} = r_{\mathbf{V}} = \mathring{r}$ *and* $\mathbf{M}_k$ *has mutually orthogonal singular space, we have* $\max\{\kappa_1, \kappa_2\} \le \kappa_0 \cdot (K/\alpha)^{1/2}$.

According to Lemma 3.2, the unfolded matrices $\mathscr{M}_1(\boldsymbol{\mathcal{M}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{M}})$ are well-conditioned if $\mathbf{U}$ and $\mathbf{V}$ are well-conditioned. Interestingly, this implies that our tensor-based spectral initialization becomes more efficient when the population center matrices $\mathbf{M}_k$'s have *either perfectly aligned singular spaces or nearly orthogonal singular spaces.*

Theorem 3.3. *Let* $\widehat{\mathbf{s}}^{(0)}$ *be the initial clustering output by Algorithm 2. There exists some absolute constant* $c, C_1, C_2 > 0$ *such that if*

$$\lambda \ge C_1 \max\{\kappa_1, \kappa_2\} rK \cdot \frac{d^{1/2}}{n^{1/4}} \quad \text{and} \quad \Delta^2 \ge C_2 \alpha^{-1} K^2 \left(\frac{dKr}{n} + 1 \right),$$

we get, with probability at least $1 - \exp(-c(n \wedge d))$, *that*

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \le \frac{K}{2\Delta^2} \left(\frac{dKr}{n} + 1 \right)$$

By Theorem 3.3, if $\kappa_1, \kappa_2 = O(1)$, the signal strength condition required by tensor-based spectral initialization method is comparable to but slightly weaker than that needed in low-rank Lloyd's algorithm. Both algorithms require a signal strength lower bound by $d^{1/2}n^{-1/4}$, but low-rank Lloyd's algorithm involves an additional lower bound related to the initial accuracy $n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*)$. Theorem 3.3 also suggests that Algorithm 2 delivers a consistent clustering if the separation strength $\Delta^2 \gg K(1 + rdK/n)$.

Finally, by combining Theorem 3.3 and Theorem 3.1, the successes of Algorithm 1 and Algorithm 2 require signal strength and separation strength conditions

$$\lambda \ge C_1 \left[\alpha^{-1/2} K^{1/2} \left(\sqrt{\frac{d}{n}} + 1 \right) + \max\{(\kappa_1 \vee \kappa_2) rK, \alpha^{-1/4}\} \frac{d^{1/2}}{n^{1/4}} \right]$$

and

$$\frac{\Delta^2}{\alpha^{-1}(\kappa_0 \vee \gamma^2)\gamma^2 K^2 \left(\frac{dKr}{n} + 1 \right)} \to \infty$$

Comparison with HOOI [64] and the condition number of $\mathscr{M}_3(\boldsymbol{\mathcal{M}})$. Algorithm 2 looks similar to HOOI [64], which uses HOSVD for *mode-wise* spectral initialization and applies power iterations to further improve the estimates of singular spaces. The *mode-wise* HOSVD and subsequent power iterations both require a lower bound on $\sigma_{\min}\big(\mathscr{M}_k(\boldsymbol{\mathcal{M}})\big), k = 1, 2, 3$. While our Theorem 3.3 also requires a lower bound (in the form of κ_1, κ_2) on $\sigma_{\min}\big(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\big)$ and $\sigma_{\min}\big(\mathscr{M}_2(\boldsymbol{\mathcal{M}})\big)$, we emphasize that a similar lower bound on $\sigma_{\min}\big(\mathscr{M}_3(\boldsymbol{\mathcal{M}})\big)$ is too strong and trivialize the whole problem. To see this, just notice via definition that $\Delta \ge \sigma_{\min}\big(\mathscr{M}_3(\boldsymbol{\mathcal{M}})\big)/2$.

3.3. Minimax lower bound

Theorem 3.1 has shown that the low-rank Lloyd's algorithm achieves the asymptotical clustering error rate $\exp(-\Delta^2/8)$. In this section, a matching minimax lower bound is derived showing that the aforesaid rate is indeed optimal in the minimax sense. A lower bound under GMM has been established by [44]. We follow the arguments in [17] to establish the minimax lower bound for LrMM. Observe that the error rate only depends on the separation strength Δ implying that the dimension d_1, d_2 and ranks r_k's play a less important role here.

Define the following parameter space for the population center matrices and arrangements of latent labels:

$$\Omega_\Delta \equiv \Omega(\Delta, d_1, d_2, n, K, \alpha) := \Big\{(\{\mathbf{M}_k\}_{k=1}^K, \mathbf{s}) : \ \mathbf{M}_k \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(\mathbf{M}_k) = r_k, \mathbf{s} \in [K]^n, \\ \min_{k\in[K]} |\{i \in [n] : s_i = k\}| \geq \alpha n/K, \min_{a\neq b} \|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}} \geq \Delta\Big\}$$

For notation simplicity, we omit its dependence on the ranks r_k's.

Theorem 3.4. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ satisfy LrMM (3) with $(\{\mathbf{M}_k\}_{k=1}^K, \mathbf{s}^*) \in \Omega_\Delta$. Suppose $\{\mathbf{E}_i\}_{i=1}^n$ has i.i.d $\mathcal{N}(0, \sigma^2)$ entries. If $\Delta^2/\left(\sigma^2 \log(K/\alpha)\right) \to \infty$ as $n \to \infty$, we have*

$$\inf_{\widehat{\mathbf{s}}} \sup_{(\{\mathbf{M}_k\}_{k=1}^n, \mathbf{s}^*)\in\Omega_\Delta} \mathbb{E}\frac{h_c(\widehat{\mathbf{s}}, \mathbf{s}^*)}{n} \geq \exp\left(-(1+o(1))\frac{\Delta^2}{8\sigma^2}\right)$$

where $\inf_{\widehat{\mathbf{s}}}$ is taken over all clustering algorithms.

Compared to Theorem 3.1 and Theorem 3.3, the minimax lower bound is established only requiring a separation strength $\Delta^2 \gg 1$ assuming $K/\alpha = O(1)$. Theorem 3.4 holds for any signal strength and the infimum is taking over all possible clustering algorithms without considering their computational feasibility. Here, an algorithm is said *computationally feasible* if it is computable within a polynomial time complexity in terms of n and d_1, d_2.

4. Computational Barriers

We now turn to the computational hardness of LrMM. For simplicity, we set $\alpha, K, r \asymp 1$ throughout this section. Note that $\Delta \gg 1 + (d/n)^{1/2}$ already implies *nearly all* [3] population center matrices must own a signal strength $\lambda \gtrsim 1 + (d/n)^{1/2}$. However, our signal strength condition (12) requires an additional lower bound $\lambda \gtrsim d^{1/2}n^{-1/4}$. The purpose of this section is to provide evidences on its necessity to guarantee computationally feasible clustering algorithms. Our evidence is built on the *low-degree likelihood ratio* framework for hypothesis

[3] More precisely, at most one population center matrix can violate the signal strength condition. See Section 5 for more details.

testing proposed by [36, 27], which has delivered convincing evidences justifying the computational hardness under sparse GMM [42] and for sparse PCA [14].

Suppose that, given i.i.d. observations $\mathbf{X}_1, \cdots, \mathbf{X}_n$, one is interested in the computational and statistical limit in distinguishing two hypothesis $\mathbb{Q}_n$ and $\mathbb{P}_n$, i.e,

$$H_0^{(n)} : \mathbf{X}_1 \sim \mathbb{Q}_n \quad \text{versus} \quad H_1^{(n)} : \mathbf{X}_1 \sim \mathbb{P}_n \tag{14}$$

The above two hypotheses are said *statistically indistinguishable* if no test can have both type I and type II error probabilities vanishing asymptotically. The famous Neyman-Pearson lemma tells us that the likelihood ratio test based on $L_n(\boldsymbol{\mathcal{X}}) := d\mathbb{P}_n/d\mathbb{Q}_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ has a preferable power and is uniformly most powerful under some scenarios. A well recognized fact is that $\mathbb{Q}_n$ and $\mathbb{P}_n$ are statistically indistinguishable if the quantity $\|L_n\|^2 := \mathbb{E}_{\mathbb{Q}_n}[L_n(\boldsymbol{\mathcal{X}})^2]$ remains bounded as $n \to \infty$. See [36] for a simple proof.

While the asymptotic magnitude of $\|L_n\|^2$ is informative for understanding the statistical limit of testing (14), it does not directly reflect the computational limit of testing (14). Towards that end, the low-degree likelihood ratio framework seeks a polynomial approximation of $L_n(\boldsymbol{\mathcal{X}})$ and investigates the magnitude of the resultant approximation. More exactly, let $L_n^{\le D}(\boldsymbol{\mathcal{X}})$ be the orthogonal projection of $L_n(\boldsymbol{\mathcal{X}})$ onto the linear space spanned by polynomials $\mathbb{R}^{d_1 \times d_2 \times n} \mapsto \mathbb{R}$ of degrees at most D. Similarly, define $\|L_n^{\le D}\|^2 := \mathbb{E}_{\mathbb{Q}_n}[L_n^{\le D}(\boldsymbol{\mathcal{X}})^2]$. [36] conjectures that the asymptotic magnitude of $\|L_n^{\le D}\|^2$ reflects the computational hardness of testing the hypothesis (14). More formally, their conjecture, slightly adapted for our purpose, can be written as follows. It has been introduced in [46]. Here, a test $\phi_n(\cdot)$ taking value 1 means rejecting the null hypothesis and takes value 0 if the null hypothesis is not rejected. Thus $\mathbb{E}_{\mathbb{Q}_n}[\phi_n(\boldsymbol{\mathcal{X}})]$ and $\mathbb{E}_{\mathbb{P}_n}[1 - \phi_n(\boldsymbol{\mathcal{X}})]$ stands for type-I and type-II error, respectively.

Conjecture 4.1 ([46]). *If there exists $\epsilon > 0$ and $D = D_n \ge (\log nd)^{1+\epsilon}$ for which $\left\|L_n^{\le D}\right\| = 1 + o(1)$ as $n \to \infty$, then there is no polynomial-time test $\phi_n : \mathbb{R}^{d_1 \times d_2 \times n} \mapsto \{0, 1\}$ such that the sum of type-I error and type-II error probabilities*

$$\mathbb{E}_{\mathbb{Q}_n}[\phi_n(\boldsymbol{\mathcal{X}})] + \mathbb{E}_{\mathbb{P}_n}[1 - \phi_n(\boldsymbol{\mathcal{X}})] \to 0 \quad \textit{as} \quad n \to \infty$$

Based on this conjecture, [36] reproduces the sharp phase transitions for the spiked Wigner matrix model and the widely-believed statistical-to-computational gap in tensor PCA, and [46] develops a computational hardness theory for estimating the population low-rank matrices under LrMM.

Note that a specific hypothesis $\mathbb{P}_n$ is necessary to apply Conjecture 4.1 and investigate the computational barriers in clustering for LrMM. Towards that end, we consider a symmetric two-component LrMM as in [46]. It is a special case of model (3) with $K = 2$, $r_1 = r_2 = 1$, $\mathbf{M}_1 = \lambda \mathbf{u}\mathbf{v}^\top$ and $\mathbf{M}_2 = -\mathbf{M}_1 = -\lambda \mathbf{u}\mathbf{v}^\top$. Here $\mathbf{u} \in \mathbb{R}^{d_1}$ and $\mathbf{v} \in \mathbb{R}^{d_2}$ have unit norms. The signal strength is $\lambda > 0$ and separation strength is 2λ, i.e., the two quantities are at the same order. Then the observations can be re-written as

$$\mathbf{X}_i = s_i^*(\lambda \mathbf{u}\mathbf{v}^\top) + \mathbf{E}_i, \quad \forall i = 1, \cdots, n, \tag{15}$$

where $s_i^* = 1$ if $\mathbf{X}_i$ is sampled from $\mathcal{N}(\mathbf{M}_1, \mathbf{I}_{d_1} \otimes \mathbf{I}_{d_2})$ and $s_i^* = -1$ if $\mathbf{X}_i$ is sampled from $\mathcal{N}(\mathbf{M}_2, \mathbf{I}_{d_1} \otimes \mathbf{I}_{d_2})$. Note that the rank-one model (15) is no more difficult than the general K-component case but it suffices for our purpose. The null hypothesis $\mathbb{Q}_n$ corresponds to the case $\lambda = 0$, i.e., all observations are pure noise. Clearly, the difficulty level of distinguishing $\mathbb{Q}_n$ and $\mathbb{P}_n$ is characterized by signal strength λ in eq. (15). Conjecture 4.1 requires the calculation of $\|L_n^{\leq D}\|^2$, which is extremely difficult for generally fixed singular vectors $\mathbf{u}, \mathbf{v}$ and deterministic latent labels $\mathbf{s}^*$. A prior distribution simplifies the calculation. Finally, our null and alternative hypothesis are formally defined as follows.

Definition 4.2 (Null and alternative hypothesis).

- Under $\mathbb{Q}_n$, we observe n matrices $\mathbf{X}_1, \cdots, \mathbf{X}_n$ generated i.i.d. from (15) with $\lambda = 0$. Equivalently, it means that each $\mathbf{X}_i$ has i.i.d. standard normal entries.
- Under $\mathbb{P}_n := \mathbb{P}_n^{\lambda_*}$, we observe n matrices $\mathbf{X}_1, \cdots, \mathbf{X}_n$ generated i.i.d. from (15) with $\lambda = \lambda_*$, and moreover, each coordinate of $\mathbf{u}$ and $\mathbf{v}$ independently uniformly take values from $\{\pm d_1^{-1/2}\}$ and $\{\pm d_2^{-1/2}\}$, respectively, and the entries of $\mathbf{s}^*$ are independent Rademacher random variables, i.e., taking ± 1 with equal probabilities.

Theorem 4.3. *Consider $\mathbb{Q}_n$ and $\mathbb{P}_n$ in Definition 4.2. If $\lambda_* = o(d^{1/2} n^{-1/4})$ as $n \to \infty$, then $\left\|L_n^{\leq D}\right\| = 1 + o(1)$.*

The proof of Theorem 4.3 can be found in [46]. If Conjecture 4.1 is true, Theorem 4.3 implies that $\mathbb{Q}_n$ and $\mathbb{P}_n^{\lambda_*}$ are statistically indistinguishable by polynomial-time algorithms as long as the signal strength $\lambda_* = o(d^{1/2} n^{-1/4})$. We now establish the connection of testing the hypothesis to the clustering problem under two-component symmetric LrMM (15).

For any fixed $\lambda_* > 0$, define the parameter space of interest by

$$\widetilde{\Omega}_{\lambda_*} \equiv \widetilde{\Omega}(\lambda_*, d_1, d_2, n) := \left\{(\mathbf{M}, \mathbf{s}) : \ \mathbf{M} = \lambda \mathbf{u}\mathbf{v}^\top, \mathbf{u} \in \mathbb{R}^{d_1}, \mathbf{v} \in \mathbb{R}^{d_2}, \mathbf{s} \in \{\pm 1\}^n, |\mathbf{1}^\top \mathbf{s}| \leq n/2, \lambda \geq \lambda_*\right\}$$

By Chernoff bound, with probability at least $1 - e^{-c_0 n}$ where $c_0 > 0$ is an absolute constant, the i.i.d. observations $\mathbf{X}_1, \cdots, \mathbf{X}_n$ generated by $\mathbb{P}_n^{\lambda_*}$ satisfy the rank-one LrMM (15) with parameters $(\mathbf{M}, \mathbf{s}) \in \tilde{\Omega}_{\lambda_*}$. The following theorem tells that if consistent clustering is possible for LrMM, so is for distinguishing the hypothesis in Definition 4.2.

Theorem 4.4. *Suppose there exists a clustering algorithm $\widehat{\mathbf{s}}_{comp} : \mathbb{R}^{d_1 \times d_2 \times n} \mapsto \{\pm 1\}^n$ for LrMM (15) with runtime poly(n, d) that is consistent under the sequence of signal strength $\left\{\lambda_*^{(n)}\right\}_{n \geq 1}$ in the sense that there exists a sequence $\{(\delta_n, \zeta_n)\}_{n \geq 1} \to 0$ such that for all large n,*

$$\sup_{(\mathbf{M}, \mathbf{s}^*) \in \widetilde{\Omega}_{\lambda_*^{(n)}}} \mathbb{P}\left(n^{-1} \cdot h_c(\widehat{\mathbf{s}}_{comp}, \mathbf{s}^*) > \delta_n\right) \leq \zeta_n \tag{16}$$

If the signal strength satisfies $\lambda_*^{(n)} \geq C_0 (1 + \epsilon^{-2})^{1/2} d^{1/2} n^{-1/2}$ with some absolute constant $C_0 > 0$ and $\epsilon \in (0, 1)$, then there exists a test $\phi_n : \mathbb{R}^{d_1 \times d_2 \times n} \mapsto \{0, 1\}$

with runtime poly(n,d) that consistently distinguishes $\mathbb{P}_n^{\lambda_*^{(n)}}$ *from* $\mathbb{Q}_n$ *so that*

$$\mathbb{E}_{Q_n}[\phi_n(\boldsymbol{\mathcal{X}})] + \sup_{((1-\epsilon)\mathbf{M},\mathbf{s}^*)\in\widetilde{\Omega}_{\lambda_*^{(n)}}} \mathbb{E}_{(\mathbf{M},\mathbf{s}^*)}[1-\phi_n(\boldsymbol{\mathcal{X}})] \to 0, \quad \text{as } n,d\to\infty.$$

Essentially, Theorem 4.4 only needs a signal strength $\lambda_* \gg d^{1/2}n^{-1/2}$ to successfully reduce a polynomial-time clustering algorithm to a polynomial-time hypothesis test. Based on Conjecture 4.1, a combination of Theorem 4.3 and Theorem 4.4 implies the following result, whose proof is straightfoward and hence omitted.

Corollary 4.5. *Suppose Conjecture 4.1 holds. If the signal strength* $\lambda_*^{(n)} = o(d^{1/2}n^{-1/4})$*, then for any polynomial-time clustering algorithm* $\widehat{\mathbf{s}}_{comp}$*, there exist absolute constants* $\delta,\zeta>0$ *such that*

$$\sup_{(\mathbf{M},\mathbf{s}^*)\in\widetilde{\Omega}_{\lambda_*^{(n)}}} \mathbb{P}\left(n^{-1}\cdot h_c(\widehat{\mathbf{s}}_{comp},\mathbf{s}^*)>\delta\right)\geq\zeta$$

as $n\to\infty$.

It is worth pointing out that even though the signal strength $\lambda_* = o(d^{1/2}n^{-1/4})$, the separation strength $\Delta = 2\lambda_*$ can still be much larger than $d^{1/2}n^{-1/2}$ that is required by Theorem 3.1. This suggests that if signal strength is not strong, consistent clustering by polynomial-time algorithms is still impossible even though the separation strength is very strong.

5. Relaxing the Signal Strength Condition

Our main theorem in Section 3 imposes a strong signal strength condition on *all* the population center matrices, i.e., their smallest non-zero singular value is lower bounded by $O(1+d^{1/2}n^{-1/4})$. While evidences in Section 4 show that this condition might be necessary for the two-component symmetric case if only polynomial-time algorithms are sought, this condition appears flawed in the general asymmetric case. This section aims to relax the signal strength condition in the sense that one population center matrix is allowed to be arbitrarily smaller, e.g., in spectral norm, than $d^{1/2}n^{-1/4}$.

Without loss of generality, we focus on the two-component LrMM, i.e., $K=2$ in model (3), whose population center matrices are denoted by $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively. For narration simplicity, assume $\|\mathbf{M}_1\|_{\mathrm{F}}$ is large so that reliable estimation is possible, and assume $\|\mathbf{M}_2\|_{\mathrm{F}}$ is small so that reliable estimation is impossible. More exactly, the following assumption is imposed.

Assumption 5.1. *There exists a large constant* $C>1$ *such that*

$$\sigma_{r_1}(\mathbf{M}_1)\geq C\left(\kappa_0\alpha^{-1/2}\sqrt{\frac{d}{n}}+\alpha^{-1/2}\frac{d^{1/2}}{n^{1/4}}+\kappa_0^{1/2}\right)$$

and

$$\sigma_1(\mathbf{M}_2) \le C^{-1}\left(\sqrt{\frac{d}{n}} + \kappa_0^{-1}\frac{d^{1/2}}{n^{1/4}} + \alpha^{-1/2}\kappa_0^{-1/2}\right)$$

.

If $\kappa_0, \alpha = O(1)$, Assumption 5.1 can be recasted as $\sigma_{r_1}(\mathbf{M}_1) \ge C(d^{1/2}n^{-1/4} + 1)$ and $\sigma_1(\mathbf{M}_2) \le C^{-1}(d^{1/2}n^{-1/4} + 1)$. Note that Assumption 5.1 puts no lower bound on $\sigma_1(\mathbf{M}_2)$. In the extreme case, $\sigma_1(\mathbf{M}_2)$ is allowed to be zero and consistent estimation of $\mathbf{M}_2$ is unavailable even if the true labels are revealed. Denote the separation distance

$$\Delta := \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}}.$$

Assumption 5.1 already implies that $\Delta \ge (C/2)\left(d^{1/2}n^{-1/4} + 1\right)$ if the ranks r_1, r_2 are both upper bounded by $O(1)$. Intuitively, although clustering shall becomes easier as the constant C in Assumption 5.1 increases, this cannot be deduced by Theorem 3.1 where the signal strength condition (12) fails.

Under Assumption 5.1, it is generally pointless to compute the center matrix $\widehat{\mathbf{M}}_2$ by SVD in Lloyd's algorithm since $\mathbf{M}_2$ cannot be reliably estimated. Moreover, the SVD procedure complicates the subsequent theoretical analysis of Lloyd's algorithm. Similarly, the spectral initialization can be mis-leading if a rank $r_{\mathbf{U}}$ larger than r_1 is adopted. For our purpose, we slightly modify the low-rank Lloyd's algorithm. Basically, only the top-r_1 singular vectors are taken during spectral initialization, i.e., effort is made only for estimating $\mathbf{M}_1$ whose left and right singular vectors are denoted by $\mathbf{U}_1$ and $\mathbf{V}_1$, respectively. Instead of estimating $\mathbf{M}_2$ via SVD, we opt to a trivial estimate by setting $\widehat{\mathbf{M}}_2^{(t)} = \mathbf{0}$. The detailed steps are enumerated in Algorithm 3, whose theoretical performance is guaranteed by Theorem 5.2.

Theorem 5.2. *Let $\widehat{\mathbf{s}}^{(t)}$ be the output at t-th iteration by Algorithm 3. Suppose Assumption 5.1 holds. If $(\kappa_0^2 \vee \kappa_0 r_1)^{-2}\alpha n \to \infty$, $\alpha\kappa_0^{-1}\Delta^2 \to \infty$ and*

$$\sqrt{\frac{r_1}{r_2}} \cdot \frac{\sigma_{r_1}(\mathbf{M}_1)}{\sigma_1(\mathbf{M}_2)} \to \infty, \quad \text{as } n \to \infty$$

, then we have

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \le \exp\left(-\big(1 - o(1)\big)\frac{\Delta^2}{8}\right) + \frac{1}{2^t}$$

with probability at least $1 - \exp(-\Delta) - \exp\big(-c_0(d \wedge n)\big)$ for a small but absolute constant $c_0 > 0$.

Finally, we remark that the low-rankness assumption for $\mathbf{M}_2$ in Theorem 5.2 is not essential, which can be dropped by instead requiring $\sqrt{r_1}\sigma_{r_1}(\mathbf{M}_1)/\|\mathbf{M}_2\|_{\mathrm{F}} \to \infty$.

Algorithm 3 Low-rank Lloyd's Algorithm under Relaxed SNR Assumption 5.1 (rlr-Lloyd)

Input: observations: $\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^{d_1 \times d_2}$ where $\mathbf{X}_i = \mathbf{M}_{s_i^*} + \mathbf{E}_i$ and $s_i^* \in \{1, 2\}$; or a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times n}$ by concatenating the matrix observations slice by slice, ranks r_1, r_2.

Spectral initialization:

(1). Obtain the estimated singular vectors $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{V}}_1$ by applying HOSVD to the tensor $\boldsymbol{\mathcal{X}}$ in mode-1 and mode-2 matricizations with rank r_1.

(2). Project $\boldsymbol{\mathcal{X}}$ onto the column space of $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{V}}_1$ by $\widehat{\boldsymbol{\mathcal{G}}} := \boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{V}}_1 \widehat{\mathbf{V}}_1^\top$

(3). Apply K-means on the rows of $\widehat{\mathbf{G}} := \mathscr{M}_3(\widehat{\boldsymbol{\mathcal{G}}}) \in \mathbb{R}^{n \times d_1 d_2}$ and obtain the initial clustering by

$$(\widehat{\mathbf{s}}^{(0)}, \{\widehat{\mathbf{M}}_1^{(0)}, \widehat{\mathbf{M}}_2^{(0)}\}) := \mathop{\arg\min}_{\mathbf{s} \in [2]^n; \mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d_2}} \sum_{i=1}^n \left\| [\widehat{\mathbf{G}}]_{i\cdot} - vec(\mathbf{M}_{s_i}) \right\|^2$$

for $t = 1, \ldots, T$ **do**

For each $k = 1, 2$:

$$\widehat{\mathbf{M}}_k^{(t)} \leftarrow \text{best rank-}r_k \text{ approximation of } \bar{\mathbf{X}}_k(\widehat{\mathbf{s}}^{(t-1)}) := \frac{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = k\right) \mathbf{X}_i}{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = k\right)}$$

Set $\widehat{\mathbf{M}}_2^{(t)} \leftarrow \mathbf{0}$ if $\sigma_1(\widehat{\mathbf{M}}_2^{(t)}) < \sigma_1(\widehat{\mathbf{M}}_1^{(t)})$; or set $\widehat{\mathbf{M}}_1^{(t)} \leftarrow \widehat{\mathbf{M}}_2^{(t)}$, $\widehat{\mathbf{M}}_2^{(t)} \leftarrow \mathbf{0}$ if $\sigma_1(\widehat{\mathbf{M}}_2^{(t)}) > \sigma_1(\widehat{\mathbf{M}}_1^{(t)})$.

Re-label by setting, for each $i \in [n]$:

$$\widehat{s}_i^{(t)} \leftarrow \mathop{\arg\min}_{k \in [2]} \|\mathbf{X}_i - \widehat{\mathbf{M}}_k^{(t)}\|_{\mathrm{F}}^2$$

end for

Output:

6. Extension to Sub-Gaussian Noise

Theorems in Section 3 rely on the assumption of Gaussian noise, which simplifies our technical proofs. In this section, we establish similar results when the noise has sub-Gaussian tails. More exactly, we assume the following condition.

Assumption 6.1. *(Sub-Gaussian Tail) The noise matrix $\mathbf{E}_i$ has i.i.d. zero-mean entries and unit variance, and for $\forall \mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, the following probability holds*

$$\mathbb{P}(\langle \mathbf{M}, \mathbf{E}_i \rangle \geq t) \leq e^{-t^2/(2\sigma_{sg}^2 \cdot \|\mathbf{M}\|_{\mathrm{F}}^2)}, \quad \forall t > 0,$$

where $\sigma_{sg} > 0$ is the sub-Gaussian constant.

Clearly, Assumption 6.1 implies that each entry of $\mathbf{E}_i$ is sub-Gaussian and has a ψ_2-norm bounded by $O(\sigma_{\mathsf{sg}})$. Recall that the ψ_2-norm of a random variable X is defined by $\|X\|_{\psi_2} := \inf\{u > 0 : \exp(X^2/u^2) \leq 2\}$. Under sub-Gaussian noise, the convergence of low-rank Lloyd's algorithm is guaranteed by Theorem 6.2.

Theorem 6.2. *Suppose Assumption 6.1 holds and $d \geq C_0 \log K$ for some absolute constant $C_0 > 0$. Assume that $\alpha n (K\kappa_0^2 r^2)^{-1} \to \infty$ as $n \to \infty$ and*

(i) initial clustering error:

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = o\left(\frac{\alpha}{(\kappa_0 \vee \gamma^2)\gamma^2 K}\right) \tag{17}$$

(ii) separation strength:

$$\frac{\Delta^2/\sigma_{sg}^2}{\alpha^{-1} K^2 r \left(\frac{dr}{n} + 1\right)} \to \infty \tag{18}$$

(iii) signal strength:

$$\frac{\lambda}{\sigma_{sg}} \geq C_1 \left[\alpha^{-1/2} K^{1/2} \sqrt{\frac{d}{n}} + \alpha^{-1} K \sqrt{\frac{h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*)}{n}} + \alpha^{-1/4} \frac{d^{1/2}}{n^{1/4}}\right] \tag{19}$$

for some absolute constant $C_1 > 0$.

Let $\widehat{\mathbf{s}}^{(t)}$ bet the cluster labels at t-th iteration output by Algorithm 1. Then, for all $t \geq 1$, we have

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma_{sg}^2}\right) + \frac{1}{2^t} \tag{20}$$

with probability at least $1 - \exp(-\Delta) - \exp(-c_1(d \wedge n))$ for some absolute constant $c_1 > 0$.

For our tensor-based initialization method, Algorithm 2 is still valid under sub-Gaussian noise Assumption 6.1. Indeed, one can replace λ with $\lambda\sigma_{\mathsf{sg}}^{-1}$ and Δ with $\Delta\sigma_{\mathsf{sg}}^{-1}$, respectively, in Theorem 3.3, and a similar initial clustering error can be derived.

7. Clustering versus Estimation

[46] investigated the minimax optimal estimation of latent low-rank matrices under two-component symmetric LrMM, which revealed multiple phase transitions and a statistical-to-computational gap. In this section, together with Theorem 3.1 and 3.3, we discuss the differences between estimation and clustering.

7.1. Example where clustering is more challenging

For simplicity, we consider the rank-one symmetric two-component LrMM (15) with $d_1 = d_2 = d$, where the separation strength Δ and signal strength λ coincides up to a constant factor. The minimax rate of estimating $\mathbf{M}$ (up to a sign flip), established in [46], is

$$\inf_{\widehat{\mathbf{M}}} \sup_{(\mathbf{M},\mathbf{s}^*)\in\widetilde{\Omega}_\lambda} \mathbb{E} \min_{\eta=\pm 1} \left\|\widehat{\mathbf{M}} - \eta\mathbf{M}\right\|_{\mathrm{F}} \asymp \min\left\{\frac{1}{\lambda}\sqrt{\frac{d}{n}} + \sqrt{\frac{d}{n}}, \lambda\right\} \tag{21}$$

The above rate is achievable by the computationally NP-hard maximum likelihood estimator with almost no constraint on signal strength and by a computationally fast spectral-aggregation estimator under the regime of strong signal strength $\lambda \gtrsim d^{1/2}n^{-1/4}$. For a fair comparison, we focus on this computationally feasible regime. The phase transitions under this regime can be summarized as in Table 2.

Sample size	Signal strength	Minimax optimal estimation error
$d^2 \lesssim n$	$\frac{\sqrt{d}}{n^{1/4}} \lesssim \lambda \lesssim 1$	$\frac{1}{\lambda}\sqrt{\frac{d}{n}}$
	$\lambda \gtrsim 1$	$\sqrt{\frac{d}{n}}$
$d^2 \gg n$	$\lambda \gtrsim \frac{\sqrt{d}}{n^{1/4}}$	$\sqrt{\frac{d}{n}}$

TABLE 2

Phase transition in minimax optimal estimation for two-component symmetric LrMM under the regime of strong signal strength $\lambda \gtrsim d^{1/2}n^{-1/4}$. See (21) and [46] for more details.

Without loss of generality, we assume the dimension $d \to \infty$ as $n \to \infty$. The case $d^2 \gg n$ is referred to as the high-dimensional setting, and $d^2 \lesssim n$ is called the low-dimensional setting. An estimator $\widehat{\mathbf{M}}$ is said *strongly consistent* if the relative estimation error $\|\widehat{\mathbf{M}}-\mathbf{M}\|_{\mathrm{F}}\|\mathbf{M}\|_{\mathrm{F}}^{-1}$ approaches to zero in expectation as $n \to \infty$. Table 2 tells that strongly consistent estimation $\mathbf{M}$ is *always* achievable as long as the signal strength is greater $d^{1/2}n^{-1/4}$. A particularly interesting regime is $d^{1/2}n^{-1/4} \lesssim \lambda \lesssim 1$. For instance, when $d^2 = o(n)$, $\mathbf{M}$ can still be consistently estimated even when the signal strength $\lambda \to 0$ as $n \to \infty$.

It is certainly not the case for clustering. Besides *consistent clustering* (see definition in Theorem 4.4), we say a clustering algorithm is *weakly efficient* if it can beat a random guess, but the mis-clustering error rate does not vanish as $n \to \infty$. When $d^2 = o(n)$, Theorem 3.4 dictates that even weakly efficient

Sample size	Signal strength	Consistent estimation	Weakly efficient clustering	Consistent clustering
$d^2 \lesssim n$	$\frac{\sqrt{d}}{n^{1/4}} \lesssim \lambda \lesssim 1$	Possible	Impossible	Impossible
	$1 \lesssim \lambda \ll 1$	Possible	Possible	Impossible
	$\lambda \gg 1$	Possible	Possible	Possible
$d^2 \gg n$	$\lambda \gtrsim \frac{\sqrt{d}}{n^{1/4}}$	Possible	Possible	Possible

TABLE 3

The differences of phase transitions in estimation and clustering for two-component symmetric LrMM under the regime of strong signal strength $\lambda \gtrsim d^{1/2}n^{-1/4}$. Here $d^2 \gg n$ is referred to as the high-dimensional setting, and $d^2 \lesssim n$ as the low-dimensional setting.

clustering is impossible, i.e., $\exp(-\lambda^2/2)$ is at least $1/2$, if the signal strength is smaller than some absolute constant $c_0 > 0$. However, the spectral aggregation estimator [46] can still consistently estimate the population center matrix $\mathbf{M}$ in the aforesaid scenario. Moreover, by Theorem 3.1, consistent clustering even requires the signal strength $\lambda \to \infty$, which is much more stringent than that required by (strongly) consistent estimation.

The differences of phase transitions in estimation and clustering are enumerated in Table 3. Basically, strongly consistent estimation is always possible as long as $\lambda \gtrsim d^{1/2}n^{-1/4}$. In contrast, weakly efficient clustering is possible only when $\lambda \gtrsim 1 + d^{1/2}n^{-1/4}$, and consistent clustering is possible only when $\lambda \gtrsim d^{1/2}n^{-1/4}$ and meanwhile $\lambda \gg 1$. Note that the gap between estimation and clustering is present only under the low-dimensional setting $n \gtrsim d^2$. The gap vanishes under the high-dimensional setting $d^2 \gg n$, in which case the signal strength condition $\lambda \gtrsim d^{1/2}n^{-1/4}$ already implies $\lambda \gg 1$.

We collect these facts to convince that, at least for the two-component symmetric LrMM (15), clustering is intrinsically more challenging than estimation. The same phenomenon also arises in GMM. See, e.g., [59].

7.2. Example where estimation is more challenging

While, generally, clustering is recognized as being more challenging than estimation, there are examples where clustering is easier than estimation. Similarly as in Section 5, consider the two-component LrMM with population center matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ so that

$$\sigma_{r_1}(\mathbf{M}_1) \geq C_1\Big(1 + \frac{d^{1/2}}{n^{1/2}} + \frac{d^{1/2}}{n^{1/4}}\Big) \quad \text{and} \quad \sigma_1(\mathbf{M}_2) \leq C_1^{-1} \cdot \frac{d^{1/2}}{n^{1/2}}$$

where $C_1 > 0$ is a large constant and, without loss of generality, we assume $\kappa_0, \alpha, r_1, r_2 = O(1)$. Observe that

$$\sqrt{\frac{r_1}{r_2}} \cdot \frac{\sigma_{r_1}(\mathbf{M}_1)}{\sigma_1(\mathbf{M}_2)} \geq \begin{cases} C_1^2 n^{1/4}, & \text{if } n \leq d^2; \\ C_1^2 (n/d)^{1/2}, & \text{if } n > d^2; \end{cases} \to \infty, \quad \text{as } n \to \infty$$

Moreover,

$$\Delta := \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}} \gtrsim C_1\Big(1 + \frac{d^{1/2}}{n^{1/4}}\Big) \to \infty$$

if the constant $C_1 > 0$ diverges to infinity. Therefore, by Theorem 5.2, if $C_1 \to \infty$, our Algorithm 3 consistently cluster all observations.

However, consistent estimation of the population center matrices is more challenging. Even if all the latent labels are correctly identified, estimation of $\mathbf{M}_2$ is still impossible because of its weak signal strength. Indeed, the low-rank approximation to

$$\bar{\mathbf{X}}_2(\mathbf{s}^*) := \frac{1}{n_2^*} \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = 2\right) \mathbf{X}_i$$

achieves the error rate (in expectation) $O(d^{1/2}n^{-1/2})$ and the relative error rate (in expectation) diverges to infinity as $C_1 \to \infty$. Similarly, the trivial estimate by a zero matrix attains the relative error rate 1 that never vanishes as $n \to \infty$. Consequently, a strongly consistent estimate of $\mathbf{M}_2$ becomes impossible.

8. Discussions

8.1. Estimation of $r_{\mathbf{U}}$, $r_{\mathbf{V}}$, K and r_k's

Our tensor-based spectral initialization method requires an input of ranks $r_{\mathbf{U}}$, $r_{\mathbf{V}}$ and the number of clusters K, which are usually unknown in practice. Under the decomposition (9), they constitute the Tucker ranks of tensor $\boldsymbol{\mathcal{M}}$. Several approaches are available to estimate the Tucker ranks for tensor PCA model. One typical approach [31, 6] is to check the scree plots [8] of $\mathscr{M}_1(\boldsymbol{\mathcal{X}})$, $\mathscr{M}_2(\boldsymbol{\mathcal{X}})$ and $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$, respectively. Under a suitable signal strength condition as in Theorem 3.3, the scree plots of $\mathscr{M}_1(\boldsymbol{\mathcal{X}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{X}})$ shall serve a reliable estimate of $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$, respectively. However, we note that it is statistically more efficient to estimate K by, instead, taking the scree plot of $\mathscr{M}_3(\boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}^\top)$, where $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are obtained in step 1 of Algorithm 2. This additional spectral projection promotes further noise reduction as in Algorithm 2. After obtaining $r_{\mathbf{U}}$, $r_{\mathbf{V}}$ and K, an initial clustering $\widehat{\mathbf{s}}^{(0)}$ can be attained by apply Algorithm 2. Similarly, we then estimate the rank r_k by the scree plot of the sample average of matrix observations whose initial labels are k. It provides a valid estimate as long as the initial clustering is sufficiently good. The aforementioned approach works nicely in real-world data applications. See Section 9 for more details.

8.2. Matrix observation with categorical entries

Oftentimes, the matrix observations consist of categorical entries. For instance, the Malaria parasite gene networks (see Section 9.2.3) have binary entries (Bernoulli distribution); the 4D-scanning transmission electron microscopy [24] produces count-type entries (Poisson distribution). Our algorithms are still applicable and deliver appealing performance on, e.g., Malaria parasite gene networks dataset. Unfortunately, our theory can not directly cover those cases, although the noise are still sub-Gaussian. Without loss of generality, let us consider multi-layer binary networks and assume $\mathbf{X}_i$ has Bernoulli entries. Then the entries of $\mathbf{X}_i$

have an equal variance only when they have the same expectation, reducing the network to a trivial Erdős-Rényi graph. Nevertheless, equal noise variance is crucial to establish Theorem 3.3. Moreover, the techniques for proving Theorem 6.2 are likely sub-optimal since the sub-Gaussian constant σ_{sg} is usually not sharp enough to characterize a Bernoulli random variable. We leave this to future works.

9. Numerical Experiments and Real Data Applications

9.1. Numerical Experiments

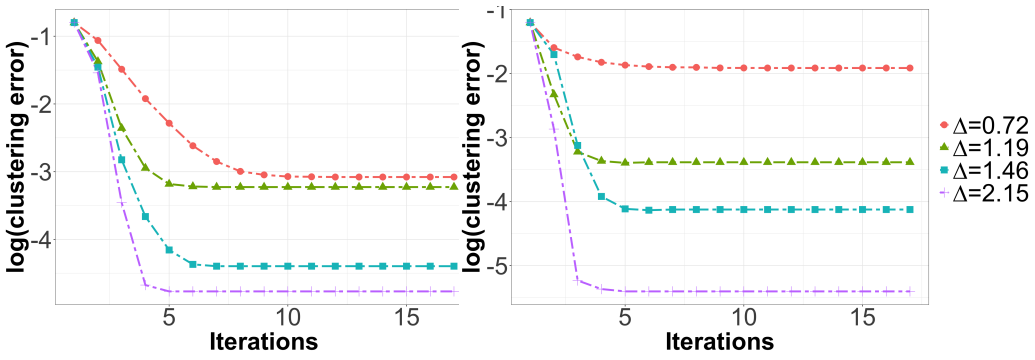
This section presents the empirical performance of lr-Lloyd's algorithm (Algorithm 1) and its relaxed variant under weak SNR (Algorithm 3) referred to as the rlr-Lloyd's algorithm. Specifically, we focus on the algorithmic convergence and final clustering error.

In the first simulation setting **S1**, we fix the dimension $d_1 = d_2 = 50$ and sample size $n = 200$. The latent labels s_i^* are generated i.i.d. from the model (2) with equal mixing probabilities, i.e., $\pi_k = 1/K$. All the presented results in **S1** are based on the average of 30 independent trials. We test the convergence of Algorithm 1 under both Gaussian (**S1-1**) and Bernoulli (**S1-2**) noise.

In **S1-1**, we set $K = 2$, $r_1 = r_2 = 2$ and standard Gaussian noise. The population center matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are generated in the following manner. For each $k = 1, 2$, we independently generate a $d_1 \times d_2$ matrix with i.i.d. standard Gaussian entries and extract its top-2 left and right singular vectors as $\mathbf{U}_k$ and $\mathbf{V}_k$, respectively. The singular values are manually set as $\mathbf{\Sigma}_k = \mathrm{diag}\{1.2\lambda, \lambda\}$ for some fix $\lambda > 0$. Then the population center matrices are constructed as $\mathbf{M}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$. Our experiment tries four levels of signal strength $\lambda \in \{1.9, 2.1, 2.3, 2.5\}$. For each λ, the population center matrices are generated as above and the separation strength is recorded. The corresponding separation strength are $\Delta \in \{4.22, 4.66, 5.11, 5.45\}$. At each level of signal strength, the observations $\{\mathbf{X}_i : i = 1, \cdots, 200\}$ are independently drawn from (4) with the obtained center matrices $\mathbf{M}_1$ and $\mathbf{M}_2$. Here we focus on the convergence behavior of Lloyd's iterations of Algorithm 1, and thus a warm initial clustering $\widehat{\mathbf{s}}^{(0)}$ is provided before hand. The same initial clustering is used for all simulations and the initial clustering error is $n^{-1}h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = 0.45$, i.e., slightly better than a random guess. Convergence of Algorithm 1 under four levels of signal strength (or, correspondingly, separation strength) is displayed in the left plot of Figure 1. The decreasing of log of clustering error is linear in first few iterations, as expected by our Theorem 3.1. The algorithm converges fast and the final clustering error is reflected by the separation strength Δ. It is worth pointing out that Figure 1(a) also shows that Algorithm 1 converges faster when Δ becomes larger. While this cannot be directly concluded from Theorem 3.1, it can be easily verified by checking the proof.

In **S1-2**, we test the effectiveness of Algorithm 1 under non-Gaussian and non-i.i.d. noise. In particular, we consider the mixture multi-layer stochastic

block model (MMSBM) introduced in [31] [4]. We set the number of clusters $K = 3$. For each $k = 1, 2, 3$, the k-th SBM is associated with a connection probability matrix $\mathbf{B}_k \in [0,1]^{K\times K}$ and a membership matrix $\mathbf{Z}_k \in \{0,1\}^{d\times K}$, which are set as $\mathbf{B}_k := \bar{p}_k \cdot \mathbf{I}_K + \bar{p}_k/2 \cdot (\mathbf{1}_K\mathbf{1}_K^\top - \mathbf{I}_K)$ with $\bar{p}_k = \bar{p}\cdot k/K$ and $\mathbf{Z}_k(i,:) = \mathbf{e}_{s_i^*}$, respectively. Thus each SBM has three cluster of nodes and the population center matrices are $\mathbf{M}_k = \mathbf{Z}_k\mathbf{B}_k\mathbf{Z}_k^\top \in [0,1]^{d\times d}$. Conditioned on the latent label $\mathbf{s}_i^*$, the i-th observation $\mathbf{X}_i$ is sampled from $\mathrm{SBM}(\mathbf{Z}_{s_i^*}, \mathbf{B}_{s_i^*})$, namely, $\mathbf{X}_i(j_1, j_2) \sim \mathrm{Bernoulli}(\mathbf{M}_{s_i^*}(j_1, j_2))$ and $\mathbf{X}_i(j_2, j_1) = \mathbf{X}_i(j_1, j_2)$ for $1 \le j_1 < j_2 \le d$. Note that $\mathbf{X}_i$ is symmetric because the network is undirected. We manually set the diagonal entries of $\mathbf{X}_i$ to zeros so that no self-loop is allowed in the observed network. Clearly, the entry-wise variances of $\mathbf{X}_i$ are not necessarily equal. Under the above MMSBM, the signal strength and separation strength are characterized by sparsity level $\bar{p}$. Four sparsity levels $\bar{p} \in \{0.05, 0.08, 0.10, 0.15\}$ are studied so that the corresponding separation strength are $\Delta \in \{0.75, 1.19, 1.46, 2.15\}$. Similarly, a fixed good initial clustering $\widehat{\mathbf{s}}^{(0)}$ is used for all simulations and the initial clustering error is $n^{-1}h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = 0.3$. Convergence behavior of Algorithm 1 is displayed in the right plot of Figure 1. Still, Lloyd's iterations converges fast and the final clustering error is decided by the separation strength Δ.



(a) Simulation **S1-1**: Log of clustering error ($K = 2$) with Δ varying under Gaussian noise. (b) Simulation **S1-2** Log of clustering error ($K = 3$) with Δ varying under Bernoulli noise (MMSBM).

Fig 1: (Convergence behavior of Algorithm 1) Log of clustering error with Δ varying under two scenarios: LrMM with Gaussian noise and MMSBM with Bernoulli noise.

In the second simulation setting **S2**, we aim to compare the final clustering error of vanilla Lloyd's algorithm and our low-rank Lloyd's algorithm. The dimensions are varied at two cases $d_1 = d_2 \in \{50, 100\}$, sample size is set as $n \in \{100, 200\}$, number of clusters $K = 2$ and ranks $r_1 = r_2 = 3$. The latent labels are generated as in **S1**. For each d_1 and n, the simulation is repeated for 100 times and their average clustering error rate is reported.

[4] We emphasize that our Theorem 6.2 is not directly applicable to MMSBM due to non-i.i.d. noise.

In **S2-1**, the population center matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are constructed such that they share identical singular spaces. More exactly, we extract singular vectors $\mathbf{U}_1$, $\mathbf{V}_1$ and singular value matrix $\mathbf{\Sigma}_1$ as is done in **S1-1**. Then the population center matrices are set as $\mathbf{M}_1 = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^\top$ and $\mathbf{M}_2 = \mathbf{U}_1(\mathbf{\Sigma}_1 + \mathrm{diag}\{\Delta/3, \Delta/3, \Delta/3\})\mathbf{V}_1^\top$. Here the signal strength is fixed at $\lambda = 10$ and the separation parameter is chosen from $\Delta \in \{1, 5, 10\}$. The final clustering error and its standard error by four methods are reported in the upper half of Table 4. Noted that the initialization of "vec-Lloyd" in [44] is attained by spectral clustering on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$. We observe that the clustering errors of four methods all decrease as Δ increases. However, lr-Lloyd initialized by Algorithm 2 achieves a much smaller clustering error compared with other methods. This is due to the fact that our proposed tensor-based spectral initialization is capable to capture the low-rank signal whereas both spectral clustering and naive K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$ ignores the low-rank structure in the other two modes of $\boldsymbol{\mathcal{M}}$. As a result, all the other three methods perform almost the same under current setting. Lastly, the bold-font column in Table 4 confirms Theorem 3.1 in that the clustering error achieved by TS-init initialized lr-Lloyd algorithm is only determined by Δ regardless of the dimension d_1, d_2 or the sample size n.

In **S2-2**, the singular vectors of $\mathbf{M}_1$ and $\mathbf{M}_2$ are generated exactly the same as in **S1-1**. The singular values of $\mathbf{M}_1$ and $\mathbf{M}_2$ are set as $\mathbf{\Sigma}_1 = \mathrm{diag}(1.2\lambda, 1.1\lambda, \lambda)$ and $\mathbf{\Sigma}_2 = \mathrm{diag}(0.36, 0.33, 0.30)$, respectively. Then $\sigma_{\min}(\mathbf{M}_1) = \lambda$ and $\sigma_1(\mathbf{M}_2) = 0.36$. Here λ is varied at $\{1.9, 2.2, 2.5\}$ for the case $d_1 = d_2 = 50$ and $\{2.7, 3.0, 3.3\}$ for the case $d_1 = d_2 = 100$. Consequently, the signal strength of $\mathbf{M}_2$ is much smaller than $\mathbf{M}_1$ that corresponds to the weak SNR setting in Section 5, and we test the performance of the relaxed lr-Lloyd's algorithm (Algorithm 3). The results are reported in the lower half of Table 4. Clearly, rlr-Lloyd's algorithm outperforms the vanilla Lloyd's algorithm (i.e., the vectorized version). In certain cases, the vanilla Lloyd's algorithm merely beats a random guess whereas the rlr-Lloyd's algorithm almost achieves zero clustering error. We also observe that rlr-Lloyd's algorithm still performs nicely if initialized by K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$.

9.2. Real Data Applications

We now demonstrate the merits of our proposed low-rank Lloyd's (lr-Lloyd) algorithm on several real-world datasets and compare with existing methods.

9.2.1. BHL dataset

The BHL (brain, heart and lung) dataset[5], which had been analyzed in [49], consists of $d_1 = 1124$ gene expression profiles of $n = 27$ brain, heart, or lung tissues. Each tissue is measured repeatedly for $d_2 = 4$ times and hence the ith sample can be constructed as $\mathbf{X}_i \in \mathbb{R}^{1124\times 4}$ for $i = 1, \cdots, 27$. Our aim is to correctly identify those $\mathbf{X}_i$'s belonging to the same type of tissue, i.e., $K = 3$. We

[5]The dataset is publicly available at https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1083.

Setting	$d_1 = d_2$	n	λ	Δ	vec-Lloyd [44]	lr-Lloyd initialized by TS-Init (Algorithm 2)	vec-Lloyd initialized by K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$	lr-Lloyd initialized by K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$
S2-1	50	100	10	1	0.461 (0.032)	**0.401 (0.058)**	0.462 (0.030)	0.459 (0.031)
			10	5	0.459 (0.033)	**0.163 (0.039)**	0.456 (0.033)	0.452 (0.034)
			10	10	0.458 (0.034)	**0.066 (0.025)**	0.441 (0.047)	0.433 (0.054)
		200	10	1	0.475 (0.019)	**0.398 (0.056)**	0.469 (0.025)	0.466 (0.025)
			10	5	0.473 (0.021)	**0.152 (0.027)**	0.462 (0.027)	0.450 (0.039)
			10	10	0.471 (0.022)	**0.063 (0.016)**	0.437 (0.041)	0.380 (0.082)
	100	100	10	1	0.461 (0.028)	**0.391 (0.069)**	0.460 (0.033)	0.461 (0.033)
			10	5	0.461 (0.029)	**0.157 (0.054)**	0.455 (0.036)	0.455 (0.036)
			10	10	0.460 (0.029)	**0.063 (0.026)**	0.458 (0.034)	0.456 (0.034)
		200	10	1	0.468 (0.023)	**0.390 (0.064)**	0.469 (0.023)	0.467 (0.023)
			10	5	0.468 (0.024)	**0.147 (0.028)**	0.469 (0.022)	0.465 (0.026)
			10	10	0.467 (0.024)	**0.062 (0.017)**	0.459 (0.030)	0.451 (0.037)
Setting	$d_1 = d_2$	n	$\sigma_{\min}(\mathbf{M}_1)$	Δ	vec-Lloyd [44]	rlr-Lloyd (Algorithm 3)	vec-Lloyd initialized by K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$	rlr-Lloyd initialized by K-means on $\mathscr{M}_3(\boldsymbol{\mathcal{X}})$
S2-2	50	100	1.9	3.68	0.434 (0.052)	**0.314 (0.138)**	0.418 (0.066)	0.327 (0.129)
			2.2	4.24	0.424 (0.061)	**0.134 (0.125)**	0.385 (0.079)	0.152 (0.138)
			2.5	4.81	0.417 (0.068)	**0.041 (0.051)**	0.309 (0.103)	0.055 (0.091)
		200	1.9	3.68	0.433 (0.052)	**0.070 (0.020)**	0.380 (0.070)	0.072 (0.046)
			2.2	4.24	0.431 (0.054)	**0.057 (0.018)**	0.351 (0.077)	0.059 (0.048)
			2.5	4.81	0.424 (0.057)	**0.035 (0.015)**	0.268 (0.088)	0.033 (0.014)
	100	100	2.7	5.19	0.422 (0.056)	**0.300 (0.169)**	0.416 (0.057)	0.301 (0.164)
			3	5.76	0.421 (0.059)	**0.131 (0.164)**	0.390 (0.077)	0.176 (0.181)
			3.3	6.33	0.426 (0.053)	**0.067 (0.139)**	0.347 (0.086)	0.065 (0.130)
		200	2.7	5.19	0.442 (0.040)	**0.019 (0.010)**	0.395 (0.071)	0.022 (0.037)
			3	5.76	0.443 (0.041)	**0.008 (0.006)**	0.301 (0.089)	0.008 (0.007)
			3.3	6.33	0.440 (0.043)	**0.003 (0.004)**	0.190 (0.069)	0.003 (0.004)

TABLE 4

Clustering error of lr-Lloyd (Algorithm 1) and rlr-Lloyd (Algorithm 3) compared with vanilla Lloyd's algorithm [44] on vectorized data (vec-Lloyd). The number in brackets represents the standard error over 100 *trials.*

	lr-Lloyd	DEEM	K-means	SKM	DTC	TBM	EM	AFPF
Clustering error	**3.70**	7.41	11.11	11.11	18.52	11.11	11.11	11.11

TABLE 5

Clustering error on BHL dataset. SKM: sparse K-means [58]; DTC: dynamic tensor clustering [53]; TBM: tensor block model (TBM) [57]; EM: standard EM implemented in [49]; AFPF: adaptive pairwise fusion penalized clustering [21].

apply Algorithm 1 together with an initial clustering $\widehat{\mathbf{s}}^{(0)}$ obtained by Algorithm 2 with $r_{\mathbf{U}} = r_{\mathbf{V}} = 1$. These ranks are chosen based on the scree plots of $\mathscr{M}_1(\boldsymbol{\mathcal{X}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{X}})$. The final clustering error attained by lr-Lloyd's algorithm is $n^{-1} \cdot h_c(\widehat{\mathbf{s}}, \mathbf{s}^*) = 0.03704$. As shown in Table 5, our lr-Lloyd's algorithm performs the best among all the competitors[6] that are reported in [49].

The improvement can be attributed to two reasons. First, DEEM in [49] is designed based on EM algorithm targeted at Gaussian probability distribution, and hence they need to first perform multiple Kolmogorov-Smirnov tests to drop the columns not following Gaussian distribution, which might lead to potential information loss. In sharp contrast, their procedure is not necessary for our method, as the low-rank Lloyd's algorithm allows for sub-Gaussian noise. Secondly, our algorithm is more suitable for the specific structure of the data. Particularly, the population center matrices are expected to be rank-one as the columns of $\mathbf{X}_i$ represent repeated measurements for the same sample. However, such planted structure is under-exploited in [49] and others.

[6]Note that all results except lr-Lloyd are directly borrowed from [49], which use $\mathbf{X}_i$'s after dimension reduction to a size of either 20×4 or 30×4, and we only report the better one here.

9.2.2. EEG dataset

The EEG dataset[7] has been extensively studied by various statistical models [39, 67, 28, 29]. The goal is to inspect EEG correlations of genetic predisposition to alcoholism. The data contains measurements which were sampled at $d_1 = 256$ Hz for 1 second, from $d_2 = 64$ electrodes placed on each scalp of $n = 122$ subjects. Each subject, either being *alcoholic* or not, completed 120 trials under different stimuli. More detailed description of the dataset can be found in [66]. For our application, we average all the trials for each subject under single stimulus condition (S1) and two matched stimuli condition (S2), respectively, and construct the data tensor as $\boldsymbol{\mathcal{X}}^{(S_1)} \in \mathbb{R}^{256\times 64\times 122}$ (or $\boldsymbol{\mathcal{X}}^{(S_2)} \in \mathbb{R}^{256\times 64\times 122}$) after standardization. Thus each subject is associated with a 256×64 matrix, and we aim to cluster these subjects into $K = 2$ groups, corresponding to alcholic group and control group. We apply rlr-Lloyd's algorithm (Algorithm 3) with $r_{\mathbf{U}} = r_{\mathbf{V}} = 3$ and $r_1 = 2, r_2 = 1$. Here $r_{\mathbf{U}}$ and $r_{\mathbf{V}}$ are selected by the scree plot of $\mathscr{M}_1(\boldsymbol{\mathcal{X}})$ and $\mathscr{M}_2(\boldsymbol{\mathcal{X}})$, and r_1 and r_2 are tuned by interpreting the final outcomes. The clustering error of our method and competitors are shown in Table 6. It is worth pointing out that our task of clustering is generally more challenging than classification, which has been investigated on the EEG dataset [39, 67, 28, 29]. Those classification approaches often achieve lower *classification* error rates. As a faithful comparison, our rlr-Lloyd's algorithm enjoys a superior performance to its competitors in terms of *clustering* error rate and time complexity.

Surprisingly, we note that the original lr-Lloyd's algorithm (Algorithm 1 + Algorithm 2) would not deliver a satisfactory result on this dataset. It can be partially explained by Figure 2, which displays the average of all trials under S2 for two groups. It is readily seen that the average matrix of control group is comparatively close to pure noise, and hence the relaxed version of lr-Lloyd's algorithm can work reasonably well in this scenario.

	rlr-Lloyd	vec-Lloyd	SKM	DTC	TBM
S1	**39.34**	42.62	44.26	45.08	43.44
S2	**28.69**	35.25	36.07	39.34	35.25

TABLE 6

Clustering error of EEG dataset under S1 and S2. Note that the methods vec-Lloyd and SKM [58] refer to directly applying Lloyd's algorithm and sparse K-means on vectorized data, i.e., on rows of $\mathscr{M}_3(\boldsymbol{\mathcal{X}}^{(S_1)})$ or $\mathscr{M}_3(\boldsymbol{\mathcal{X}}^{(S_2)})$, whereas DTC[53] and TBM [57] are both tensor-based clustering methods.

9.2.3. Malaria parasite genes networks dataset

We then consider the *var* genes networks of the human malaria parasite *Plasmodium falciparum* constructed by [37] via mapping $n = 9$ highly variable regions (HVRs) to a multi-layer network. Following the practice in [31], we focus

[7]The dataset is publicly available at https://archive.ics.uci.edu/ml/datasets/EEG+Database.

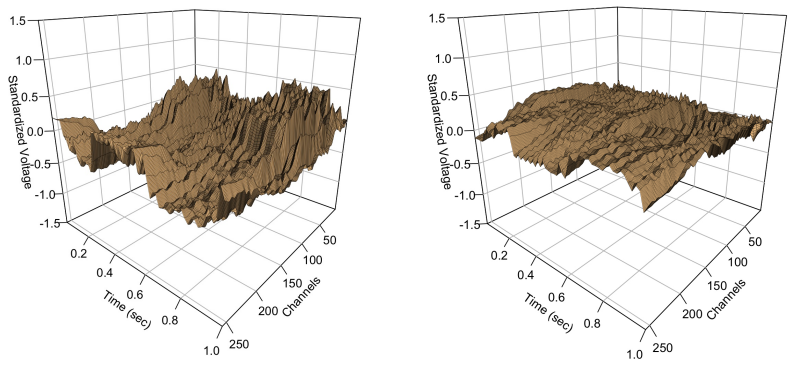


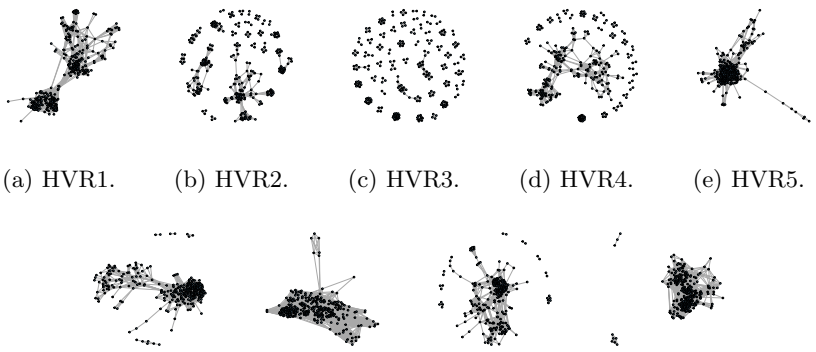
Fig 2: EEG dataset: average of matrix observations for alcoholic group (left) and control group (right) under S2.

on $d_1 = d_2 = 212$ common nodes appearing on all 9 layers and obtain a multi-layer network adjacency tensor $\boldsymbol{\mathcal{X}} \in \{0,1\}^{212\times212\times9}$ with each layer being the associated adjacency matrix. Unfortunately, the method in [37] needs to discard 3 out of 9 HVRs due to their extreme sparse structures, referring to region $\{2,3,4\}$ in Figure 3. This later had been remedied by the tensor-decomposition-based method TWIST in [31]. In term of clustering all layers, we expect our algorithm would have a comparable performance in contrast with the results in [31]. Specifically, [31] obtain a hierarchical structure with 6 clusters of all layers by repeatedly clustering the embedding vectors. Following their practice, by setting $(r_{\mathbf{U}}, r_{\mathbf{V}}, K) = (15, 15, 6)$, we apply Algorithm 2 on $\boldsymbol{\mathcal{X}}$, and find that the 9 HVRs fall in to the following clusters: $\{1\}, \{2,3,4,5\}, \{6\}, \{7\}, \{8\}, \{9\}$. The result is exactly the same as that in [31] but our method avoid repeated clustering. We remark that our tensor-based spectral initialization already produces a good initial clustering on this dataset, and thus further low-rank Lloyd's iterations seem unnecessary. In sharp contrast, it would lead to unsatisfactory result if we directly apply K-means with $K = 6$ on the embedding matrix obtained by TWIST. This further demonstrates the validity and flexibility of our proposed lr-Lloyd's algorithm.

9.2.4. UN comtrade trade flow networks dataset

In the last example, we consider the international commodity trade flow data in 2019 in terms of countries/regions and different types of commodities, collected by [47] from *UN comtrade Database*[8]. Following the data processing procedure in [47], we pick out top $d_1 = d_2 = 48$ countries/regions ranked by exports and obtain a weighted adjacency tensor $\widetilde{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{48\times48\times97}$, where $n = 97$ layers

[8]The dataset is publicly available at https://comtrade.un.org.



(a) HVR1. (b) HVR2. (c) HVR3. (d) HVR4. (e) HVR5.

(f) HVR6. (g) HVR7. (h) HVR8. (i) HVR9.

Fig 3: Malaria parasite genes networks dataset: 9 highly variable regions (HVRs) represented by their adjacency matrices [31]

represent different categories of commodities[9]. The entry $\widetilde{\boldsymbol{\mathcal{X}}}(i_1, i_2, i_3)$ indicates the amount of exports from country i_1 to country i_2 in terms of commodity type i_3. To have a comparable magnitude across different entries, our data tensor is obtained after transformation $\boldsymbol{\mathcal{X}} = \log(\widetilde{\boldsymbol{\mathcal{X}}} + 1)$. We emphasize that in [47] the edges of $\boldsymbol{\mathcal{X}}$ have to be further converted to binary under their framework, which might cause undesirable information loss. We apply Algorithm 1 that is initialized by Algorithm 2 with parameters $(r_{\mathbf{U}}, r_{\mathbf{V}}, K) = (3, 3, 2)$ and $(r_1, r_2) = (2, 2)$. These choices produce most interpretable result as summarized in Table 7. It is intriguing to notice that cluster 1 mainly consists of products of low durability including animal & vegetable products and part of foodstuffs, whereas cluster 2 contains most industrial products that might indicate a trend of global trading. These findings are consistent with [47].

Funding

Dong Xia's research was partially supported by Hong Kong RGC Grant GRF 16300121 and GRF 16301622.

References

[1] Abbe, E., Fan, J. and Wang, K. (2020). An ℓ_p theory of PCA and spectral clustering. *arXiv preprint arXiv:2006.14062*.

[2] Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E. and Vogelstein, J. T. (2021). Inference for multiple heterogeneous networks

[9]The categories are based on 2-digit HS code in https://www.foreign-trade.com/reference/hscode.htm.

Commodity cluster 1	Commodity cluster 2
01-05 Animal & Animal Products (100%)	15 Vegetable Products (13.73%)
06-14 Vegetable Products (86.27%)	19-22 Foodstuffs (60.82%)
16-18, 23-24 Foodstuffs (39.18%)	**25,27 Mineral Products (86.68%)**
26 Mineral Products (13.32%)	**28-30,32-35,38 Chemicals & Allied Industries (96.46%)**
31,36-37 Chemicals & Allied Industries (3.54%)	**39-40 Plastics / Rubbers (100%)**
41,43 Raw Hides, Skins, Leather, & Furs (23.01%)	**42 Raw Hides, Skins, Leather, & Furs (76.99%)**
45-47 Wood & Wood Products (15.13%)	**44,48-49 Wood & Wood Products (84.87%)**
50-55,57-58,60 Textiles (23.40%)	**56,59,61-63 Textiles (65.97%)**
65-67 Footwear / Headgear (17.45%)	**64 Footwear / Headgear (82.55%)**
75,78-81 Metals (6.44%)	**68-71 Stone / Glass (100%)**
86,89 Transportation (5.50%)	**72-74,76,82-83 Metals (93.56%)**
91-93,97 Miscellaneous (8.19%)	**84-85 Machinery / Electrical (100%)**
	87-88 Transportation (94.50%)
	90,94-96,99 Miscellaneous (91.81%)

TABLE 7

Clustering result of UN comtrade network. The number in brackets is the percentage of the amount of exports in the corresponding type of commodity.

with a common invariant subspace. *Journal of Machine Learning Research* **22** 1–49.

[3] ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V. and QIN, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* **18** 8393–8484.

[4] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* **45** 77–120.

[5] CAI, B., ZHANG, J. and SUN, W. W. (2021). Jointly Modeling and Clustering Tensors in High Dimensions. *arXiv preprint arXiv:2104.07773*.

[6] CAI, J.-F., LI, J. and XIA, D. (2022). Generalized low-rank plus sparse tensor estimation by fast Riemannian optimization. *Journal of the American Statistical Association* **just-accepted** 1–39.

[7] CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* **46** 60–89.

[8] CATTELL, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research* **1** 245–276.

[9] CHEN, S., LIU, S. and MA, Z. (2020). Global and individualized community detection in inhomogeneous multilayer networks. *arXiv preprint arXiv:2012.00933*.

[10] CHEN, X. and YANG, Y. (2021). Cutoff for exact recovery of Gaussian mixture models. *IEEE Transactions on Information Theory* **67** 4223–4238.

[11] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics* 493–507.

[12] DASGUPTA, S. (2008). *The hardness of k-means clustering*. Department of Computer Science and Engineering, University of California ....

[13] DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* **21** 1253–1278.

[14] Ding, Y., Kunisky, D., Wein, A. S. and Bandeira, A. S. (2019). Subexponential-time algorithms for sparse PCA. *arXiv preprint arXiv:1907.11635*.

[15] Dong, X., Frossard, P., Vandergheynst, P. and Nefedov, N. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing* **60** 5820–5831.

[16] Fei, Y. and Chen, Y. (2018). Hidden integrality of SDP relaxations for sub-Gaussian mixture models. In *Conference On Learning Theory* 1931–1965. PMLR.

[17] Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics* **46** 2153–2185.

[18] Gao, C. and Zhang, A. Y. (2022). Iterative algorithm for discrete structure recovery. *The Annals of Statistics* **50** 1066–1094.

[19] Gao, X., Shen, W., Zhang, L., Hu, J., Fortin, N. J., Frostig, R. D. and Ombao, H. (2021). Regularized matrix data clustering and its application to image analysis. *Biometrics* **77** 890–902.

[20] Gavish, M. and Donoho, D. L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory* **63** 2137–2152.

[21] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66** 793–804.

[22] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory* **62** 2788–2797.

[23] Han, Q., Xu, K. and Airoldi, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning* 1511–1520. PMLR.

[24] Han, R., Willett, R. and Zhang, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics* **50** 1–29.

[25] Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association* **97** 1090–1098.

[26] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks* **5** 109–137.

[27] Hopkins, S. (2018). Statistical inference and the sum of squares method, PhD thesis, Cornell University.

[28] Hu, W., Shen, W., Zhou, H. and Kong, D. (2020). Matrix linear discriminant analysis. *Technometrics* **62** 196–205.

[29] Huang, H.-H., Yu, F., Fan, X. and Zhang, T. (2022). Robust Regularized Low-Rank Matrix Models for Regression and Classification. *arXiv preprint arXiv:2205.07106*.

[30] Jin, J., Ke, Z. T. and Wang, W. (2017). Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics* **45** 2151–2189.

[31] Jing, B.-Y., Li, T., Lyu, Z. and Xia, D. (2021). Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics* **49** 3181–3205.
[32] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.
[33] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133.
[34] Kumar, A., Rai, P. and Daume, H. (2011). Co-regularized multi-view spectral clustering. *Advances in neural information processing systems* **24**.
[35] Kumar, A., Sabharwal, Y. and Sen, S. (2004). A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science* 454–462. IEEE.
[36] Kunisky, D., Wein, A. S. and Bandeira, A. S. (2019). Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*.
[37] Larremore, D. B., Clauset, A. and Buckee, C. O. (2013). A network approach to analyzing highly recombinant malaria parasite genes. *PLoS computational biology* **9** e1003268.
[38] Levin, K., Lodhia, A. and Levina, E. (2019). Recovering low-rank structure from multiple networks with unknown edge distributions. *arXiv preprint arXiv:1906.07265*.
[39] Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics* **38** 1094–1121.
[40] Liu, T., Yuan, M. and Zhao, H. (2022). Characterizing Spatiotemporal Transcriptome of the Human Brain Via Low-Rank Tensor Decomposition. *Statistics in Biosciences* 1–29.
[41] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory* **28** 129–137.
[42] Löffler, M., Wein, A. S. and Bandeira, A. S. (2020). Computationally efficient sparse clustering. *arXiv preprint arXiv:2005.10817*.
[43] Löffler, M., Zhang, A. Y. and Zhou, H. H. (2021). Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics* **49** 2506–2530.
[44] Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.
[45] Luo, Y. and Zhang, A. R. (2022). Tensor clustering with planted structures: Statistical optimality and computational limits. *The Annals of Statistics* **50** 584–613.
[46] Lyu, Z. and Xia, D. (2022). Optimal Estimation and Computational Limit of Low-rank Gaussian Mixtures. *arXiv preprint arXiv:2201.09040*.
[47] Lyu, Z., Xia, D. and Zhang, Y. (2021). Latent Space Model for Higher-order Networks and Generalized Tensor Decomposition. *arXiv preprint arXiv:2106.16042*.
[48] Mahajan, M., Nimbhorkar, P. and Varadarajan, K. (2009). The pla-

nar k-means problem is NP-hard. In *International workshop on algorithms and computation* 274–285. Springer.

[49] Mai, Q., Zhang, X., Pan, Y. and Deng, K. (2021). A doubly enhanced em algorithm for model-based tensor clustering. *Journal of the American Statistical Association* 1–15.

[50] Ndaoud, M. (2018). Sharp optimal recovery in the two-component Gaussian mixture model. *arXiv preprint arXiv:1812.08078*.

[51] Richard, E. and Montanari, A. (2014). A statistical model for tensor PCA. *Advances in neural information processing systems* **27**.

[52] Stanley, N., Shai, S., Taylor, D. and Mucha, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE transactions on network science and engineering* **3** 95–105.

[53] Sun, W. W. and Li, L. (2019). Dynamic tensor clustering. *Journal of the American Statistical Association* **114** 1894–1907.

[54] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68** 841–860.

[55] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.

[56] Verzelen, N. and Arias-Castro, E. (2017). Detection and feature selection in sparse mixture models. *The Annals of Statistics* **45** 1920–1950.

[57] Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. *Advances in neural information processing systems* **32**.

[58] Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105** 713–726.

[59] Wu, Y. and Zhou, H. H. (2019). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*.

[60] Xia, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics* **15** 3798–3851.

[61] Xia, D. and Yuan, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 58–77.

[62] Xia, D., Zhang, A. R. and Zhou, Y. (2022). Inference for low-rank tensors—no need to debias. *The Annals of Statistics* **50** 1220–1245.

[63] Xia, D. and Zhou, F. (2019). The sup-norm perturbation of hosvd and low rank tensor denoising. *The Journal of Machine Learning Research* **20** 2206–2247.

[64] Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory* **64** 7311–7338.

[65] Zhang, A. Y. and Zhou, H. H. (2022). Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering. *arXiv preprint arXiv:2205.14855*.

[66] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain research bulletin* **38** 531–538.

[67] Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 463–483.

Supplementary Material

10. Proofs of Main Theorems

10.1. Proof of Theorem 3.1

Step 1: Notations and Good Initialization We need to introduce some notations to simplify the presentation of our proof. Define the following loss related to the parameter tensor $\boldsymbol{\mathcal{M}}$:

$$\ell(\mathbf{s}, \mathbf{s}^*) := \sum_{i=1}^{n} \left\| \mathbf{M}_{s_i} - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2$$

and the hamming loss with respect to the true label $\mathbf{s}^*$:

$$h(\mathbf{s}, \mathbf{s}^*) := \sum_{i=1}^{n} \mathbb{I}\left(s_i \neq s_i^*\right)$$

A simple relation is that $h(\mathbf{s}, \mathbf{s}^*) \leq \Delta^{-2} \cdot \ell(\mathbf{s}, \mathbf{s}^*)$ due to the fact

$$\sum_{i=1}^{n} \|\mathbf{M}_{s_i} - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \geq \sum_{i=1}^{n} \mathbb{I}\left(s_i \neq s_i^*\right) \Delta^2.$$

Note that, by definition the Hamming clustering error $h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = \sum_{i=1}^{n} \mathbb{I}\left(\pi(s_i^{(0)}) \neq s_i^*\right)$ for some permutation π, we can always relabel our $\mathbf{M}_1, \cdots, \mathbf{M}_K$ to $\mathbf{M}_{\pi(1)}, \cdots, \mathbf{M}_{\pi(K)}$ after initialization. Therefore, without loss of generality we can assume $\pi = \mathrm{Id}$ and hence $h(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*)$. On the other hand, by eq. (10) we have

$$\ell(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \leq \gamma^2 \Delta^2 h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = o\left(\frac{\alpha n \Delta^2}{(\kappa_0 \vee \gamma^2) K}\right) \tag{22}$$

Note that (22) implies that $\ell(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \leq \tau$ for some $\tau = o\left((\kappa_0 \vee \gamma^2)^{-1} \alpha n \Delta^2 / K\right)$ and hence $\Delta^2 \gg (\kappa_0 \vee \gamma^2) K \tau / (\alpha n)$.

Step 2: Iterative Convergence We then analyze the convergence property of low-rank Lloyd algorithm. Without loss of generality, given the labelling $\widehat{\mathbf{s}}^{(t-1)}$ at the $(t-1)$-th iteration, we investigate the behavior of $\widehat{\mathbf{s}}^{(t)}$, i.e., after one iteration of Lloyd algorithm.

To simplify the presentation, the subsequent analysis is conditioned on the following events, where $C > 0$ is some absolute constant.

$$\mathcal{Q}_1 = \bigcup_{k \in [K]} \left\{ \left\| \frac{\sum_{i=1}^{n} \mathbb{I}\left(s_i^* = k\right) \mathbf{E}_i}{\sum_{i=1}^{n} \mathbb{I}\left(s_i^* = k\right)} \right\| \leq C \sqrt{\frac{d}{n_k^*}} \right\}$$

$$\mathcal{Q}_2 = \bigcup_{I \in [n]} \left\{ \left\| \frac{1}{\sqrt{|I|}} \sum_{i \in I} \mathbf{E}_i \right\| \leq C \left(\sqrt{d} + \sqrt{n}\right) \right\}$$

The following lemma dictates that $\mathcal{Q}_1 \cap \mathcal{Q}_2$ occurs with high probability.

Lemma 10.1. *There exists some absolute constants $C_0, c_0 > 0$ such that if $d \geq C_0 \log K$, then*

$$\mathbb{P}\left(Q_1^c \cup Q_2^c\right) \leq \exp(-c_0 d)$$

Our goal is to establish the following relation between two sucessive iterations:

$$\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}) \leq 2n \cdot \exp\left\{-\big(1 - o(1)\big)\frac{\Delta^2}{8}\right\} + \frac{1}{2}\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}) \tag{23}$$

and prove that it holds with high probability for all positive integer t.

Suppose for iteration $t-1$, $h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ satisfies (10) and $\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ satisfies (22). By the definition of $\widehat{\boldsymbol{s}}^{(t)}$, we have for each $i \in [n]$:

$$\left\|\mathbf{X}_i - \widehat{\mathbf{M}}^{(t)}_{\widehat{s}_i^{(t)}}\right\|_{\mathrm{F}}^2 \leq \left\|\mathbf{X}_i - \widehat{\mathbf{M}}^{(t)}_{s_i^*}\right\|_{\mathrm{F}}^2$$

Rearranging terms above, we obtain

$$\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}^{(t)}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{\widehat{s}_i^{(t)}} \right\rangle \leq -\frac{1}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_{\widehat{s}_i^{(t)}}\right\|_{\mathrm{F}}^2 + \mathcal{R}\left(\widehat{s}_i^{(t)}; \widehat{\mathbf{s}}^{(t-1)}\right) \tag{24}$$

where

$$\mathcal{R}\left(a; \widehat{\mathbf{s}}^{(t-1)}\right) := \frac{1}{2}\left[\left\|\mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{s_i^*}\right\|_{\mathrm{F}}^2 - \left\|\mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{a}\right\|_{\mathrm{F}}^2 + \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right]$$

Without loss of generality, suppose $\widehat{s}_i^{(t)} = a$ for some $a \in [K]$. Set $\delta = o(1)$ that is to be determined later. The following fact is obvious.

$$\begin{aligned}
&\mathbb{I}\left(\widehat{s}_i^{(t)} = a\right) \\
&= \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}^{(t)}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{a} \right\rangle \leq -\frac{1}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 + \mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)})\right) \\
&\leq \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{s_i^*} \right\rangle + \left\langle \mathbf{E}_i, \widehat{\mathbf{M}}^{(t)}_{a} - \mathbf{M}_a \right\rangle + \mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\leq \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}^{(t)}_{s_i^*} \right\rangle + \left\langle \mathbf{E}_i, \widehat{\mathbf{M}}^{(t)}_{a} - \mathbf{M}_a \right\rangle \geq \frac{\delta}{4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\mathbb{I}\left(\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}$$

By the definition of $\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)$, we have

$$\begin{aligned}
\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*) &= \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} = a \right) \\
&\leq \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right) \\
&+ \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} = a \right) \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)} \right\rangle + \left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a \right\rangle \geq \frac{\delta}{4} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right) \\
&+ \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} = a \right) \mathbb{I}\left(\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{4} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right) \\
&=: \xi_{\mathsf{err}} + \beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) + \beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})
\end{aligned}$$

where we define

$$\xi_{\mathsf{err}} := \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right)$$

and

$$\begin{aligned}
\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} & \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a \right) \\
& \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)} \right\rangle + \left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a \right\rangle \geq \frac{\delta}{4} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right)
\end{aligned}$$

and

$$\beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a \right) \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{I}\left(\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{4} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right)$$

It suffices to bound $\xi_{\mathsf{err}}, \beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ and $\beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$, respectively.

Step 2.1: Bounding ξ_{err}. Let us begin with $\mathbb{E}\xi_{\mathsf{err}}$. By definition,

$$\mathbb{E}\xi_{\mathsf{err}} = \sum_{i=1}^{n} \sum_{a \in [K] \setminus \{s_i^*\}} \left\| \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\|_{\mathrm{F}}^2 \mathbb{P}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right)$$

Note that $\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle$ is normal distribution with mean zero and variance $\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2$. The standard concentration inequality of normal random variable yields

$$\mathbb{P}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*} \right\rangle \geq \frac{1-\delta}{2} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right) \leq \exp\left(-\frac{(1-\delta)^2}{8} \left\| \mathbf{M}_{s_i^*} - \mathbf{M}_a \right\|_{\mathrm{F}}^2 \right)$$

Therefore,

$$\mathbb{E}\xi_{\mathsf{err}} \le \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \exp\left(-\frac{(1-\delta)^2}{8}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right).$$

Assume $n \gg K$, $\Delta^2 \gg \log K$ and let δ converge to 0 as slow as possible, we can get

$$\mathbb{E}\xi_{\mathsf{err}} \le n \cdot \exp\left\{-\big(1-o(1)\big)\frac{\Delta^2}{8}\right\}$$

By Markov inequality,

$$\mathbb{P}\left(\xi_{\mathsf{err}} \ge \exp(\Delta)\mathbb{E}\xi_{\mathsf{err}}\right) \le \exp(-\Delta)$$

We conclude that, with probability at least $1-\exp(-\Delta)$,

$$\xi_{\mathsf{err}} \le \exp(\Delta)\mathbb{E}\xi_{\mathsf{err}} \le n \cdot \exp\left\{-\big(1-o(1)\big)\frac{\Delta^2}{8}\right\}$$

Step 2.2: Bounding $\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ By definition,

$$\begin{aligned}\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) =& \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \ne a\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\rangle \ge \frac{\delta}{8}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ &+ \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \ne a\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a\right\rangle \ge \frac{\delta}{8}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ &=: \beta_{1,1}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) + \beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})\end{aligned}$$

Without loss of generality, we only prove the upper bound of the second term $\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$. Notice that the labels $\widehat{\mathbf{s}}^{(t)}$ depend on all the noise matrices $\{\mathbf{E}_i\}_{i=1}^n$, thus $\widehat{\mathbf{M}}_a^{(t)}$ is dependent on $\mathbf{E}_i$. Delicate treatment is necessary to establish a sharp upper bound for $\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$.

Recall the definition that $\widehat{\mathbf{M}}_a^{(t)}$ is computed by the best rank-r_a approximation of $\bar{\mathbf{X}}_a(\widehat{\mathbf{s}}^{(t-1)}) := (n_a^{(t-1)})^{-1}\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right)\mathbf{X}_i$ with $n_a^{(t-1)} := \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right)$. Denote $\widehat{\mathbf{U}}_a^{(t)}$ and $\widehat{\mathbf{V}}_a^{(t)}$ the left and right singular vectors of $\widehat{\mathbf{M}}_a^{(t)}$. Then we have $\widehat{\mathbf{M}}_a^{(t)} = \widehat{\mathbf{U}}_a^{(t)}(\widehat{\mathbf{U}}_a^{(t)})^\top \bar{\mathbf{X}}_a(\widehat{\mathbf{s}}^{(t-1)})\widehat{\mathbf{V}}_a^{(t)}(\widehat{\mathbf{V}}_a^{(t)})^\top$. *For notation simplicity, we now drop the superscript* (t) *in* $\widehat{\mathbf{U}}_a^{(t)}$, $\widehat{\mathbf{V}}_a^{(t)}$ *and write* $\widehat{\mathbf{U}}_a, \widehat{\mathbf{V}}_a$ *instead.*

Now write

$$\begin{aligned}\widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a &= \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top \bar{\mathbf{X}}_a(\widehat{\mathbf{s}}^{(t-1)})\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{M}_a \\ &= \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top \left(\frac{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right)(\mathbf{M}_{s_i^*} + \mathbf{E}_i)}{\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right)}\right)\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{M}_a\end{aligned}$$

Recall that $n_a^* = \sum_{i=1}^n \mathbb{I}(s_i^* = a)$. Denote

$$\bar{\mathbf{E}}_a^* := (n_a^*)^{-1} \sum_{i=1}^n \mathbb{I}(s_i^* = a)\, \mathbf{E}_i \quad \text{and} \quad \bar{\mathbf{E}}_a^{(t-1)} := (n_a^{(t-1)})^{-1} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) \mathbf{E}_i$$

Then we can proceed as

$$\begin{aligned}
\widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a &= \widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top \left(\frac{1}{n_a^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) \mathbf{M}_{s_i^*} + \bar{\mathbf{E}}_a^{(t-1)} \right) \widehat{\mathbf{V}}_a \widehat{\mathbf{V}}_a^\top - \mathbf{M}_a \\
&= \widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top \left[\mathbf{M}_a + \frac{1}{n_a^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) (\mathbf{M}_{s_i^*} - \mathbf{M}_a) + \bar{\mathbf{E}}_a^* + (\bar{\mathbf{E}}_a^{(t-1)} - \bar{\mathbf{E}}_a^*) \right] \widehat{\mathbf{V}}_a \widehat{\mathbf{V}}_a^\top - \mathbf{M}_a \\
&= \widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top \left(\mathbf{M}_a + \bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)} \right) \widehat{\mathbf{V}}_a \widehat{\mathbf{V}}_a^\top - \mathbf{M}_a
\end{aligned}$$

where we've defined

$$\Delta_{\mathbf{M}}^{(t-1)} := \frac{1}{n_a^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) (\mathbf{M}_{s_i^*} - \mathbf{M}_a) \quad \text{and} \quad \Delta_{\mathbf{E}}^{(t-1)} := \bar{\mathbf{E}}_a^{(t-1)} - \bar{\mathbf{E}}_a^*$$

For simplicity, we denote $\Delta^{(t-1)} := \bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}$ and write

$$\widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a = \widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top \left(\mathbf{M}_a + \Delta^{(t-1)} \right) \widehat{\mathbf{V}}_a \widehat{\mathbf{V}}_a^\top - \mathbf{M}_a \tag{25}$$

Notice that since $h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ satisfies (10), we have that

$$\begin{aligned}
n_a^{(t-1)} = \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) &\geq \sum_{i=1}^n \mathbb{I}(s_i^* = a) - \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} \neq s_i^*\right) \\
&\geq n_a^* - h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \geq \frac{\alpha n}{K} - \frac{\alpha n}{8K} \geq \frac{7\alpha n}{8K}
\end{aligned}$$

The following lemma is useful whose proof is postponed to Section 11.

Lemma 10.2. *Suppose that $h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ satisfies (10). Then,*

$$\|\Delta_{\mathbf{M}}^{(t-1)}\| \leq \frac{16K}{7\alpha n} h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \cdot \min\{\kappa_0 \lambda, \gamma \Delta\}$$

where we define

$$h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) := \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a, s_i^* \neq a\right) + \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} \neq a, s_i^* = a\right)$$

Moreover, under event $\mathcal{Q}_1 \cap \mathcal{Q}_2$, there exist absolute constants $C_1, C_2 > 0$ such that

$$\left\|\bar{\mathbf{E}}_a^*\right\| \leq C_1 \sqrt{\frac{dK}{\alpha n}} \quad \text{and} \quad \|\Delta_{\mathbf{E}}^{(t-1)}\| \leq C_2 \frac{K\sqrt{(d+n) \cdot h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}}{\alpha n}$$

By Lemma 10.2, we obtain, under $\mathcal{Q}_1 \cap \mathcal{Q}_2$, that

$$\left\|\Delta^{(t-1)}\right\| \le c\lambda + C\left(\alpha^{-1/2}K^{1/2}\sqrt{\frac{d}{n}} + \alpha^{-1}K\sqrt{\frac{h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{n}}\right)$$

Recall the condition $\sigma_{\min}(\mathbf{M}_a) \ge \lambda > C\left(\alpha^{-1/2}K^{1/2}(d/n)^{1/2} + \alpha^{-1}Kn^{-1/2}\cdot h_a^{1/2}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)\right)$. Together with the bound for $\Delta^{(t-1)}$, we have

$$\sigma_{\min}(\mathbf{M}_a) \ge \lambda > 2\left\|\Delta^{(t-1)}\right\|$$

provided that eq. (10) holds.

Continuing from eq. (25), we need a delicate representation formula for $\widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a$, which is guaranteed by the following lemma whose proof is deferred to Section 11.

Lemma 10.3. *For any rank-r matrix $\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$ with compact SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{O}_{d_1,r}$ and $\mathbf{V} \in \mathbb{O}_{d_2,r}$ and $\mathbf{\Sigma} = diag(\sigma_1, \cdots, \sigma_r)$ with $\sigma_1 \ge \cdots \ge \sigma_r > 0$. Let Δ be an arbitrary $d_1 \times d_2$ perturbation matrix and $\mathbf{X} = \mathbf{M} + \Delta$. Denote $\widehat{\mathbf{U}} \in \mathbb{O}_{d_1,r}, \widehat{\mathbf{V}} \in \mathbb{O}_{d_2,r}$ the top-r left and right singular vectors of $\mathbf{X}$. Suppose that $\sigma_r > 3\|\Delta\|$, then we have the following relation:*

$$\begin{bmatrix} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top \end{bmatrix} = \begin{bmatrix} \sum_{k\ge1}\mathcal{S}^{\mathbf{U}}_{\mathbf{M},k}(\Delta) & \mathbf{0} \\ \mathbf{0} & \sum_{k\ge1}\mathcal{S}^{\mathbf{V}}_{\mathbf{M},k}(\Delta) \end{bmatrix} = \sum_{k\ge1}\mathcal{S}_{\mathbf{M},k}(\Delta)$$

Here the k-th order perturbation term $\mathcal{S}_{\mathbf{M},k}(\Delta)$ is defined as

$$\mathcal{S}_{\mathbf{M},k}(\Delta) := \sum_{\mathbf{m}:m_1+\cdots+m_{k+1}=k} (-1)^{1+\tau(\mathbf{m})}\cdot \mathfrak{P}^{-m_1}\Delta^*\mathfrak{P}^{-m_2}\Delta^*\cdots\Delta^*\mathfrak{P}^{-m_{k+1}}$$

where $\mathbf{m} = (m_1, \cdots, m_{k+1})$ contains non-negative integers, $\tau(\mathbf{m}) = \sum_{i=1}^{k+1}\mathbb{I}(m_i > 0)$ and

$$\Delta^* := \begin{bmatrix} \mathbf{0} & \Delta \\ \Delta^\top & \mathbf{0} \end{bmatrix}, \quad \mathfrak{P}^{-k} := \begin{cases} \begin{pmatrix} \mathbf{0} & \mathbf{U}\mathbf{\Sigma}^{-k}\mathbf{V}^\top \\ \mathbf{V}\mathbf{\Sigma}^{-k}\mathbf{U}^\top & \mathbf{0} \end{pmatrix} & \text{if } k \text{ is odd} \\ \begin{pmatrix} \mathbf{U}\mathbf{\Sigma}^{-k}\mathbf{U}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{V}\mathbf{\Sigma}^{-k}\mathbf{V}^\top \end{pmatrix} & \text{if } k \text{ is even.} \end{cases}$$

for all $k \ge 1$. Specifically, $\mathfrak{P}^0 = \mathfrak{P}^\perp$ denotes the orthogonal spectral projector defined by

$$\mathfrak{P}^\perp = \begin{pmatrix} \mathbf{U}_\perp\mathbf{U}_\perp^\top & 0 \\ 0 & \mathbf{V}_\perp\mathbf{V}_\perp^\top \end{pmatrix}$$

By Lemma 10.3, we have the following decomposition

$$\begin{aligned} \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a &= \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\left(\mathbf{M}_a + \Delta^{(t-1)}\right)\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{M}_a \\ &= \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a + \mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right) \\ &+ \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right) + \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\Delta^{(t-1)}\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top \end{aligned}$$

so that we can re-write

$$\begin{aligned}\beta_{1,2}(\mathbf{s}^*,\widehat{\mathbf{s}}^{(t)}) \le &\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\right\rangle \ge \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \ge \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \ge \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\Delta^{(t-1)}\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top\right\rangle \ge \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\end{aligned} \tag{26}$$

It suffices to bound each term in the RHS of above equation.

Step 2.2.1: Treating the terms of $\langle \mathbf{E}_i, (\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top)\mathbf{M}_a\rangle$. By Lemma 10.3, we have

$$\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\right\rangle = \sum_{k\ge 1}\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle \tag{27}$$

The RHS of (27) is the sum of infinite series. It turns out that delicate treatments are only necessary for the leading two terms. Now we write

$$\begin{aligned}&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\right\rangle \ge \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ \le&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle \ge \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle \ge \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\ge 3}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle \ge \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\end{aligned} \tag{28}$$

The first and second perturbation terms $\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}$ and $\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}$ can be explicitly determined by Lemma 10.3. Indeed, we have

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle = \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle = \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathbf{V}_a \right\rangle \\
&= \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}^*_a\mathbf{V}_a \right\rangle + \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\
&= \frac{1}{n^*_a}\mathbb{I}\left(s^*_i = a\right)\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 + \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\frac{1}{n^*_a}\sum_{j\neq i}^n \mathbb{I}\left(s^*_j = a\right)\mathbf{E}_j\right)\mathbf{V}_a \right\rangle \\
&\quad + \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle
\end{aligned}$$

We then bound the first term on RHS of eq. (28) by

$$\begin{aligned}
&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s^*_i\}} \left\|\mathbf{M}_a - \mathbf{M}_{s^*_i}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\leq \sum_{i=1}^n \sum_{a\in[K]\setminus\{s^*_i\}} \left\|\mathbf{M}_a - \mathbf{M}_{s^*_i}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\frac{1}{n^*_a}\mathbb{I}\left(s^*_i = a\right)\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{288}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \sum_{i=1}^n \sum_{a\in[K]\setminus\{s^*_i\}} \left\|\mathbf{M}_a - \mathbf{M}_{s^*_i}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\frac{1}{n^*_a}\sum_{j\neq i}^n \mathbb{I}\left(s^*_j = a\right)\mathbf{E}_j\right)\mathbf{V}_a \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \sum_{i=1}^n \sum_{a\in[K]\setminus\{s^*_i\}} \left\|\mathbf{M}_a - \mathbf{M}_{s^*_i}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned} \tag{29}$$

We bound the first two terms on RHS of eq. (29) by Markov inequality and thus their expectation is needed. Since $\mathbf{E}_i$ has i.i.d. $N(0,1)$ entries, $\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2$ follows a Chi-squared distribution with degrees of freedom $(d_1 - r_a)r_a$. By the concentration inequality of Chi-squared random variable, we get

$$\mathbb{P}\left(\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq (d_1 - r_a)r_a + 2\sqrt{u(d_1 - r_a)r_a} + 2u\right) \leq \exp(-u)$$

for any $u > 0$. As a result, there exists an absolute constant $c_2 > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n^*_a}\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{288}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) = \mathbb{P}\left(\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq \frac{\delta n^*_a}{288}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\leq \exp\left(-c_2\frac{\delta\alpha n}{K}\left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}$$

where the last inequality holds due to the condition $\Delta^2 \gg \alpha^{-1}dKr/n$ and by setting $\delta = o(1)$ in the way that it converges to 0 sufficiently slowly.

For the second term in RHS of eq. (29), due to the property of Gaussian

distribution, we observe that

$$\widetilde{\mathbf{E}}_i := \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a \quad \text{and} \quad \widetilde{\mathbf{E}}_{-i} := \mathbf{U}_{a\perp}^\top \left(\sum_{j\neq i}^n \mathbb{I}\left(s_j^* = a\right) \mathbf{E}_j \right) \mathbf{V}_a$$

are two independent matrices, hence $\left\langle \widetilde{\mathbf{E}}_i, \widetilde{\mathbf{E}}_{-i} \right\rangle | \widetilde{\mathbf{E}}_i \sim N\big(0, (n_a^* - 1)\|\widetilde{\mathbf{E}}_i\|_{\mathrm{F}}^2\big)$. Then we can proceed as follows. There exists an absolute constant $c_3 > 0$ such that for any $u > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\sum_{j\neq i}^n \mathbb{I}\left(s_j^* = a\right) \mathbf{E}_j \right) \mathbf{V}_a \right\rangle \geq \frac{\delta n_a^*}{288} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \\
&\leq \mathbb{P}\left(\left\langle \widetilde{\mathbf{E}}_i, \widetilde{\mathbf{E}}_{-i} \right\rangle \geq \frac{\delta n_a^*}{288} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \,\middle|\, \left\|\widetilde{\mathbf{E}}_i\right\|_{\mathrm{F}}^2 \leq (d_1 - r_a) r_a + 2\sqrt{u(d_1 - r_a) r_a} + 2u \right) \\
&\cdot \mathbb{P}\left(\left\|\widetilde{\mathbf{E}}_i\right\|_{\mathrm{F}}^2 \leq (d_1 - r_a) r_a + 2\sqrt{u(d_1 - r_a) r_a} + 2u \right) + \mathbb{P}\left(\left\|\widetilde{\mathbf{E}}_i\right\|_{\mathrm{F}}^2 \geq (d_1 - r_a) r_a + 2\sqrt{u(d_1 - r_a) r_a} + 2u \right) \\
&\leq \exp\left(-c_3 \frac{\delta^2 \alpha n \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^4}{K\left(dr + 2\sqrt{dru} + 2u\right)} \right) + \exp(-u)
\end{aligned}$$

It suffices to choose $u = C_2 \delta (\alpha n / K)^{1/2} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$, then the above probability can be bounded as

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\sum_{j\neq i}^n \mathbb{I}\left(s_j^* = a\right) \mathbf{E}_j \right) \mathbf{V}_a \right\rangle \geq \frac{\delta n_a^*}{288} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \\
&\leq \exp\left(-c_2 \frac{\delta^2 \alpha n \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^4}{dKr} \right) + \exp\left(-c_2 \frac{\delta \sqrt{\alpha n} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2}{\sqrt{K}} \right) \\
&\leq 2\exp\left(-c_2 \frac{\delta \sqrt{\alpha n} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2}{\sqrt{K}} \right)
\end{aligned}$$

where the last inequality holds since $\Delta^2 \gg \alpha^{-1} r d K / n$, $\alpha n \gg K$ and by setting $\delta \to 0$ sufficiently slow. Therefore, the expectation of the first two terms on RHS of eq. (29) is bounded by

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\frac{1}{n_a^*} \mathbb{I}\left(s_i^* = a\right) \left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{288} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \\
&+ \sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\frac{1}{n_a^*} \sum_{j\neq i}^n \mathbb{I}\left(s_j^* = a\right) \mathbf{E}_j \right) \mathbf{V}_a \right\rangle \geq \frac{\delta}{288} \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \Bigg] \\
&\leq C_2 n K \Delta^2 \exp\left(-\delta (\alpha n / K)^{1/2} \Delta^2 \right) \leq n \exp\left(-c_2 \delta (\alpha n / K)^{1/2} \Delta^2 \right)
\end{aligned}$$

Now the first two terms on RHS of eq. (29) can be bounded by Markov inequality. We get, with probability at least $1-\exp\big(-\delta(\alpha n/K)^{1/4}\Delta\big)$ that

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^*=a\right)\left\|\mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$+\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a,\mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\sum_{j\neq i}^{n}\mathbb{I}\left(s_j^*=a\right)\mathbf{E}_j\right)\mathbf{V}_a\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq n\cdot\exp\left(-\delta(\alpha n/K)^{1/2}\Delta^2\right)$$

which holds as long as $\delta\to 0$ sufficiently slowly compared with $\alpha n/K\to\infty$.

We now bound the third term on RHS of eq. (29). Denote $\mathbf{\Xi}_1^{(t-1)}:=\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\left(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top$. Then write

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{\Xi}_1^{(t-1)}\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq\sum_{i=1}^{n}\sum_{a\in[K]}\sum_{b\in[K]\setminus\{a\}}\mathbb{I}\left(s_i^*=b\right)\left\|\mathbf{M}_a-\mathbf{M}_b\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{\Xi}_1^{(t-1)}\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_b-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq C_3\sum_{i=1}^{n}\sum_{a\in[K]}\sum_{b\in[K]\setminus\{a\}}\mathbb{I}\left(s_i^*=b\right)\left\|\mathbf{M}_a-\mathbf{M}_b\right\|_{\mathrm{F}}^2\cdot\frac{\langle\mathbf{E}_i,\mathbf{\Xi}_1^{(t-1)}\rangle^2}{\delta^2\|\mathbf{M}_b-\mathbf{M}_a\|_{\mathrm{F}}^4}$$
$$\leq C_3\sum_{a\in[K]}\sum_{b\in[K]\setminus\{a\}}\|\mathbf{\Xi}_1^{(t-1)}\|^2\cdot\frac{\sum_{i=1}^{n}\mathbb{I}\left(s_i^*=b\right)\langle\mathbf{E}_i,\mathbf{\Xi}_1^{(t-1)}/\|\mathbf{\Xi}_1^{(t-1)}\|\rangle^2}{\delta^2\|\mathbf{M}_b-\mathbf{M}_a\|_{\mathrm{F}}^2}\qquad(30)$$

The following lemma is needed whose proof is deferred to Section 11.

Lemma 10.4. *There exist absolute constants $c_1,C_1>0$ such that, for any fixed $b\in[K]$ and d_1,d_2 and r, the following inequality holds with probability at least $1-\exp(-c_1dr)$:*

$$\sup_{\substack{\mathbf{\Xi}\in\mathbb{R}^{d_1\times d_2},\mathrm{rank}(\mathbf{\Xi})\leq r\\\|\mathbf{\Xi}\|\leq 1}}\sum_{i=1}^{n}\mathbb{I}\left(s_i^*=b\right)\langle\mathbf{E}_i,\mathbf{\Xi}\rangle^2\leq C_1r(dr+n_b^*)$$

We denote the event in Lemma 10.4 by $\mathcal{Q}_3$ and proceed by conditioning on $\mathcal{Q}_3$. By Lemma 10.4 and eq. (30) we obtain that:

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{\Xi}_1^{(t-1)}\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\overset{(a)}{\leq}C_3'\sum_{a\in[K]}\sum_{b\in[K]\setminus\{a\}}\frac{r(dr+n_b^*)}{\delta^2\Delta^2}\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2h_a^2(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right)$$
$$\overset{(b)}{\leq}\frac{1}{16}\ell(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$$

where (a) holds due to Lemma 10.2 and (b) holds provided $\Delta^2 \gg \alpha^{-1}K^2r\,(dr/n+1)$ and (10).

We then bound the second perturbation term, i.e., the second term on RHS of eq. (28). Observe that

$$\begin{aligned}\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle &= \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Delta^{(t-1)\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \\ &- \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\ &- \left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \Delta^{(t-1)\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \end{aligned} \tag{31}$$

Then write

$$\begin{aligned}&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ \leq &\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Delta^{(t-1)\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ + &\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ + &\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \Delta^{(t-1)\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \end{aligned} \tag{32}$$

For the indicator function in the first term on the RHS of eq. (32), we further have the decomposition

$$\begin{aligned}&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Delta^{(t-1)\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ \leq &\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ + &\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top (\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ + &\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top (\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\ + &\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top (\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top (\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \end{aligned} \tag{33}$$

We again calculate the expectation of the first term on RHS of eq. (33) to utilize Markov inequality. Notice that $\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle = \left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1} \right\rangle$ and due the property of Gaussian random matrices (see [55]), $\mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a$ is independent of $\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}$.

Then we write $\boldsymbol{\mathcal{I}}_1 := \mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_{a\perp} \mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top} \mathbf{U}_a \boldsymbol{\Sigma}_a^{-1}$ and by the random matrix theory, we have that for any $u > 0$

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
\leq &\mathbb{P}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \boldsymbol{\mathcal{I}}_1\right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \,\middle|\, \|\boldsymbol{\mathcal{I}}_1\|_{\mathrm{F}}^2 \leq \frac{r}{\lambda^2}\left(\frac{d+u^2}{n_a^*}\right)^2\right) \\
&\cdot \mathbb{P}\left(\|\boldsymbol{\mathcal{I}}_1\|_{\mathrm{F}}^2 \leq \frac{r}{\lambda^2}\left(\frac{d+u^2}{n_a^*}\right)^2\right) + \exp(-u^2) \\
\leq &\exp\left(-c_4 \frac{\delta^2 \left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^4 \lambda^2\alpha^2 n^2}{rK^2(d^2+u^4)}\right) + \exp(-u^2)
\end{aligned}$$

for some absolute constant $c_4 > 0$. Choosing $u^2 = C_4\delta^{1/2}(\alpha n/K)^{1/2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$ we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
\leq &\exp\left(-c_4 \frac{\delta\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \alpha n}{\kappa_0^2 K r^2}\right) + \exp\left(-c_4 \frac{\delta^{1/2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \sqrt{\alpha n}}{\sqrt{K}}\right)
\end{aligned}$$

where the last inequality holds as $\lambda^2 \geq (1/4)\cdot r^{-1}\kappa_0^{-2}\max_{a,b\in[K],a\neq b}\|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}}^2$, $\lambda^2 \geq dK/(\alpha n)$, $\Delta^2 \gg rdK/n$, $\alpha n/K \gg \kappa_0^2 r^2$ and by setting $\delta \to 0$ sufficiently slow. Therefore, by Markov inequality, we get with probability at least $1 - \exp\big(-\delta^{1/2}(\alpha n/K)^{1/4}\Delta\big) - \exp\big(-\delta(\kappa_0 r)^{-1}(\alpha n/K)^{1/2}\Delta\big)$ that

$$\begin{aligned}
&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\quad \leq n\cdot\left[\exp\left(-\delta^{1/2}(\alpha n/K)^{1/2}\Delta^2\right) + \exp\left(-\delta(\kappa_0 r)^{-2}(\alpha n/K)\Delta^2\right)\right]
\end{aligned}$$

which holds as long as $\delta \to 0$ sufficiently slowly compared with $\alpha n/(K\kappa_0^2 r^2) \to \infty$.

It suffices to consider the remaining terms on RHS of eq. (33). For the second term of eq. (33), by Lemma 10.2 we have

$$\begin{aligned}
&\left\|\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\|^2 \\
&\leq C_5\frac{dK/(\alpha n)}{\lambda^2}\left[\frac{K}{\alpha n}\Delta^2 h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^2 n^2}\right]
\end{aligned}$$

for some absolute constant $C_5 > 0$, where we've used condition (10). Then

Lemma 10.4 implies

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot$$
$$\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq C_5\sum_{a=1}^{K}\sum_{b=1}^{K}\frac{r(dr+n)}{\delta^2\Delta^2}\left[\frac{K}{\alpha n}\Delta^2h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right]\frac{dK/(\alpha n)}{\lambda^2}$$
$$\leq\frac{1}{96}\ell(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$$

where the last inequality holds given $\Delta^2\gg\alpha^{-1}K^2r\,(dr/n+1)$, $\lambda^2\geq dK/(\alpha n)$. The same bound holds for the third term of eq. (33). For the last term, by Lemma 10.2, we obtain

$$\left\|\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\|^2$$
$$\leq C_6\frac{1}{\lambda^2}\left[\frac{\kappa_0^2\gamma^2K}{\alpha n}\lambda^2\Delta^2h_a^4(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^3(d^2+n^2)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^3n^3}\right]$$

for some absolute constant $C_6>0$. As a result, using (10)we have that

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2$$
$$\cdot\,\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq C_6\sum_{a=1}^{K}\sum_{b=1}^{K}\frac{r(dr+n)}{\delta^2\Delta^2}\frac{1}{\lambda^2}\left[\frac{K}{\alpha n}\lambda^2\Delta^2h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^3(d^2+n^2)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^3n^3}\right]$$
$$\leq\frac{1}{96}\ell(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$$

which holds for $\lambda^2\geq d/(\alpha n)$, $\Delta^2\gg\alpha^{-1}K^2r\,(dr/n+1)$.
We then consider the second term on the RHS of eq. (32), in which the indicator

function admits

$$
\begin{aligned}
&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\left(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\left(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}
\tag{34}
$$

The last three terms on RHS of (34) can be bounded exactly the same as those of (33), and we only need to treat the first term using Markov inequality. Then write

$$
\begin{aligned}
\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\right\rangle =& \left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\sum_{j\neq i}\mathbb{I}\left(s_j^*=a\right)\mathbf{E}_i\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\right\rangle\\
&+\left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^*=a\right)\mathbf{E}_i\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\right\rangle
\end{aligned}
$$

The bound for the first term above is the same as the first term of (33) and therefore we only consider the second term. By random matrix theory, there exists some absolute constant $C_7>0$ such that for any $u>0$:

$$
\mathbb{P}\left(\left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^*=a\right)\mathbf{E}_i\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\right\rangle \geq C_7\frac{r}{\lambda n_a^*}\left(d+u^2\right)\sqrt{\frac{d+u^2}{n_a^*}}\right)\leq\exp\left(-u^2\right)
$$

The we can choose $u^2=c_5\delta(\alpha n/K)^{1/3}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2$ to obtain

$$
\begin{aligned}
C_7\frac{r}{\lambda n_a^*}\left(d+u^2\right)\sqrt{\frac{d+u^2}{n_a^*}}\leq& C_7'\left(\frac{r}{\lambda}\left(\frac{dK}{\alpha n}\right)^{3/2}+\frac{\delta^{3/2}rK\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^3}{\lambda\alpha n}\right)\\
\leq&\delta\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2
\end{aligned}
$$

where the last inequality is due to $\lambda\geq d/(\alpha n)$, $\lambda\geq(1/2)\cdot r^{-1/2}\kappa_0^{-1}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}$, $\Delta^2\gg\alpha^{-1}K^{3/2}dr/n$ and $\alpha n/K\gg\kappa_0 r^{3/2}$. Hence we obtain that

$$
\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^*=a\right)\mathbf{E}_i\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\right\rangle \geq \frac{\delta}{2304}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
&\leq\exp\left(-c_5\frac{\delta\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2(\alpha n)^{1/3}}{K^{1/3}}\right)
\end{aligned}
$$

By Markov inequality, we get with probability at least $1-\exp\left(-\delta(\alpha n/K)^{1/6}\Delta\right)$ that

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot \mathbb{I}\left(\left\langle \mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a,\mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^*=a\right)\mathbf{E}_i\right)\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\right\rangle\geq\frac{\delta}{2304}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \leq n\cdot\exp\left(-\delta(\alpha n/K)^{1/3}\Delta^2\right)$$

It remains to consider the indicator function in the second term on the RHS of eq. (32), which reads

$$\begin{aligned}
&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\Delta^{(t-1)\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}\tag{35}$$

The treatment, as can be readily seen, is essentially the same as (33).
It suffices to bound the high-order perturbation term, i.e., the last term on RHS of eq. (28). Write

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle\geq\frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\rangle\geq\frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\langle\mathbf{E}_i,\mathbf{\Xi}_2\rangle\geq\frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}\tag{36}$$

where $\mathbf{\Xi}_2:=\sum_{k\geq3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*+\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{M}_a-\sum_{k\geq3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a$. By random matrix theory, the first term of on RHS of eq. (36) can be directly bounded such that for any $u>0$, there exists some absolute constant $C_8>0$ with

$$\mathbb{P}\left(\left\langle\mathbf{E}_i,\sum_{k\geq3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\rangle\geq C_8\frac{r}{\lambda^2}\sqrt{d+u^2}\left(\sqrt{\frac{d+u^2}{n_a^*}}\right)^3\right)\leq\exp\left(-u^2\right)$$

Choosing $u^2 = c_6\delta(\alpha n/K)^{1/4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$ for some absolute constant $c_6 > 0$, we have that

$$C_8\frac{r}{\lambda^2}\sqrt{d+u^2}\left(\sqrt{\frac{d+u^2}{n_a^*}}\right)^3 \le C_8'\left(\frac{rd^2K^{3/2}}{\lambda^2\alpha^{3/2}n^{3/2}} + \frac{\delta^2 rK\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^4}{\lambda^2\alpha n}\right)$$
$$\le \delta\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$$

where last inequality is due to $\lambda^2 \ge d/\sqrt{\alpha n}$, $\lambda^2 \ge (1/4)\cdot r^{-1}\kappa_0^{-2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$, $\Delta^2 \gg \alpha^{-1}K^{3/2}dr/n$ and $\alpha n/K \gg \kappa_0^2 r^2$. Hence we obtain that

$$\mathbb{P}\left(\left\langle \mathbf{E}_i, \sum_{k\ge 3}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\rangle \ge \frac{\delta}{192}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \le \exp\left(-c_6\frac{\delta\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2(\alpha n)^{1/4}}{K^{1/4}}\right)$$

By Markov's inequality, we have with probability greater than $1 - \exp\big(-\delta(\alpha n/K)^{1/8}\Delta\big)$ that

$$\sum_{i=1}^n\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\ge 3}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\rangle \ge \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\le n\cdot\exp\left(-\delta(\alpha n/K)^{1/4}\Delta^2\right)$$

For the second term on RHS of eq. (36), by carefully inspecting the perturbation term defined in Lemma 10.3 we obtain, there exists some absolute constant $C_9 > 0$ such that the following bound holds:

$$\|\mathbf{\Xi}_2\|^2 = \left\|\sum_{k\ge 3}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{M}_a - \sum_{k\ge 3}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\|^2$$
$$\le C_9\frac{1}{\lambda^4}\left[\frac{\kappa_0^4\gamma^2K^6}{\alpha^6n^6}\lambda^4\Delta^2h_a^6(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^6(d^3+n^3)h_a^3(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^6n^6}\right.$$
$$\left. + \left(\frac{dK}{\alpha n}\right)^2\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2h_a^2(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right)\right]$$

Using (10), $\lambda^2 \ge d/(\alpha n)$ and $\Delta^2 \gg \alpha^{-1}K^2r\,(dr/n+1)$., we have that

$$\sum_{i=1}^n\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\langle\mathbf{E}_i, \mathbf{\Xi}_2\rangle \ge \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\le C_9\sum_{a=1}^K\sum_{b=1}^K\frac{r(dr+n)}{\delta^2\Delta^2}\frac{1}{\lambda^4}\left[\frac{\kappa_0^4\gamma^2K^6}{\alpha^6n^6}\lambda^4\Delta^2h_a^6(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^6(d^3+n^3)h_a^3(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^6n^6}\right.$$
$$\left. + \left(\frac{dK}{\alpha n}\right)^2\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2h_a^2(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right)\right]$$
$$\le \frac{1}{64}\ell(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$$

Step 2.2.2: Treating the terms of $\langle \mathbf{E}_i, \mathbf{M}_a(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top)\rangle$. By symmetry, we can bound $\langle \mathbf{E}_i, \mathbf{M}_a(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top)\rangle$ the same way as $\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\rangle$ and hence the proof is omitted.

Step 2.2.3: Treating the terms of $\langle \mathbf{E}_i, (\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top)\mathbf{M}_a(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top)\rangle$. By Lemma 10.3, we obtain that

$$
\begin{aligned}
&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \geq \frac{\delta}{32}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\right)\\
\leq &\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\mathcal{S}_{\mathbf{M},1}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle \geq \frac{\delta}{128}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle \geq \frac{\delta}{128}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\mathcal{S}_{\mathbf{M},1}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle \geq \frac{\delta}{128}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \|\mathbf{M}_a - \mathbf{M}_{s_i^*}\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle \geq \frac{\delta}{128}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\right)
\end{aligned}
\tag{37}
$$

Since $\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\mathcal{S}_{\mathbf{M},1}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle = \langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{U}_a^\top\Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\rangle$, hence we can treat the first term on RHS of eq. (37) the same as the first term on the RHS of eq. (32). Moreever, since

$$
\begin{aligned}
&\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle\\
&= \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\rangle + \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_a\mathbf{V}_a^\top\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)}) - \mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\right\rangle\\
&+ \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\right\rangle
\end{aligned}
$$

From the above decomposition, it can be easily recognized that the above term is analogous to the last term on RHS of eq. (28), i.e., $\left\langle \mathbf{E}_i, \sum_{k\geq 3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle$. By symmetry, the term $\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\mathcal{S}_{\mathbf{M},1}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle$ can be bounded in a similar fashion and the details are omitted.

It remains to consider the last term on RHS of eq. (37). Observe that

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})\right\rangle\geq\frac{\delta}{128}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\left[\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\right.\\
&+\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
&+\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
&\left.+\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\right]
\end{aligned}\tag{38}$$

We then bound the expectation of the first term inside the bracket on RHS of eq. (38). By random matrix theory, there exists some absolute constant $C_{10}>0$ such that for any $u>0$:

$$\mathbb{P}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\rangle\geq C_{10}\frac{r}{\lambda^3}\sqrt{d+u^2}\left(\sqrt{\frac{d+u^2}{n_a^*}}\right)^4\right)\leq\exp\left(-u^2\right)$$

Choosing $u^2=c_7\delta(n/K)^{1/5}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2$ for some absolute constant $c_7>0$, we have that

$$\begin{aligned}
C_{10}\frac{r}{\lambda^3}\sqrt{d+u^2}\left(\sqrt{\frac{d+u^2}{n_a^*}}\right)^4&\leq C_{10}'\left(\frac{rd^{5/2}K^2}{\lambda^3\alpha^2n^2}+\frac{\delta^{5/2}rK^{3/2}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^5}{\lambda^3\alpha^{3/2}n^{3/2}}\right)\\
&\leq\delta\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2
\end{aligned}$$

where we've used $\lambda^2\geq d/\sqrt{\alpha n}$ and $\lambda\geq(1/2)\cdot r^{-1/2}\kappa_0^{-1}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}$, $\Delta^2\gg\alpha^{-1}K^2dr/n$ and $\alpha n/K\gg\kappa_0^2r^{5/3}$. Hence we obtain that

$$\mathbb{P}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\leq\exp\left(-c_7\frac{\delta\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\alpha^{1/5}n^{1/5}}{K^{1/5}}\right)$$

By Markov's inequality, we have with probability greater than $1-\exp\big(-\delta(\alpha n/K)^{1/10}\Delta\big)$ that

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\sum_{k\geq 3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\right\rangle\geq\frac{\delta}{512}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq n\cdot\exp\left(-\delta(\alpha n/K)^{1/5}\Delta^2\right)$$

For the last three terms inside the bracket on RHS of eq. (38), we have that

$$\begin{aligned}
&\left\|\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\right\|^2\\
+&\left\|\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right\|^2\\
+&\left\|\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a\sum_{k\geq 2}\left(\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\Delta^{(t-1)})-\mathcal{S}_{\mathbf{M},k}^{\mathbf{V}_a}(\bar{\mathbf{E}}_a^*)\right)\right\|^2\\
\leq&C_{11}\frac{1}{\lambda^6}\left[\frac{\kappa_0^6\gamma^2K^8}{\alpha^8n^8}\lambda^6\Delta^2h_a^8(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^8(d^4+n^4)h_a^4(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^8n^8}\right.\\
+&\left.\left(\frac{dK}{\alpha n}\right)^3\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2h_a^2(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)+\frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right)\right]
\end{aligned}$$

for some absolute constant $C_{11} > 0$. As a result, using (10) we obtain

$$
\begin{aligned}
&\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \\
&\cdot\Bigg[\mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\mathbf{M}_a \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\right\rangle \geq \frac{\delta}{512}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+\mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a \sum_{k\geq 2}\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right\rangle \geq \frac{\delta}{512}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+\mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\mathbf{M}_a \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\right\rangle \geq \frac{\delta}{512}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\Bigg] \\
&\leq C_{11}\sum_{a=1}^{K}\sum_{b=1}^{K}\frac{r(dr+n)}{\delta^2\Delta^2}\frac{1}{\lambda^6}\Bigg[\frac{\kappa_0^6\gamma^2K^8}{\alpha^8n^8}\lambda^6\Delta^2h_a^8(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^8(d^4+n^4)h_a^4(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^8n^8} \\
&+\left(\frac{dK}{\alpha n}\right)^3\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2h_a^2(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*) + \frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)}{\alpha^2n^2}\right)\Bigg] \\
&\leq \frac{1}{64}\ell(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)
\end{aligned}
$$

where the last inequality holds for $\lambda^2 \geq Kd/(\alpha n)$ and $\Delta^2 \gg \alpha^{-1}K^2r\,(dr/n+1)$.

Step 2.2.4: Treating the terms of $\left\langle \mathbf{E}_i, \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\Delta^{(t-1)}\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top\right\rangle$. The following decomposition is obvious:

$$
\begin{aligned}
&\left\langle \mathbf{E}_i, \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\Delta^{(t-1)}\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top\right\rangle \\
&= \left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{U}_a^\top(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \\
&+ \left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \\
&+ \left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{U}_a^\top(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \\
&+ \left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \quad (39)
\end{aligned}
$$

The first term above, i.e., $\left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{U}_a^\top(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle$, is essentially the same as $\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle$. For the second term of eq. (39), we further have

$$
\begin{aligned}
&\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \\
&= \left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle + \left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_a\mathbf{V}_a^\top\right\rangle
\end{aligned}
$$

Note that

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}})\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\
=&\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}})\mathbf{V}_a\mathbf{\Sigma}_a^{-1}\mathbf{U}_a^\top(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}})\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\
+&\left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}})^\top\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}})\mathbf{V}_a\mathbf{V}_a^\top \right\rangle
\end{aligned}$$

which can be decomposed of the same structure as the first term of eq. (31). On the other hand, we observe that

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta)(\bar{\mathbf{E}}^*_a + \Delta_{\mathbf{M}} + \Delta_{\mathbf{E}})\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\
=&\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}\left(\bar{\mathbf{E}}^*_a + \Delta_{\mathbf{M}} + \Delta_{\mathbf{E}}\right)\bar{\mathbf{E}}^*_a\mathbf{V}_a\mathbf{V}_a^\top \right\rangle + \left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}\left(\bar{\mathbf{E}}^*_a + \Delta_{\mathbf{M}} + \Delta_{\mathbf{E}}\right)(\Delta_{\mathbf{M}} + \Delta_{\mathbf{E}})\mathbf{V}_a\mathbf{V}_a^\top \right\rangle
\end{aligned}$$

which can be treated in the same manner as the last term on RHS of eq. (28). By symmetry, we can similarly treat the term $\left\langle \mathbf{E}_i, \mathbf{U}_a\mathbf{U}_a^\top(\bar{\mathbf{E}}^*_a + \Delta_{\mathbf{M}} + \Delta_{\mathbf{E}})\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle$. It suffices to consider the last term of eq. (39):

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \\
=&\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},1}(\Delta^{(t-1)}) \right\rangle \\
+&\left\langle \mathbf{E}_i, \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},1}(\Delta^{(t-1)})\left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\sum_{k\geq 2}\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) \right\rangle \\
+&\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},1}(\Delta^{(t-1)}) \right\rangle \\
+&\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right)\sum_{k\geq 2}\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) \right\rangle
\end{aligned}$$

Again, the first term above is analogous to the last term on RHS of eq. (28) and the second and third term are analogous to the term $\left\langle \mathbf{E}_i, \sum_{k\geq 2}\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\mathbf{M}_a\sum_{k\geq 2}\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)})\right\rangle$. It remains to consider the last term above, for which we have a further decom-

position:

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) \left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right) \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) \right\rangle \\
&= \left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\bar{\mathbf{E}}^*_a \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle + \left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \left(\Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right) \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle \\
&+ \left\langle \mathbf{E}_i, \sum_{k\geq 2} \left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\right) \left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right) \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle \\
&+ \left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right) \sum_{k\geq 2} \left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\right) \right\rangle \\
&+ \left\langle \mathbf{E}_i, \sum_{k\geq 2} \left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\right) \left(\bar{\mathbf{E}}^*_a + \Delta^{(t-1)}_{\mathbf{M}} + \Delta^{(t-1)}_{\mathbf{E}}\right) \sum_{k\geq 2} \left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\right) \right\rangle
\end{aligned} \tag{40}$$

For the first term above, by random matrix theory, there exists some absolute constant $C_{12} > 0$ such that for any $u > 0$:

$$\mathbb{P}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\bar{\mathbf{E}}^*_a \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle \geq C_{12} \frac{r}{\lambda^4} \sqrt{d+u^2} \left(\sqrt{\frac{d+u^2}{n^*_a}} \right)^5 \right) \leq \exp\left(-u^2\right)$$

Choosing $u^2 = c_8 \delta (\alpha n/K)^{1/6} \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2$ for some absolute constant $c_8 > 0$, we have that

$$\begin{aligned}
C_{12} \frac{r}{\lambda^4} \sqrt{d+u^2} \left(\sqrt{\frac{d+u^2}{n^*_a}} \right)^5 &\leq C'_{12} \left(\frac{r d^3 K^{5/2}}{\lambda^4 \alpha^{5/2} n^{5/2}} + \frac{\delta^3 r K^2 \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^6}{\lambda^4 \alpha^2 n^2} \right) \\
&\leq \delta \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2
\end{aligned}$$

where we've used $\lambda^2 \geq d/\sqrt{\alpha n}$ and $\lambda \geq (1/2) \cdot r^{-1/2} \kappa_0^{-1} \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}$, $\Delta^2 \gg K^{5/2} dr/(\alpha n)^{3/2}$ and $\alpha n/K \gg \kappa_0^2 r^{3/2}$. Hence we obtain that

$$\mathbb{P}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\mathbf{M}_a \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle \geq \frac{\delta}{512} \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \leq \exp\left(-c_7 \frac{\delta \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \alpha^{1/5} n^{1/5}}{K^{1/5}} \right)$$

By Markov's inequality, we have with probability greater than $1 - \exp\big(-\delta(\alpha n/K)^{1/12}\Delta\big)$ that

$$\begin{aligned}
\sum_{i=1}^n \sum_{a\in[K]\setminus\{s^*_i\}} \left\|\mathbf{M}_a - \mathbf{M}_{s^*_i}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a)\bar{\mathbf{E}}^*_a \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}^*_a) \right\rangle \geq \frac{\delta}{2560} \left\|\mathbf{M}_{s^*_i} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \right) \\
\leq n \cdot \exp\left(-\delta(\alpha n/K)^{1/6}\Delta^2\right)
\end{aligned}$$

For the last four terms on RHS of eq. (40), there exists some absolute constant $C_{12} > 0$ such that

$$
\begin{aligned}
&\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \\
&\Bigg[\mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\left(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right) \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right\rangle \geq \frac{\delta}{2560}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\Delta^{(t-1)} \sum_{k\geq 2} \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right\rangle \geq \frac{\delta}{2560}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2} \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\Delta^{(t-1)} \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\right\rangle \geq \frac{\delta}{2560}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\Delta^{(t-1)} \sum_{k\geq 2}\left(\mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\Delta^{(t-1)}) - \mathcal{S}^{\mathbf{V}_a}_{\mathbf{M},k}(\bar{\mathbf{E}}_a^*)\right)\right\rangle \geq \frac{\delta}{2560}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\Bigg] \\
&\leq C_{13}\sum_{a=1}^{K}\sum_{b=1}^{K} \frac{r(dr+n)}{\delta^2\Delta^2}\frac{1}{\lambda^8}\Bigg[\frac{\kappa_0^8\gamma^2K^{10}}{\alpha^{10}n^{10}}\lambda^8\Delta^2 h_a^{10}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^{10}(d^5+n^5)h_a^5(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^{10}n^{10}} \\
&+ \left(\frac{dK}{\alpha n}\right)^4\left(\frac{\gamma^2K^2}{\alpha^2n^2}\Delta^2 h_a^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^2(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^2n^2}\right)\Bigg] \\
&\leq \frac{1}{1024}\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)
\end{aligned}
$$

save that $\lambda^2 \geq Kd/(\alpha n)$, $\Delta^2 \gg \alpha^{-1}K^2r\,(dr/n+1)$ and (10) holds.
So far we finish the analysis of $\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ and by symmetry the term $\beta_{1,1}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ can be handled in a similar way.

Step 2.3: Bounding $\beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$. Recall the definition of $\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)})$, we have that

$$
\begin{aligned}
\beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) =& \sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
\leq& \sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_a - \widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&+ \sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}\left\|\mathbf{M}_a - \widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}} \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}
\tag{41}
$$

We need to bound three terms on RHS of eq. (41) separately. It follows that for some absolute constant $C_{14} > 0$:

$$\left\|\mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\|_{\mathrm{F}}^2 \leq C_{14}\left(\frac{\gamma^2 K^2}{\alpha^2 n^2}\Delta^2 h_{s_i^*}^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^2(d+n)h_{s_i^*}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^2 n^2} + \frac{dK}{\alpha n}\right)$$

Then for the first term on RHS of eq. (41), there exists some absolute constants $C_{15} > 0$:

$$\begin{aligned}
&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\leq C_{15}\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t)} \neq s_i^*\right) \max_{a\in[K]\setminus\{s_i^*\}} \frac{\frac{\gamma^4 K^4}{\alpha^4 n^4}\Delta^4 h_{s_i^*}^4(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^4(d^2+n^2)h_{s_i^*}^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^4 n^4} + \frac{d^2K^2}{\alpha^2 n^2}}{\delta^2\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2} \\
&\leq C_{15}\cdot h(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)\cdot \frac{\max_{b\in[K]}\left(\frac{\gamma^4 K^4}{\alpha^4 n^4}\Delta^4 h_b^4(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^4(d^2+n^2)h_b^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^4 n^4} + \frac{d^2K^2}{\alpha^2 n^2}\right)}{\delta^2\Delta^2} \\
&\leq \frac{1}{6}\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)
\end{aligned}$$

where in the last inequality we've used $\Delta^2 \gg \tau K/(\alpha n)$, $\Delta^2 \gg \alpha^{-1}K\left(d/n+1\right)$ and $\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \leq \tau$. Similarly, we can bound the second term on RHS of eq. (41) as

$$\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_a - \widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \leq \frac{1}{24}\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)$$

It remains to consider the last term on RHS of eq. (41), which has the following bound for some absolute constant $C_{16} > 0$:

$$\begin{aligned}
&\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}\left\|\mathbf{M}_a - \widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}} \\
\leq& C_{16}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}\left(\frac{\gamma K}{\alpha n}\Delta h_{s_i^*}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K\sqrt{(d+n)h_{s_i^*}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}}{\alpha n} + \sqrt{\frac{dK}{\alpha n}}\right)
\end{aligned}$$

Hence we can obtain that for some absolute constant $C_{17} > 0$:

$$\begin{aligned}
&\sum_{i=1}^n \sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} = a\right)\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}\left\|\mathbf{M}_a - \widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}} \geq \frac{\delta}{12}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
&\leq C_{17}\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t)} \neq s_i^*\right) \max_{a\in[K]\setminus\{s_i^*\}} \frac{\frac{\gamma^2 K^2}{\alpha^2 n^2}\Delta^2 h_{s_i^*}^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^2(d+n)h_{s_i^*}(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^2 n^2} + \frac{dK}{\alpha n}}{\delta^2} \\
&\leq C_{17}\cdot h(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)\cdot \frac{\max_{b\in[K]}\left(\frac{\gamma^2 K^2}{\alpha^2 n^2}\Delta^2 h_b^2(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) + \frac{K^2(d+n)h_b(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{\alpha^2 n^2} + \frac{dK}{\alpha n}\right)}{\delta^2} \\
&\leq \frac{1}{6}\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*)
\end{aligned}$$

provided that $\Delta^2 \gg \tau K/(\alpha n)$, $\Delta^2 \gg \alpha^{-1}K\,(d/n+1)$ and $\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \le \tau$.

Step 3: Obtaining contraction property. Collecting all pieces, we arrive at with probability at least $1-\exp(-\Delta)$:

$$\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*) \le n \exp\left(-(1-o(1))\frac{\Delta^2}{8}\right) + \frac{1}{4}\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t-1)}) + \frac{1}{2}\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$$

as $\alpha n/(K\kappa_0^2 r^2) \to \infty$. As a consequence, we obtain the contraction property (23).

To finish the proof, we use a mathematical induction step, which essentially requires the initialization conditions (10) and (22) hold. Notice that $\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le 2n \exp\left(-(1-o(1))\frac{\Delta^2}{8}\right) + \ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t-1)})/2 \le \tau$ as long as $\Delta^2 \gg \log(\tau/n)$. Moreover, we also have

$$h(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le \Delta^{-2}\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le \frac{\tau}{\Delta^2} = o\left(\frac{n}{K} \cdot \frac{\alpha}{\kappa_0 \vee \gamma^2}\right)$$

Hence the condition $\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le \tau$ and $h(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le (\kappa_0 \vee \gamma^2)^{-1}\alpha n/8K$ hold for all $t \ge 0$ and hence (23) holds for all $t \ge 1$. Using the relation $h(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \le \Delta^{-2}\ell(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ and the condition $\Delta^2 \gg \kappa_0^2 K\tau/(\alpha n)$, with probability greater than $1-\exp(-\Delta)$, for each $t \ge 0$ we have that

$$n^{-1} \cdot h(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}) \le \exp\left(-(1-o(1))\frac{\Delta^2}{8}\right) + 2^{-t}$$

The proof is completed by applying a union bound accounting for the events $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$.

10.2. Proof of Theorem 3.3

We first characterize the error of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ and without loss of generality, we only consider $\widehat{\mathbf{U}}$. Following the same argument in the proof of Theorem 1 in [64], one can obtain that there exists some absolute constant $c_0, C_0 > 0$ such that if $\sigma_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}})) \ge C_0(dr_{\mathbf{U}})^{1/2}n^{1/4}$, then with probability at least $1-\exp(-c_0(n \wedge d))$:

$$\left\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U}^*)\right\|_{\mathrm{F}} \le \frac{C(dr_{\mathbf{U}})^{1/2}\left[\sigma_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}})) + (dn)^{1/2}\right]}{\sigma^2_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}}))} \le \frac{1}{4\sqrt{2}}$$

We need the following lemma to relate $\sigma_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}}))$ to λ.

Lemma 10.5. *For $j \in \{1,2\}$, $\sigma_{\min}(\mathscr{M}_j(\boldsymbol{\mathcal{M}})) \ge \kappa_j^{-1}(Kr)^{-1/2}\sqrt{n}\lambda$*

By Lemma 10.5 and $r_{\mathbf{U}} \le Kr$, it suffices to recast the condition as $\lambda \ge C_0\kappa_1 rKd^{1/2}n^{-1/4}$. Combined with the bound for $\widehat{\mathbf{V}}$, we conclude that if $\lambda \ge C_0(\kappa_1 \vee \kappa_2)rKd^{1/2}n^{-1/4}$, then with probability at least $1-\exp(-c_0(n \wedge d))$:

$$\max\left\{\left\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U}^*)\right\|_{\mathrm{F}}, \left\|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V}^*)\right\|_{\mathrm{F}}\right\} \le \frac{1}{4\sqrt{2}} \tag{42}$$

Denote the above event by $\mathcal{Q}_{0,1}$ and we proceed by conditioning on $\mathcal{Q}_{0,1}$.
We then analyze the performance of spectral clustering based on $\widehat{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top$. Our proof is based on the proof for Lemma 4.2 in [43] with slight modification. Let $\boldsymbol{\mathcal{G}} := \boldsymbol{\mathcal{M}} \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top$ denote the signal part of $\widehat{\boldsymbol{\mathcal{G}}}$ (also $\mathbf{G} := \mathscr{M}_3(\boldsymbol{\mathcal{G}})$) and $\mathfrak{M} = [vec(\widehat{\mathbf{M}}_{\widehat{s}_1^{(0)}}), \cdots, vec(\widehat{\mathbf{M}}_{\widehat{s}_n^{(0)}})]^\top \in \mathbb{R}^{n\times d_1 d_2}$ denote the corresponding k-means solution. We claim the following lemma, whose proof is deferred to Section 11.

Lemma 10.6. *Suppose $\mathcal{Q}_{0,1}$ holds. Then we have the following facts:*

(I) $\mathfrak{M}$, the k-means solution, is close $\mathbf{G}$, i.e., there exists some absolute constants $c_0, C_0 > 0$ such that with probability at least $1 - \exp(-c_0 d)$:

$$\|\mathfrak{M} - \mathbf{G}\|_{\mathrm{F}} \le C_0\sqrt{K}\left(\sqrt{dKr + n}\right)$$

(II) The rows of $\mathbf{G}$ belonging to different clusters is well-separated, i.e.

$$\left\|\boldsymbol{\mathcal{G}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right\|_{\mathrm{F}} \ge \frac{\Delta}{2}$$

for any $i, j \in [n], s_i^* \neq s_j^*$.

We proceed by conditioning on $\mathcal{Q}_{0,2} := \{\text{(II) holds}\}$. Define the following set

$$S = \left\{ i \in [n] : \|[\mathfrak{M}]_{i\cdot} - [\mathbf{G}]_{i\cdot}\| \ge \frac{\Delta}{4} \right\}$$

Then by construction we have

$$|S| \le \frac{\|\mathfrak{M} - \mathbf{G}\|_{\mathrm{F}}^2}{(\Delta/4)^2} \le \frac{\alpha n}{2K}$$

where the last inequality is due to the condition $\Delta^2 \ge 32C_0^2\alpha^{-1}K^2\,(dKr/n + 1)$. We claim that all indices in S^c are correctly clustered. To see this, let

$$N_k = \{i \in [n] : s_i^* = k, i \in S^c\}$$

The following two facts hold:

- For each $k \in [K]$, $|N_k| \ge n_k^* - |S| \ge \alpha n/(2K) > 0$
- For each pair $a, b \in [K], a \neq b$, there cannot exist some $i \in N_a$ and $j \in N_b$ such that $\widehat{s}_i^{(0)} = \widehat{s}_j^{(0)}$. Otherwise we have $\widehat{\mathbf{M}}_{\widehat{s}_i^{(0)}} = \widehat{\mathbf{M}}_{\widehat{s}_j^{(0)}}$ and it follows that

$$\begin{aligned}\|[\mathbf{G}]_{i\cdot} - [\mathbf{G}]_{j\cdot}\| &\le \|[\mathbf{G}]_{i\cdot} - [\mathfrak{M}]_{i\cdot}\| + \|[\mathfrak{M}]_{i\cdot} - [\mathfrak{M}]_{j\cdot}\| + \|[\mathfrak{M}]_{j\cdot} - [\mathbf{G}]_{j\cdot}\| \\ &< \frac{\Delta}{2}\end{aligned}$$

which contradicts (II).

The above two facts imply that sets $\{\widehat{s}_i^{(0)} : i \in N_k\}$ are disjoint for all $k \in [K]$. Therefore, there exists a permutation π such that $\sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(0)} \neq \pi(s_i^*)\right) = 0$, i.e., indices in S^c are correctly clustered. As a consequence,

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \leq n^{-1} \cdot |S| \leq \frac{16C_0^2 K}{\Delta^2}\left(\frac{dKr}{n} + 1\right)$$

The proof is completed by taking union bound over $\mathcal{Q}_0^c := \mathcal{Q}_{0,1}^c \bigcup \mathcal{Q}_{0,2}^c$.

10.3. Proof of Theorem 3.4

We essentially follow a similar argument of [17]. Without loss of generality we assume $\|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}} = \Delta$. Consider the $\mathbf{s}^* \in [K]^n$ such that $n_1^* \leq n_2^* \leq \cdots \leq n_K^*$ and $n_1^* = n_2^* = \lfloor \alpha n/K \rfloor$. For every $k \in [K]$, we can choose a subset $\mathfrak{N}_k \subset \{i \in [n] : s_i^* = k\}$ with cardinality $\lceil n_k^* - \frac{\alpha n}{4K^2} \rceil$. And let $\mathfrak{N} = \bigcup_{k=1}^K \mathfrak{N}_k$ denote the collection of samples in $\mathfrak{N}_k$'s. Define the following parameter space for $\mathbf{s}$:

$$\mathbf{S}^* = \{\mathbf{s} \in [K]^n : s_i = s_i^* \text{ for } i \in \mathfrak{N}\}$$

For any two $\mathbf{s}, \mathbf{s}' \in \mathbf{S}^*$ such that $\mathbf{s} \neq \mathbf{s}'$, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(s_i \neq s_i') \leq \frac{K}{n}\frac{\alpha n}{4K^2} = \frac{\alpha}{4K}$$

Meanwhile, for any permutation $\pi \neq \mathrm{Id}$ from $[K]$ to $[K]$, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\pi(s_i) \neq s_i') \geq \frac{K}{n}\left(\frac{\alpha n}{K} - \frac{\alpha n}{4K^2}\right) \geq \frac{3\alpha}{4K}$$

Therefore, we conclude that $h_c(\mathbf{s}, \mathbf{s}') = h(\mathbf{s}, \mathbf{s}') = \sum_{i=1}^n \mathbb{I}(s_i \neq s_i')$ for any $\mathbf{s}, \mathbf{s}' \in \mathbf{S}^*$. Define the parameter space

$$\Omega(d_1, d_2, n, K, \alpha) = \Big\{(\{\mathbf{M}_k\}_{k=1}^K, \mathbf{s}) : \ \mathbf{M}_k \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(\mathbf{M}_k) = r_k, \forall k \in [K], \mathbf{s} \in [K]^n, \min_{k \in [K]} |\{i \in [n] : s_i = k\}| \geq \alpha n/K, \min_{a \neq b} \|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}} \geq \Delta\Big\}$$

and

$$\Omega_0(d_1, d_2, n, K, \alpha) = \Big\{(\{\mathbf{M}_k\}_{k=1}^K, \mathbf{s}) : \ \mathbf{M}_k \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(\mathbf{M}_k) = r_k, \forall k \in [K], \mathbf{s} \in \mathbf{S}^*, \min_{k \in [K]} |\{i \in [n] : s_i = k\}| \geq \alpha n/K, \min_{a \neq b} \|\mathbf{M}_a - \mathbf{M}_b\|_{\mathrm{F}} \geq \Delta\Big\}$$

Since $\Omega_0 \subset \Omega$, we have

$$\inf_{\widehat{\mathbf{s}}} \sup_{\Omega} \mathbb{E} h_c(\widehat{\mathbf{s}}, \mathbf{s}) \geq \inf_{\widehat{\mathbf{s}}} \sup_{\Omega_0} \mathbb{E} h_c(\widehat{\mathbf{s}}, \mathbf{s}) \geq \inf_{\widehat{\mathbf{s}}} \frac{1}{|\mathbf{S}^*|}\sum_{\mathbf{s} \in \mathbf{S}^*} \mathbb{E} h_c(\widehat{\mathbf{s}}, \mathbf{s}) \geq \sum_{i \in \mathfrak{N}^c} \inf_{\widehat{s}_i} \frac{1}{|\mathbf{S}^*|}\sum_{\mathbf{s} \in \mathbf{S}^*} \mathbb{P}(\widehat{s}_i \neq s_i) \tag{43}$$

where we consider a uniform prior on $\mathbf{S}^*$ and hence the second inequality holds as minimax risk is lower bounded by Bayes risk, and the last inequality holds since the infimum can be taken over all $\widehat{\mathbf{s}}$ such that $\widehat{s}_i = s_i^*$ for $i \in \mathfrak{N}$. Then it suffices to consider $\inf_{\widehat{s}_i} \frac{1}{|\mathbf{S}^*|}\sum_{\mathbf{s}\in\mathbf{S}^*}\mathbb{P}(\widehat{s}_i \neq s_i)$ for $i \in \mathfrak{N}^c$. Without loss generality, we assume $1 \in \mathfrak{N}^c$ and for any $k \in [K]$ we denote $\mathbf{S}_k^* = \{\mathbf{s} \in \mathbf{S}^* : s_1 = k\}$. It's obvious that $\mathbf{S}^* = \bigcup_{k=1}^K \mathbf{S}_k^*$ and $\mathbf{S}_a^* \bigcap \mathbf{S}_b^* = \phi$ for $a \neq b$. In addition, by the definition of such partition, for any $a \neq b \in [K]$ and $\mathbf{s} \in \mathbf{S}_a^*$, there exists a unique $\mathbf{s}' \in \mathbf{S}_b^*$ such that $s_i = s_i'$ for all $i \neq 1$, which implies that $|\mathbf{S}_a^*| = |\mathbf{S}_b^*|$ for all $a, b \in [K]$. Then we have

$$\begin{aligned}
\inf_{\widehat{s}_1}\frac{1}{|\mathbf{S}^*|}\sum_{\mathbf{s}\in\mathbf{S}^*}\mathbb{P}(\widehat{s}_1 \neq s_1) &= \inf_{\widehat{s}_1}\frac{1}{|\mathbf{S}^*|}\frac{1}{K-1}\sum_{a<b}\left(\sum_{\mathbf{s}\in\mathbf{S}_a^*}\mathbb{P}(\widehat{s}_1 \neq a) + \sum_{\mathbf{s}\in\mathbf{S}_b^*}\mathbb{P}(\widehat{s}_1 \neq b)\right) \\
&\geq \frac{1}{K(K-1)}\sum_{a<b}\inf_{\widehat{s}_1}\left(\frac{1}{|\mathbf{S}_a^*|}\sum_{\mathbf{s}\in\mathbf{S}_a^*}\mathbb{P}(\widehat{s}_1 \neq a) + \frac{1}{|\mathbf{S}_b^*|}\sum_{\mathbf{s}\in\mathbf{S}_b^*}\mathbb{P}(\widehat{s}_1 \neq b)\right) \\
&\geq \frac{1}{K(K-1)}\inf_{\widehat{s}_1}\left(\frac{1}{|S_1^*|}\sum_{\mathbf{s}\in S_1^*}\mathbb{P}(\widehat{s}_1 \neq 1) + \frac{1}{|S_2^*|}\sum_{\mathbf{s}\in S_2^*}\mathbb{P}(\widehat{s}_1 \neq 2)\right) \\
&\geq \frac{1}{K(K-1)}\frac{1}{|\mathbf{S}_{-1}^*|}\sum_{\mathbf{s}_{-1}\in\mathbf{S}_{-1}^*}\inf_{\widehat{s}_1}\left(\mathbb{P}_{\mathbf{s}=(1,\mathbf{s}_{-1})}(\widehat{s}_1 \neq 1) + \mathbb{P}_{\mathbf{s}=(2,\mathbf{s}_{-1})}(\widehat{s}_1 \neq 2)\right) \\
&\geq \frac{1}{K(K-1)}\inf_{\widehat{s}_1}\left(\mathbb{P}_{H_0^{(1)}}(\widehat{s}_1 = 2) + \mathbb{P}_{H_1^{(1)}}(\widehat{s}_1 = 1)\right) \quad (44)
\end{aligned}$$

where $\mathbf{S}_{-1}^*$ is the collection of the subvectors in $\mathbf{S}^*$ excluding the first coordinate, and we define a simple hypothesis testing for each $i \in [n]$:

$$H_0^{(i)} : s_i = 1 \quad \text{vs.} \quad H_1^{(i)} : s_i = 2$$

Hence in (44), we have the form of Type-I error + Type-II error of the above test. Notice that $|\{i \in [n] : s_i^* = k\}\backslash\mathfrak{N}_k| \geq \lfloor \alpha n/(4K^2)\rfloor$ and hence $|\mathfrak{N}^c| \geq c_0\alpha n/K$ for some constant $c_0 > 0$. Combining this with (43), (44), we proceed that

$$\inf_{\widehat{\mathbf{s}}}\sup_{\Omega}\mathbb{E}h_c(\widehat{\mathbf{s}}, \mathbf{s}) \geq c_0\frac{\alpha n}{K^3}\frac{1}{|\mathfrak{N}^c|}\sum_{i\in\mathfrak{N}^c}\inf_{\widehat{s}_i}\left(\mathbb{P}_{H_0^{(l)}}(\widehat{s}_i = 2) + \mathbb{P}_{H_1^{(l)}}(\widehat{s}_i = 1)\right)$$

According to the Neyman-Pearson lemma, for each $i \in [n]$, the optimal test of $H_0^{(l)}$ vs. $H_1^{(l)}$ is given by the likelihood ratio test with threshold 1. Let $p_0(\mathbf{X}_i)$ and $p_1(\mathbf{X}_i)$ denote the likelihood of $\mathbf{X}_i$ under H_0 and H_1, respectively. Then $\frac{p_1(\mathbf{X}_i)}{p_0(\mathbf{X}_i)} = \frac{\exp(\|\mathbf{X}_i-\mathbf{M}_1\|_{\mathrm{F}}^2/2)}{\exp(\|\mathbf{X}_i-\mathbf{M}_2\|_{\mathrm{F}}^2/2)}$ and hence the infimum is achieved by $\widehat{s}_i = \arg\min_{k\in\{1,2\}}\|\mathbf{X}_i - \mathbf{M}_k\|_{\mathrm{F}}^2$.

Therefore,

$$\begin{aligned}
&\inf_{\widehat{s}_i}\left(\frac{1}{2}\mathbb{P}_{H_0^{(l)}}(\widehat{s}_i=2)+\frac{1}{2}\mathbb{P}_{H_1^{(l)}}(\widehat{s}_i=1)\right)\\
&=\frac{1}{2}\left(\mathbb{P}\left(\|\mathbf{M}_1+\mathbf{E}_i-\mathbf{M}_2\|_{\mathrm{F}}^2\le\|\mathbf{E}_i\|_{\mathrm{F}}^2\right)+\mathbb{P}\left(\|\mathbf{M}_2+\mathbf{E}_i-\mathbf{M}_1\|_{\mathrm{F}}^2\le\|\mathbf{E}_i\|_{\mathrm{F}}^2\right)\right)\\
&=\frac{1}{2}\left(\mathbb{P}\left(\frac{1}{2}\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}^2\le\langle\mathbf{M}_2-\mathbf{M}_1,\mathbf{E}_i\rangle\right)+\mathbb{P}\left(\frac{1}{2}\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}^2\le\langle\mathbf{M}_1-\mathbf{M}_2,\mathbf{E}_i\rangle\right)\right)
\end{aligned}$$

Notice that $\langle\mathbf{M}_2-\mathbf{M}_1,\mathbf{E}_i\rangle\overset{\mathrm{d}}{=}\langle\mathbf{M}_1-\mathbf{M}_2,\mathbf{E}_i\rangle\overset{\mathrm{d}}{=}\mathcal{N}(0,\sigma^2\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}^2)$, we can proceed as

$$\inf_{\widehat{s}_i}\left(\frac{1}{2}\mathbb{P}_{H_0^{(l)}}(\widehat{s}_i=2)+\frac{1}{2}\mathbb{P}_{H_1^{(l)}}(\widehat{s}_i=1)\right)\ge\frac{\sigma}{\sqrt{2\pi}\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}}\exp\left(-\frac{\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}^2}{8\sigma^2}\right)$$

where the inequality holds as $\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}/\sigma\ge 1$. Hence we conclude that

$$\inf_{\widehat{\mathbf{s}}}\sup_{\Omega}\mathbb{E}n^{-1}\cdot h_c(\widehat{\mathbf{s}},\mathbf{s})\ge\exp\left(-\frac{\Delta^2}{8\sigma^2}-C\log\frac{\Delta K}{\alpha\sigma}\right)=\exp\left(-(1+o(1))\frac{\Delta^2}{8\sigma^2}\right)$$

provided that $\frac{\Delta^2}{\sigma^2\log(K/\alpha)}\to\infty$.

10.4. Proof of Theorem 4.4

Suppose we are given the data $\{\mathbf{X}_i\}_{i=1}^n$ generated by eq:rank-one-model with $((1-\epsilon)\mathbf{M},\mathbf{s}^*)\in\widetilde{\Omega}_{\lambda_*^{(n)}}$ for any $\epsilon\in(0,1]$. We utilize the sample splitting trick, similar to that in Theorem 2.4 in [42], to generate two independent copies $\{\mathbf{X}_i^{(1)}\}_{i=1}^n$ and $\{\mathbf{X}_i^{(2)}\}_{i=1}^n$ by

$$\mathbf{X}_i^{(1)}=\frac{\mathbf{X}_i+\epsilon^{-1}\widetilde{\mathbf{E}}_i}{\sqrt{1+\epsilon^{-2}}},\quad \mathbf{X}_i^{(2)}=\frac{\mathbf{X}_i-\epsilon\widetilde{\mathbf{E}}_i}{\sqrt{1+\epsilon^2}}$$

for $i=1,\cdots,n$ where $\{\widetilde{\mathbf{E}}_i\}_{i=1}^n$ are Gaussian noise matrices independent of $\{\mathbf{E}_i\}_{i=1}^n$. As a consequence, we have $\mathbf{X}_i^{(1)}=\frac{s_i^*\mathbf{M}}{\sqrt{1+\epsilon^{-2}}}+\mathbf{E}_i^{(1)}$ and $\mathbf{X}_i^{(2)}=\frac{s_i^*\mathbf{M}}{\sqrt{1+\epsilon^2}}+\mathbf{E}_i^{(2)}$ with $\mathbf{E}_i^{(1)}=\frac{\mathbf{E}_i+\epsilon^{-1}\widetilde{\mathbf{E}}_i}{\sqrt{1+\epsilon^{-2}}}$ and $\mathbf{E}_i^{(2)}=\frac{\mathbf{E}_i-\epsilon\widetilde{\mathbf{E}}_i}{\sqrt{1+\epsilon^2}}$. Due to the property of Gaussian, $\{\mathbf{E}_i^{(1)}\}_{i=1}^n$ and $\{\mathbf{E}_i^{(2)}\}_{i=1}^n$ are independent. We define the following test statistic:

$$T_n=\left\|\sum_{i=1}^n\frac{\widehat{s}_i\mathbf{X}_i^{(1)}}{n}\right\|$$

where $(\widehat{s}_1,\cdots,\widehat{s}_n)=\widehat{\mathbf{s}}_{\mathsf{comp}}(\boldsymbol{\mathcal{X}}^{(2)})$ with $\boldsymbol{\mathcal{X}}^{(2)}$ being the data tensor by stacking $\left\{\mathbf{X}_i^{(2)}\right\}_{i=1}^n$. By construction, $\{\widehat{s}_i\}_{i=1}^n$ is independent of $\left\{\mathbf{E}_i^{(1)}\right\}_{i=1}^n$ and hence

$\sum_{i=1}^n \frac{\widehat{s}_i \mathbf{E}_i^{(1)}}{n} \overset{d}{=} \sum_{i=1}^n \frac{\mathbf{E}_i^{(1)}}{n}$. Under H_0, with probability at least $1-\exp(-d)$:

$$T_n = \left\| \sum_{i=1}^n \frac{\widehat{s}_i \mathbf{X}_i^{(1)}}{n} \right\| \leq \frac{C_0}{2}\sqrt{\frac{d}{n}}$$

for some absolute constant $C_0 > 0$. Under H_1, we have $((1+\epsilon^2)^{-1/2}\mathbf{M}, \mathbf{s}^*) \in \widetilde{\Omega}_{\lambda_*^{(n)}}$ since $(1-\epsilon) \leq (1+\epsilon^2)^{-1/2}$. By (16) we have that with probability greater than $1-\zeta_n$:

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}_{\mathsf{comp}}, \mathbf{s}^*) \leq \delta_n \tag{45}$$

Without loss of generality we assume $h_c(\widehat{\mathbf{s}}_{\mathsf{comp}}, \mathbf{s}^*) = h(\widehat{\mathbf{s}}_{\mathsf{comp}}, \mathbf{s}^*)$. Hence we can obtain with probability at least $1-\zeta_n-\exp(-d)$:

$$\begin{aligned} T_n &\geq \left\| \sum_{i=1}^n \frac{\widehat{s}_i s_i^*}{n\sqrt{1+\epsilon^{-2}}} \mathbf{M} \right\| - \left\| \sum_{i=1}^n \frac{\widehat{s}_i \mathbf{E}_i^{(1)}}{n} \right\| \\ &\geq \frac{\lambda_*^{(n)}(1-2n^{-1}h(\widehat{\mathbf{s}}_{\mathsf{comp}}, \mathbf{s}^*))}{\sqrt{1+\epsilon^{-2}}} - \frac{C_0}{2}\sqrt{\frac{d}{n}} \\ &> \frac{C_0}{2}\sqrt{\frac{d}{n}} \end{aligned}$$

where we've used (45) and $\lambda_*^{(n)} > C_0(1-2\delta_n)^{-1}\sqrt{1+\epsilon^{-2}}d^{1/2}n^{-1/2}$ in the last inequality. Then the test ϕ_n can be defined as

$$\phi_n(\boldsymbol{\mathcal{X}}) = \begin{cases} 1 & \text{if } T_n > C_0\sqrt{\frac{d}{n}}, \\ 0 & \text{otherwise.} \end{cases}$$

It turns out that

$$\mathbb{E}_{Q_n}[\phi_n(\boldsymbol{\mathcal{X}})] + \sup_{((1-\epsilon)\mathbf{M}, \mathbf{s}^*) \in \widetilde{\Omega}_{\lambda_*^{(n)}}} \mathbb{E}_{(\mathbf{M}, \mathbf{s}^*)}[1-\phi_n(\boldsymbol{\mathcal{X}})] \leq \zeta_n + \exp(-d)$$

Notice that computing T_n requires only $poly(d, n)$ and the proof is completed by setting $n, d \to \infty$.

10.5. Proof of Theorem 5.2

Theorem 5.2 can be obtained by modifying the proofs of Theorem 3.3 and Theorem 3.1, and hence we only sketch the necessary modifications here. For notational simplicity, we denote the smallest non-trivial singular value of $\mathbf{M}_1$ as λ_1 and the condition number of $\mathbf{M}_1$ as κ_0. We use C and c to represent generic absolute constants, whose actual values may vary in different formulas.

We consider the spectral initialization in Algorithm 3. Denote the following decomposition of tensor $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{M}}_1 + \boldsymbol{\mathcal{M}}_2$, where for $k \in [2]$, the i-th slice of $\boldsymbol{\mathcal{M}}_k$ is defined as $[\boldsymbol{\mathcal{M}}_k]_{\cdot\cdot i} = \mathbb{I}(\mathbf{s}_i^* = k)\mathbf{M}_k$. It turns out that $\mathbf{U}_1$ is the leading-r_1

left singular vectors of $\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)$ and $\mathbf{V}_1$ is the leading-r_1 left singular vectors of $\mathscr{M}_2(\boldsymbol{\mathcal{M}}_1)$. We first show that $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{V}}_1$ are close to $\mathbf{U}_1$ and $\mathbf{V}_1$, respectively. Without loss of generality, we only consider $\widehat{\mathbf{U}}_1$. A key observation is that $\widehat{\mathbf{U}}_1$ is also the leading-r_1 left eigenvectors $\mathscr{M}_1(\boldsymbol{\mathcal{X}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{X}})$. Then write

$$\begin{aligned}\mathscr{M}_1(\boldsymbol{\mathcal{X}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{X}}) &= \mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}) + \mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) + \mathscr{M}_1(\boldsymbol{\mathcal{E}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}) + \mathscr{M}_1(\boldsymbol{\mathcal{E}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) \\ &= \mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_2) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1) \\ &\quad + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_2) + [\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)]\,\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) \\ &\quad + \mathscr{M}_1(\boldsymbol{\mathcal{E}})\,[\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)]^\top + \mathscr{M}_1(\boldsymbol{\mathcal{E}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) \qquad (46)\end{aligned}$$

We are going to bound each term on RHS of eq. (46). The first term $\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1)$ is the signal part and we have

$$\sigma_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1)) = \sigma_{r_1}(\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1)) \geq n_1^*\lambda_1^2$$

For the 2nd, 3rd and 4th term of (46), we can have

$$\left\|\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_2) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_1)\right\| \leq 2\kappa_0\sqrt{n_1^*n_2^*}\lambda_1\left\|\mathbf{M}_2\right\|$$

$$\left\|\mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}_2)\right\| \leq n_2^*\left\|\mathbf{M}_2\right\|^2$$

The 5th and 6th term of eq. (46) can be together bounded as

$$\left\|[\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)]\,\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) + \mathscr{M}_1(\boldsymbol{\mathcal{E}})\,[\mathscr{M}_1(\boldsymbol{\mathcal{M}}_1) + \mathscr{M}_1(\boldsymbol{\mathcal{M}}_2)]^\top\right\| \leq C\left(\kappa_0\sqrt{n_1^*}\lambda_1 + \sqrt{n_2^*}\left\|\mathbf{M}_2\right\|\right)\sqrt{d}$$

with probability at least $1-\exp(-cd)$, for some absolute constant $c, C > 0$. Lastly, we notice that $\mathbb{E}\left(\mathscr{M}_1(\boldsymbol{\mathcal{E}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}})\right) = nd_2\mathbf{I}_{d_1}$, then by [33], with probability at least $1-\exp(-d)$ we have

$$\left\|\mathscr{M}_1(\boldsymbol{\mathcal{E}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{E}}) - nd_2\mathbf{I}_{d_1}\right\| \leq C\sqrt{n}d$$

Note that $n_2^*/n_1^* \leq 2(1-\alpha/2)/\alpha \leq 2\alpha^{-1}$. Collecting all pieces above, if $\lambda_1 \geq \kappa_0\alpha^{-1/2}\left\|\mathbf{M}_2\right\|$ and

$$\lambda_1 \geq C\left(\kappa_0\alpha^{-1/2}\sqrt{\frac{d}{n}} + \alpha^{-1/2}\frac{d^{1/2}}{n^{1/4}}\right)$$

for some large constant $C > 0$, which is satisfied by Assumption 5.1, then with probability greater than $1-\exp(-cd)$ we can have $\left\|\widehat{\mathbf{U}}_1\widehat{\mathbf{U}}_1^\top - \mathbf{U}_1\mathbf{U}_1^\top\right\| \leq 1/4$. Using same analysis on $\widehat{\mathbf{V}}_1$, we can conclude with probability at least $1-\exp(-cd)$:

$$\max\left\{\left\|\widehat{\mathbf{U}}_1\widehat{\mathbf{U}}_1^\top - \mathbf{U}_1\mathbf{U}_1^\top\right\|, \left\|\widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1^\top - \mathbf{V}_1\mathbf{V}_1^\top\right\|\right\} \leq \frac{1}{6} \qquad (47)$$

Define $\widehat{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}_1\widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1^\top$, $\boldsymbol{\mathcal{G}} := \boldsymbol{\mathcal{M}} \times_1 \widehat{\mathbf{U}}_1\widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{V}}_1\widehat{\mathbf{V}}_1^\top$ (also $\mathbf{G} := \mathscr{M}_3(\boldsymbol{\mathcal{G}})$) and $\mathfrak{M} := [vec(\widehat{\mathbf{M}}_{\widehat{s}_1^{(0)}}), \cdots, vec(\widehat{\mathbf{M}}_{\widehat{s}_n^{(0)}})]^\top \in \mathbb{R}^{n\times d_1d_2}$. We can have the following lemma, which is an analogue to Lemma 10.6.

Lemma 10.7. *Suppose (47) holds. Then we have the following facts:*

(I) $\mathfrak{M}$*, the k-means solution, is close* $\mathbf{G}$*, i.e., there exists some absolute constants* $c_0, C_0 > 0$ *such that with probability at least* $1 - \exp(-c_0 d)$*:*

$$\|\mathfrak{M} - \mathbf{G}\|_{\mathrm{F}} \leq C_0 \left(\sqrt{dr_1 + n}\right)$$

(II) The rows of $\mathbf{G}$ *belonging to different clusters is well-separated, i.e.*

$$\left\|\boldsymbol{\mathcal{G}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right\|_{\mathrm{F}} \geq \frac{\Delta}{2}$$

for any $i, j \in [n], s_i^* \neq s_j^*$.

Following the almost identical argument in the proof of Theorem 3.3 but replacing $\widehat{\mathbf{U}}$ with $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{V}}$ with $\widehat{\mathbf{V}}_1$, then under Assumption 5.1, with probability at least $1 - \exp(-c(n \wedge d))$ we have

$$n^{-1} \cdot h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \leq \frac{C}{\Delta^2}\left(\frac{dr_1}{n} + 1\right) = o\left(\frac{\alpha}{\kappa_0}\right) \tag{48}$$

if $\Delta^2 \gg \kappa_0 \alpha^{-1}(dr_1/n + 1)$. Note that Assumption 5.1 and the condition in Theorem 5.2 already imply that

$$\Delta^2 \geq C\alpha^{-1}\left(\kappa_0^2 \frac{dr_1}{n} + \frac{dr_1}{\sqrt{n}}\right)$$

Then if $n/\kappa_0^2 \to \infty$ and $\alpha\Delta^2/\kappa_0 \to \infty$, the condition $\Delta^2 \gg \kappa_0\alpha^{-1}(dr_1/n + 1)$ automatically holds and we have the following holds with probability at least $1 - \exp(-c(n \wedge d))$:

$$\ell(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) \leq \Delta^2 h_c(\widehat{\mathbf{s}}^{(0)}, \mathbf{s}^*) = o\left(\frac{\alpha n \Delta^2}{\kappa_0}\right)$$

which is an analogue to (22).

We then consider the iterative convergence of Algorithm 3. Following the same argument of Step 2 in the proof of Theorem 3.1 line by line and adopting the same notation therein, we have the following inequality:

$$\ell(\widehat{\mathbf{s}}^{(t)}, \mathbf{s}^*) \leq \xi_{\mathsf{err}} + \beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) + \beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$$

We can bound ξ_{err} the same as **Step 2.1** in the proof of Theorem 3.1. To bound $\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$, it turns out that, by symmetry, we only need to bound

$$\begin{aligned}\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := &\sum_{i=1}^n \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq 1\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_1^{(t)} - \mathbf{M}_1 \right\rangle \geq \frac{\delta}{8}\|\mathbf{M}_2 - \mathbf{M}_1\|_{\mathrm{F}}^2\right) \\ &+ \sum_{i=1}^n \|\mathbf{M}_2 - \mathbf{M}_1\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq 2\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_2^{(t)} - \mathbf{M}_2 \right\rangle \geq \frac{\delta}{8}\|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}}^2\right)\end{aligned} \tag{49}$$

The argument in **Step 2.2** in the proof of Theorem 3.1 can be directly applied to the analysis of $\widehat{\mathbf{M}}_1 - \mathbf{M}_1$, i.e., the first term on RHS of eq. (49), whereas it fails for $\widehat{\mathbf{M}}_2 - \mathbf{M}_2$ since $\sigma_{\min}(\mathbf{M}_2)$ *can* be arbitrarily close to 0 and Lemma 10.3 no longer holds. Observe that

$$\begin{aligned}
\widehat{\mathbf{M}}_2^{(t)} &= \widehat{\mathbf{U}}_2\widehat{\mathbf{U}}_2^\top \left(\frac{1}{n_2^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = 2\right) \mathbf{M}_{s_i^*} + \bar{\mathbf{E}}_2^{(t-1)} \right) \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2^\top \\
&= \widehat{\mathbf{U}}_2\widehat{\mathbf{U}}_2^\top \left[\mathbf{M}_2 + \frac{1}{n_a^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = 2\right) (\mathbf{M}_{s_i^*} - \mathbf{M}_2) + \bar{\mathbf{E}}_2^* + (\bar{\mathbf{E}}_2^{(t-1)} - \bar{\mathbf{E}}_2^*) \right] \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2^\top \\
&= \widehat{\mathbf{U}}_2\widehat{\mathbf{U}}_2^\top \left(\mathbf{M}_2 + \bar{\mathbf{E}}_2^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)} \right) \widehat{\mathbf{V}}_2\widehat{\mathbf{V}}_2^\top
\end{aligned}$$

where

$$\Delta_{\mathbf{M}}^{(t-1)} = \frac{1}{n_2^{(t-1)}} \sum_{i=1}^n \mathbb{I}\left(\widehat{s}_i^{(t-1)} = 2\right) (\mathbf{M}_{s_i^*} - \mathbf{M}_2) \quad \text{and} \quad \Delta_{\mathbf{E}}^{(t-1)} = \bar{\mathbf{E}}_2^{(t-1)} - \bar{\mathbf{E}}_2^*$$

Notice that since $h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)$ satisfies (48), we have $n_2^{(t-1)} \geq 7\alpha n/16$. Lemma 10.2 implies that under event $\mathcal{Q}_1 \cap \mathcal{Q}_2$, we have

$$\begin{aligned}
\left\|\widehat{\mathbf{M}}_2^{(t)}\right\| &\leq (1+c)\left\|\mathbf{M}_2\right\| + C\left(\alpha^{-1/2}\sqrt{\frac{d}{n}} + \alpha^{-1}\sqrt{\frac{h(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}{n}} \right) \\
&\leq c\left(\alpha^{-1/2}\sqrt{\frac{d}{n}} + \kappa_0^{-1}\frac{d^{1/2}}{n^{1/4}} + \alpha^{-1/2}\kappa_0^{-1/2} \right)
\end{aligned}$$

where the second inequality is due to Assumption 5.1. On the other hand, under event $\mathcal{Q}_1 \cap \mathcal{Q}_2$ and Assumption 5.1 we also have

$$\left\|\widehat{\mathbf{M}}_1^{(t)}\right\| \geq (1-c)\left\|\mathbf{M}_1\right\| - c\left(\alpha^{-1/2}\sqrt{\frac{d}{n}} + \alpha^{-1/2}\kappa_0^{-1/2} \right) > \left\|\widehat{\mathbf{M}}_2^{(t)}\right\|$$

By taking a union bound over $\mathcal{Q}_1 \cap \mathcal{Q}_2$, we conclude that with probability at least $1 - \exp(-cd)$ we have $\left\|\widehat{\mathbf{M}}_2^{(t)}\right\| < \left\|\widehat{\mathbf{M}}_1^{(t)}\right\|$ and hence we set $\widehat{\mathbf{M}}_2^{(t)} = 0$ afterwards. Then for the second term on RHS of eq. (49), we have

$$\begin{aligned}
&\mathbb{P}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_2^{(t)} - \mathbf{M}_2 \right\rangle \geq \frac{\delta}{8}\left\|\mathbf{M}_1 - \mathbf{M}_2\right\|_{\mathrm{F}}^2 \right) = \mathbb{P}\left(\langle \mathbf{E}_i, -\mathbf{M}_2 \rangle \geq \frac{\delta}{8}\left\|\mathbf{M}_1 - \mathbf{M}_2\right\|_{\mathrm{F}}^2 \right) \\
&\leq \exp\left(-\frac{\delta^2\left\|\mathbf{M}_1 - \mathbf{M}_2\right\|_{\mathrm{F}}^4}{128\left\|\mathbf{M}_2\right\|_{\mathrm{F}}^2} \right) \leq \exp\left(-c\frac{\lambda_1^2 r_1}{\left\|\mathbf{M}_2\right\|^2 r_2}\delta^2\left\|\mathbf{M}_1 - \mathbf{M}_2\right\|_{\mathrm{F}}^2 \right)
\end{aligned}$$

where the last inequality is due to Assumption 5.1. Hence the expecatation can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \left\|\mathbf{M}_2 - \mathbf{M}_1\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq 2\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_2^{(t)} - \mathbf{M}_2 \right\rangle \geq \frac{\delta}{8}\left\|\mathbf{M}_1 - \mathbf{M}_2\right\|_{\mathrm{F}}^2 \right) \right] \\
&\quad \leq n\Delta^2 \exp\left[-c\delta^2 r_1 r_2^{-1} \left(\lambda_1 / \left\|\mathbf{M}_2\right\|\right)^2 \Delta^2 \right]
\end{aligned}$$

By Markov inequality, with probability at least $1-\exp\left[-\delta\left(\sqrt{r_1/r_2}\lambda_1/\left\|\mathbf{M}_2\right\|\right)\Delta\right]$ we get

$$\sum_{i=1}^{n}\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}^2\mathbb{I}\left(\widehat{s}_i^{(t)}\neq 2\right)\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\widehat{\mathbf{M}}_2^{(t)}-\mathbf{M}_2\right\rangle\geq\frac{\delta}{8}\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\right)\quad\leq n\cdot\exp\left(-\delta(\alpha n/K)^{1/2}\Delta^2\right)$$
$$\leq n\cdot\exp\left[-\delta^2 r_1 r_2^{-1}\left(\lambda_1/\left\|\mathbf{M}_2\right\|\right)^2\Delta^2\right]$$

which holds as long as $\delta\to 0$ sufficiently slowly compared with $\lambda_1^2 r_1 r_2^{-1}/\left\|\mathbf{M}_2\right\|^2\to\infty$.
It remains to consider $\beta_2(\mathbf{s}^*,\widehat{\mathbf{s}}^{(t)})$. Observe that

$$\begin{aligned}\beta_2(\mathbf{s}^*,\widehat{\mathbf{s}}^{(t)})\leq&\sum_{i=1}^{n}\sum_{a\in[2]\setminus\{s_i^*\}}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq a\right)\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_{s_i^*}-\widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\|_{\mathrm{F}}^2\geq\frac{\delta}{12}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\&+\sum_{i=1}^{n}\sum_{a\in[2]\setminus\{s_i^*\}}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq a\right)\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_{s_i^*}-\widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}}^2\geq\frac{\delta}{12}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\&+\sum_{i=1}^{n}\sum_{a\in[2]\setminus\{s_i^*\}}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq a\right)\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\mathbb{I}\left(\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}\left\|\mathbf{M}_a-\widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}}\geq\frac{\delta}{12}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\end{aligned}\tag{50}$$

The first term on RHS of eq. (50) can be written as

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{a\in[2]\setminus\{s_i^*\}}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq a\right)\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_{s_i^*}-\widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\|_{\mathrm{F}}^2\geq\frac{\delta}{12}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\&=\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq 1\right)\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_2-\widehat{\mathbf{M}}_2^{(t)}\right\|_{\mathrm{F}}^2\geq\frac{\delta}{12}\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}^2\right)\\&+\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq 2\right)\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}^2\mathbb{I}\left(\frac{1}{2}\left\|\mathbf{M}_1-\widehat{\mathbf{M}}_1^{(t)}\right\|_{\mathrm{F}}^2\geq\frac{\delta}{12}\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\right)\end{aligned}\tag{51}$$

The second term of (51) can be bounded the same way as that in **Step 2.3** of the proof of Theorem 3.1. Note that $\left\|\mathbf{M}_2-\widehat{\mathbf{M}}_2^{(t)}\right\|_{\mathrm{F}}^2=\left\|\mathbf{M}_2\right\|_{\mathrm{F}}^2\leq r_2\left\|\mathbf{M}_2\right\|^2=o(r_1\lambda_1^2)=o\left(\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\right)$ and hence the first term of (51) vanishes by setting δ slowly converging to 0. It suffices to consider the last term on RHS of eq. (50).

Observe that

$$\sum_{i=1}^{n}\sum_{a\in[2]\setminus\{s_i^*\}}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq a\right)\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\mathbb{I}\left(\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}\left\|\mathbf{M}_a-\widehat{\mathbf{M}}_a^{(t)}\right\|_{\mathrm{F}}\geq\frac{\delta}{12}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$=\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq 1\right)\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\mathbb{I}\left(\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}\left\|\mathbf{M}_1-\widehat{\mathbf{M}}_1^{(t)}\right\|_{\mathrm{F}}\geq\frac{\delta}{12}\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}^2\right)$$
$$+\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t)}\neq 2\right)\left\|\mathbf{M}_2-\mathbf{M}_1\right\|_{\mathrm{F}}^2\mathbb{I}\left(\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}\left\|\mathbf{M}_2-\widehat{\mathbf{M}}_2^{(t)}\right\|_{\mathrm{F}}\geq\frac{\delta}{12}\left\|\mathbf{M}_1-\mathbf{M}_2\right\|_{\mathrm{F}}^2\right)\tag{52}$$

The first term of (52) can be bounded the same way as that in **Step 2.3** of the proof of Theorem 3.1, and the second term vanishes as $\left\|\mathbf{M}_2-\widehat{\mathbf{M}}_2^{(t)}\right\|_{\mathrm{F}}=\|\mathbf{M}_2\|_{\mathrm{F}}=o\left(\|\mathbf{M}_1-\mathbf{M}_2\|_{\mathrm{F}}\right)$.
By following the remaining proofs of Theorem 3.1, we can finish the proof of Theorem 5.2.

10.6. Proof of Theorem 6.2

The outline of proof is based on the proof of Theorem 3.1, except that some delicate treatments are necessary when noise are sub-Gaussian.

Step 1: Good initialization. Following the same argument in the proof of Theorem 3.1 and by condition (17), we obtain the initial clustering error

$$\ell(\widehat{\mathbf{s}}^{(0)},\mathbf{s}^*)\leq\gamma^2\Delta^2h_c(\widehat{\mathbf{s}}^{(0)},\mathbf{s}^*)=o\left(\frac{\alpha n\Delta^2}{(\kappa_0\vee\gamma^2)K}\right)\tag{53}$$

which is analogous to (22).

Step 2: Iterative convergence. Similarly, define the following two events:

$$\mathcal{Q}_1:=\bigcup_{k\in[K]}\left\{\left\|\frac{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)\mathbf{E}_i}{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)}\right\|\leq C\sqrt{\frac{d}{n_k^*}}\right\}$$

and

$$\mathcal{Q}_2:=\bigcup_{I\subset[n]}\left\{\left\|\frac{1}{\sqrt{|I|}}\sum_{i\in I}\mathbf{E}_i\right\|\leq C\left(\sqrt{d}+\sqrt{n}\right)\right\}$$

where $C>0$ is some constant depending only on σ_{sg}. Following the same argument as in the proof Lemma 10.1, we get $\mathbb{P}(\mathcal{Q}_1^{\mathrm{c}}\cup\mathcal{Q}_2^{\mathrm{c}})\leq\exp(-c_0d)$ if $d\geq C_0\log K$ with some absolute constants $c_0,C_0>0$.

Thus we obtain

$$\ell(\widehat{\mathbf{s}}^{(t)},\mathbf{s}^*)\leq\xi_{\mathsf{err}}+\beta_1(\mathbf{s}^*,\widehat{\mathbf{s}}^{(t)})+\beta_2(\mathbf{s}^*,\widehat{\mathbf{s}}^{(t)})$$

where

$$\xi_{\mathsf{err}} := \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a - \mathbf{M}_{s_i^*}\right\rangle \geq \frac{1-\delta}{2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$

and

$$\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_{s_i^*} - \widehat{\mathbf{M}}_{s_i^*}^{(t)}\right\rangle + \left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a\right\rangle \geq \frac{\delta}{4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$

and

$$\beta_2(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right) \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\mathcal{R}(a; \widehat{\mathbf{s}}^{(t-1)}) \geq \frac{\delta}{4}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$

Step 2.1: Bounding ξ_{err}. Suppose Assumption 6.1 holds, $n \gg K$, $\Delta^2 \gg \log K$ and let δ converge to 0 slowly, we conclude that, with probability at least $1 - \exp(-\Delta)$,

$$\xi_{\mathsf{err}} \leq n \cdot \exp\left\{-\big(1 - o(1)\big) \cdot \frac{\Delta^2}{8\sigma_{\mathsf{sg}}^2}\right\}$$

Step 2.2: Bounding $\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$. As in the proof of Theorem 3.1, we write

$$\beta_1(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) = \beta_{1,1}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) + \beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$$

where we only focus on $\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)})$ defined as

$$\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) := \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \mathbb{I}\left(\widehat{s}_i^{(t)} \neq a\right) \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{M}}_a^{(t)} - \mathbf{M}_a\right\rangle \geq \frac{\delta}{8}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$

which can be bounded as

$$\begin{aligned}
\beta_{1,2}(\mathbf{s}^*, \widehat{\mathbf{s}}^{(t)}) \leq & \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\right\rangle \geq \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
& + \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \geq \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
& + \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top - \mathbf{U}_a\mathbf{U}_a^\top\right)\mathbf{M}_a\left(\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top - \mathbf{V}_a\mathbf{V}_a^\top\right)\right\rangle \geq \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
& + \sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \widehat{\mathbf{U}}_a\widehat{\mathbf{U}}_a^\top\Delta^{(t-1)}\widehat{\mathbf{V}}_a\widehat{\mathbf{V}}_a^\top\right\rangle \geq \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned} \tag{54}$$

Step 2.2.1: Treating the terms of $\langle \mathbf{E}_i, (\widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top - \mathbf{U}_a \mathbf{U}_a^\top)\mathbf{M}_a \rangle$. Similarly, we write

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \left(\widehat{\mathbf{U}}_a \widehat{\mathbf{U}}_a^\top - \mathbf{U}_a \mathbf{U}_a^\top\right)\mathbf{M}_a \right\rangle \geq \frac{\delta}{32}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
\leq &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
+ &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},2}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
+ &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \sum_{k\geq 3}\mathcal{S}_{\mathbf{M},k}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned} \tag{55}$$

Again, we bound the first term on RHS of (55) by

$$\begin{aligned}
&\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle = \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top \right\rangle = \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\bar{\mathbf{E}}_a^* + \Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a \right\rangle \\
&= \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_a \right\rangle + \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \\
&= \frac{1}{n_a^*}\mathbb{I}\left(s_i^* = a\right)\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a\right\|_{\mathrm{F}}^2 + \left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\frac{1}{n_a^*}\sum_{j\neq i}^{n}\mathbb{I}\left(s_j^* = a\right)\mathbf{E}_j\right)\mathbf{V}_a \right\rangle \\
&\quad + \left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle
\end{aligned}$$

We then bound the first term on RHS of eq. (28) by

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathcal{S}_{\mathbf{M},1}^{\mathbf{U}_a}(\Delta^{(t-1)})\mathbf{M}_a \right\rangle \geq \frac{\delta}{96}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
\leq &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^* = a\right)\left\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a\right\|_{\mathrm{F}}^2 \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
+ &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a, \mathbf{U}_{a\perp}^\top \left(\frac{1}{n_a^*}\sum_{j\neq i}^{n}\mathbb{I}\left(s_j^* = a\right)\mathbf{E}_j\right)\mathbf{V}_a \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right) \\
+ &\sum_{i=1}^{n} \sum_{a\in[K]\setminus\{s_i^*\}} \left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2 \cdot \mathbb{I}\left(\left\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \left(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top \right\rangle \geq \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned} \tag{56}$$

We bound the first two terms on RHS of eq. (56) by Markov inequality and thus their expectation is needed. Note the entries of $\mathbf{E}_i$ are i.i.d with zero mean, unit variance, and sub-Gaussian constant bounded by $O(\sigma_{\mathsf{sg}})$. Clearly,

$\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a\|_{\mathrm{F}} \le r_a^{1/2}\|\mathbf{E}_i\|$. By the standard concentration property of sub-Gaussian random matrix (e.g., [55, Theorem 4.4.5]), there exist absolute constants $c_1, C_1 > 0$ such that

$$\mathbb{P}\Big(\|\mathbf{E}_i\| \ge C_1\sigma_{\mathsf{sg}}(d^{1/2}+u^{1/2})\Big) \le e^{-c_1 u}, \quad \forall u > 0.$$

Therefore, we get

$$\begin{aligned}\mathbb{P}\bigg(\frac{1}{n_a^*}\|\mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a\|_{\mathrm{F}}^2 \ge \frac{\delta}{288}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\bigg) &\le \mathbb{P}\bigg(\frac{r_a}{n_a^*}\|\mathbf{E}_i\|^2 \ge \frac{\delta}{288}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\bigg)\\ &\le \exp\bigg(-c_1\frac{\delta\alpha n}{\sigma_{\mathsf{sg}}^2 K}\cdot\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\bigg)\end{aligned}$$

which holds if $\Delta^2 \gg \alpha^{-1}\sigma_{\mathsf{sg}}^2 rdK/n$ and by setting $\delta \to 0$ sufficiently slowly.

The second term in RHS of (56) can be dealt with as in the proof of Theorem 3.1 due to the independence between $\mathbf{E}_i$ and $\sum_{j\neq i}\mathbf{E}_j$. Thus, we get

$$\begin{aligned}&\mathbb{P}\bigg(\Big\langle \mathbf{U}_{a\perp}^\top \mathbf{E}_i \mathbf{V}_a,\ \mathbf{U}_{a\perp}^\top\Big(\sum_{j\neq i}^n \mathbb{I}\left(s_j^* = a\right)\mathbf{E}_j\Big)\mathbf{V}_a\Big\rangle \ge \frac{\delta n_a^*}{288}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2\bigg)\\ &\le 2\exp\bigg(-c_2\frac{\delta\sqrt{\alpha n}\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\|_{\mathrm{F}}^2}{\sigma_{\mathsf{sg}}^2\sqrt{K}}\bigg)\end{aligned}$$

if $\Delta^2 \gg \alpha^{-1}\sigma_{\mathsf{sg}}^2 rdK/n, \alpha n \gg K$ and setting $\delta \to 0$ sufficiently slowly. Therefore, we can bound the expectation of the first two terms on RHS of (56), and by Markov inequality, we get with probability at least $1-\exp\left(-\delta(\alpha n/K)^{1/4}\Delta\sigma_{\mathsf{sg}}^{-1}\right)$ that

$$\begin{aligned}&\sum_{i=1}^n\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\frac{1}{n_a^*}\mathbb{I}\left(s_i^* = a\right)\left\|\mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a\right\|_{\mathrm{F}}^2 \ge \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &+\sum_{i=1}^n\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{U}_{a\perp}^\top\mathbf{E}_i\mathbf{V}_a, \mathbf{U}_{a\perp}^\top\left(\frac{1}{n_a^*}\sum_{j\neq i}^n\mathbb{I}\left(s_j^* = a\right)\mathbf{E}_j\right)\mathbf{V}_a\right\rangle \ge \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &\le n\cdot\exp\left(-\delta(\alpha n/K)^{1/2}\Delta^2\sigma_{\mathsf{sg}}^{-2}\right)\end{aligned}$$

The third term on RHS of (56) can be handled in the same way as in the proof of Theorem 3.1 since Lemma 10.4 does not rely on Gaussian assumption. Recall the event $\mathcal{Q}_3$ defined after Lemma 10.4. On the event $\mathcal{Q}_3$, we get

$$\begin{aligned}&\sum_{i=1}^n\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a - \mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\left(\Delta_{\mathbf{M}}^{(t-1)} + \Delta_{\mathbf{E}}^{(t-1)}\right)\mathbf{V}_a\mathbf{V}_a^\top\right\rangle \ge \frac{\delta}{288}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\ &\le \frac{1}{16}\ell(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)\end{aligned}$$

as long as $\Delta^2 \gg \alpha^{-1}\sigma_{\mathsf{sg}}^2 K^2 r(rdn^{-1}+1)$.

We then bound the second term on RHS of (55), i.e., the one involving $\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}(\Delta^{(t-1)})$. Write

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathcal{S}^{\mathbf{U}_a}_{\mathbf{M},2}(\Delta^{(t-1)})\mathbf{M}_a\right\rangle\geq\frac{\delta}{96}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\Delta^{(t-1)\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{U}_a^\top\Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\Delta^{(t-1)\top}\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_a\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}
\tag{57}
$$

For the indicator function in the first term on the RHS of eq. (57), we further have the decomposition

$$
\begin{aligned}
&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\Delta^{(t-1)}\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\Delta^{(t-1)\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{288}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
\leq&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)\\
+&\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top(\Delta_{\mathbf{M}}^{(t-1)}+\Delta_{\mathbf{E}}^{(t-1)})^\top\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)
\end{aligned}
\tag{58}
$$

Similarly, we bound the expectation of the first term on the RHS of (58). Denote

$$
\bar{\mathbf{E}}_{-i}:=\frac{1}{n_a^*}\sum_{j\neq i}^{n}\mathbb{I}\left(s_j^*=a\right)\mathbf{E}_j\quad\text{ so that }\quad\bar{\mathbf{E}}_a^*=\bar{\mathbf{E}}_{-i}+\frac{\mathbf{E}_i}{n_a^*}\cdot\mathbb{I}\left(s_i^*=a\right)
$$

The first term on RHS of (58) has an expectation bounded by

$$
\begin{aligned}
&\mathbb{P}\bigg(\big\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_a^* \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_a^{*\top} \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \big\rangle \geq \frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&\leq \mathbb{P}\bigg(\big\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_{-i} \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_{-i}^\top \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \big\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&+ \mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Big(\frac{\mathbf{E}_i}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_{-i}^\top \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&+ \mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_{-i} \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Big(\frac{\mathbf{E}_i^\top}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&+ \mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Big(\frac{\mathbf{E}_i}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Big(\frac{\mathbf{E}_i^\top}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg)
\end{aligned} \tag{59}
$$

Note that the first three terms of eq. (59) can be bounded as in the proof of Theorem 3.1 due to the independence between $\mathbf{E}_i$ and $\bar{\mathbf{E}}_{-i}$ after decoupling. We obtain

$$
\begin{aligned}
&\mathbb{P}\bigg(\big\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_{-i} \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_{-i}^\top \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \big\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&+ \mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Big(\frac{\mathbf{E}_i}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \bar{\mathbf{E}}_{-i}^\top \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&+ \mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \bar{\mathbf{E}}_{-i} \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Big(\frac{\mathbf{E}_i^\top}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&\leq 3\exp\bigg(-c\frac{\delta\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \alpha n/K}{\kappa_0^2 r^2 \sigma_{\mathsf{sg}}^2} \bigg) + 3\exp\bigg(-c\frac{\delta^{1/2}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \sqrt{\alpha n/K}}{\sigma_{\mathsf{sg}}^2} \bigg)
\end{aligned}
$$

where we've used $\lambda^2 \geq (1/4)\cdot r^{-1}\kappa_0^{-2}\max_{a,b\in[K],a\neq b}\left\|\mathbf{M}_a - \mathbf{M}_b\right\|_{\mathrm{F}}^2$, $\lambda^2 \geq \alpha^{-1}\sigma_{\mathsf{sg}}^2 dK/n$, $\Delta^2 \gg \alpha^{-1}\sigma_{\mathsf{sg}}^2 rdK/n$, $\alpha n/K \gg \kappa_0^2 r^2$ and set $\delta \to 0$ sufficiently slowly. The last term of eq. (59) can be handled as

$$
\begin{aligned}
&\mathbb{P}\bigg(\bigg\langle \mathbf{E}_i, \mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top \Big(\frac{\mathbf{E}_i}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top \Big(\frac{\mathbf{E}_i^\top}{n_a^*}\cdot \mathbb{I}\left(s_i^* = a\right)\Big) \mathbf{U}_a \mathbf{\Sigma}_a^{-1}\mathbf{V}_a^\top \bigg\rangle \geq \frac{\delta}{4608}\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) \\
&\overset{(a)}{\leq} \mathbb{P}\bigg(\frac{r}{(n_a^*)^{3/2}}\sigma_{\mathsf{sg}}^3 d + \frac{\kappa_0 r^{3/2}}{(n_a^*)^{3/2}}\delta^3\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \geq c\delta\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 \bigg) + \exp\bigg(-c\frac{\delta^2\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 (\alpha n)^{1/3}}{\sigma_{\mathsf{sg}}^2 K^{1/3}} \bigg) \\
&\overset{(b)}{=} \exp\bigg(-c\frac{\delta^2\left\|\mathbf{M}_{s_i^*} - \mathbf{M}_a\right\|_{\mathrm{F}}^2 (\alpha n)^{1/3}}{\sigma_{\mathsf{sg}}^2 K^{1/3}} \bigg)
\end{aligned}
$$

where (a) holds if $\lambda^2 \geq (1/4)\cdot r^{-1}\kappa_0^{-2}\max_{a,b\in[K],a\neq b}\left\|\mathbf{M}_a - \mathbf{M}_b\right\|_{\mathrm{F}}^2$ and (b) holds provided that $\lambda \geq \alpha^{-1/2}\sigma_{\mathsf{sg}}\sqrt{dK/n}$ $\Delta^2 \gg \alpha^{-1}\sigma_{\mathsf{sg}}^2 rdK/n$, $\alpha n/K \geq \kappa_0^{2/3} r$ and by setting $\delta \to 0$ sufficiently slowly. Therefore, by Markov inequality, we get with probability at least $1 - \exp\left(-\delta^2(\alpha n/K)^{1/6}\Delta\sigma_{\mathsf{sg}}^{-1}\right) - \exp\left(-\delta(\kappa_0 r)^{-1}(\alpha n/K)^{1/2}\Delta\sigma_{\mathsf{sg}}^{-1}\right)$

that

$$\sum_{i=1}^{n}\sum_{a\in[K]\setminus\{s_i^*\}}\left\|\mathbf{M}_a-\mathbf{M}_{s_i^*}\right\|_{\mathrm{F}}^2\cdot\mathbb{I}\left(\left\langle\mathbf{E}_i,\mathbf{U}_{a\perp}\mathbf{U}_{a\perp}^\top\bar{\mathbf{E}}_a^*\mathbf{V}_{a\perp}\mathbf{V}_{a\perp}^\top\bar{\mathbf{E}}_a^{*\top}\mathbf{U}_a\boldsymbol{\Sigma}_a^{-1}\mathbf{V}_a^\top\right\rangle\geq\frac{\delta}{1152}\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\|_{\mathrm{F}}^2\right)$$
$$\leq n\cdot\exp\left(-\delta^2(\alpha n/K)^{1/3}\Delta^2\sigma_{\mathsf{sg}}^{-2}\right)+\exp\left(-\delta(\kappa_0 r)^{-2}(\alpha n/K)\Delta^2\sigma_{\mathsf{sg}}^{-2}\right)$$

The remaining three terms on RHS of (58) can be handled in the same way as (34) which does not rely on Gaussian assumption.
The proof for the last two terms on RHS of eq. (57) the last term on RHS of eq. (55) remains untouched as the proof of Theorem 3.1 and hence omitted.
Finally, we can mimic the proof of Theorem 3.1 line by line starting from **Step 2.2.2** till the end, which completes the proof of Theorem 6.2.

11. Proof of Technical Lemmas

11.1. Proof of Lemma 3.2

By definition we have that

$$\mathbf{U}^\top\mathbf{U}=\begin{bmatrix}\mathbf{I}_{r_1}&\mathbf{U}_1^\top\mathbf{U}_2&\cdots&\mathbf{U}_1^\top\mathbf{U}_K\\ \mathbf{U}_2^\top\mathbf{U}_1&\mathbf{I}_{r_2}&\cdots&\mathbf{U}_2^\top\mathbf{U}_K\\ \vdots&\vdots&\ddots&\vdots\\ \mathbf{U}_K^\top\mathbf{U}_1&\mathbf{U}_K^\top\mathbf{U}_2&\cdots&\mathbf{I}_{r_K}\end{bmatrix}$$

and $\mathbf{W}^\top\mathbf{W}=\mathrm{diag}(n_1^*,\cdots,n_K^*)$. Hence we have

$$\mathbf{W}^\top\mathbf{W}\otimes\mathbf{V}^\top\mathbf{V}=\begin{bmatrix}n_1^*\mathbf{U}^\top\mathbf{U}&\mathbf{0}&\cdots&\mathbf{0}\\ \mathbf{0}&n_2^*\mathbf{U}^\top\mathbf{U}&\cdots&\mathbf{0}\\ \vdots&\vdots&\ddots&\vdots\\ \mathbf{0}&\mathbf{0}&\cdots&n_K^*\mathbf{U}^\top\mathbf{U}\end{bmatrix}$$

Simple calculations give that

$$\begin{aligned}\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}})&=\mathbf{U}\mathscr{M}_1(\boldsymbol{\mathcal{S}})(\mathbf{W}^\top\mathbf{W}\otimes\mathbf{V}^\top\mathbf{V})\mathscr{M}_1^\top(\boldsymbol{\mathcal{S}})\mathbf{U}^\top\\&=\mathbf{U}\cdot\mathrm{diag}(n_1^*\boldsymbol{\Sigma}_1^2,\cdots,n_K^*\boldsymbol{\Sigma}_K^2)\cdot\mathbf{U}^\top\end{aligned}$$

As a result, we obtain

$$\sigma_1(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}))\leq\sigma_1^2(\mathbf{U})\cdot\max_{1\leq k\leq K}n_k^*\sigma_{\mathsf{max}}^2(\boldsymbol{\Sigma}_k)$$

$$\sigma_{r_{\mathbf{U}}}(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}))\geq\sigma_{r_{\mathbf{U}}}^2(\mathbf{U})\cdot\min_{1\leq k\leq K}n_k^*\sigma_{\mathsf{min}}^2(\boldsymbol{\Sigma}_k)$$

Hence we conclude that

$$\kappa_1=\sqrt{\frac{\sigma_1(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}))}{\sigma_{r_{\mathbf{U}}}(\mathscr{M}_1(\boldsymbol{\mathcal{M}})\mathscr{M}_1^\top(\boldsymbol{\mathcal{M}}))}}\leq\kappa_0\kappa(\mathbf{U})\cdot\sqrt{\frac{n_{\mathsf{max}}^*}{n_{\mathsf{min}}^*}}$$

Similarly we can prove that $\mathscr{M}_2(\boldsymbol{\mathcal{M}})\mathscr{M}_2^\top(\boldsymbol{\mathcal{M}}) = \mathbf{V}\cdot\mathrm{diag}(n_1^*\boldsymbol{\Sigma}_1^2,\cdots,n_K^*\boldsymbol{\Sigma}_K^2)\cdot\mathbf{V}^\top$ and $\kappa_1 \le \kappa_0\kappa(\mathbf{U})\cdot(n^*_{\max}/n^*_{\min})^{1/2}$.
If $r_{\mathbf{U}} = r_{\mathbf{V}} = r_1$, by min-max principle for singular values we have

$$\sigma_{\min}(\mathbf{U}) = \sigma_{r_1}(\mathbf{U}) = \max_{S\subset\mathbb{R}^n,\dim(S)=r_1}\min_{x\in S,\|x\|=1}\left\|\begin{bmatrix}\mathbf{U}_1^\top x\\ \vdots\\ \mathbf{U}_K^\top x\end{bmatrix}\right\| \ge \max_{S\subset\mathbb{R}^n,\dim(S)=r_1}\min_{x\in S,\|x\|=1}\left\|\mathbf{U}_1^\top x\right\| = \sigma_{\min}(\mathbf{U}_1) = 1$$

and

$$\sigma_{\max}(\mathbf{U}) = \max_{x\in\mathbb{R}^n,\|x\|=1}\left\|\begin{bmatrix}\mathbf{U}_1^\top x\\ \vdots\\ \mathbf{U}_K^\top x\end{bmatrix}\right\| \le \sqrt{\sum_{k=1}^K\max_{x\in\mathbb{R}^n,\|x\|=1}\left\|\mathbf{U}_k^\top x\right\|} = \sqrt{K}$$

Therefore, we have $\kappa(\mathbf{U}) \le K^{1/2}$ and similarly $\kappa(\mathbf{V}) \le K^{1/2}$, from which we can conclude that $\max\{\kappa_1,\kappa_2\} \le \kappa_0(K^2/\alpha)^{1/2}$.
If $r_{\mathbf{U}} = r_{\mathbf{V}} = \mathring{r}$ and $\mathbf{U}_k$'s are mutually orthogonal, then $\mathbf{U},\mathbf{V}$ has orthonormal columns and $\kappa(\mathbf{U}) = \kappa(\mathbf{V}) = 1$. Hence we have $\max\{\kappa_1,\kappa_2\} \le \kappa_0(K/\alpha)^{1/2}$.

11.2. Proof of Lemma 10.1

Note that for fixed $k\in[K]$, we have $\frac{\sum_{i=1}^n\mathbb{I}(s_i^*=k)\mathbf{E}_i}{\sum_{i=1}^n\mathbb{I}(s_i^*=k)}$ has i.i.d. sub-gaussian entries with mean zero and variance $(n_k^*)^{-1}$. By random matrix theory there exists some absolute constants $c, C > 0$ such that

$$\mathbb{P}\left(\left\|\frac{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)\mathbf{E}_i}{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)}\right\| \ge C\sqrt{\frac{d}{n_k^*}}\right) \le \exp(-cd)$$

Applying a union bound gives

$$\mathbb{P}(Q_1^c) = \mathbb{P}\left(\bigcup_{k=1}^K\left\{\left\|\frac{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)\mathbf{E}_i}{\sum_{i=1}^n\mathbb{I}\left(s_i^*=k\right)}\right\| \ge C\sqrt{\frac{d}{n_k^*}}\right\}\right) \le K\exp(-cd) \le \exp(-c_0 d)$$

for some absolute constant $c_0 > 0$, provided that $d \gtrsim \log K$. To prove the tail bound for Q_2, consider fixed set $I\subseteq[n]$, we have for any $t > 0$:

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{|I|}}\sum_{i\in I}\mathbf{E}_i\right\| \le C\left(\sqrt{d}+t\right)\right) \le 2\exp(-t^2)$$

Applying a union bound gives

$$\mathbb{P}(Q_2^c) = \mathbb{P}\left(\bigcup_{I\subseteq[n]}\left\{\left\|\frac{1}{\sqrt{|I|}}\sum_{i\in I}\mathbf{E}_i\right\| \le C\left(\sqrt{d}+t\right)\right\}\right) \le 2\exp(-t^2+n)$$

By choosing $t = C\sqrt{n}$ we obtain the desired result.

11.3. Proof of Lemma 10.2

By definition, $\left\|\Delta_{\mathbf{M}}^{(t-1)}\right\|$ can be bounded by

$$\begin{aligned}\left\|\Delta_{\mathbf{M}}^{(t-1)}\right\| &= \left\|\frac{1}{n_a^{(t-1)}}\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t-1)}=a\right)\left(\mathbf{M}_{s_i^*}-\mathbf{M}_a\right)\right\| \\ &= \left\|\frac{1}{n_a^{(t-1)}}\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t-1)}=a, s_i^*\neq a\right)\left(\mathbf{M}_{s_i^*}-\mathbf{M}_a\right)\right\| \\ &\leq \frac{8K}{7\alpha n}\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t-1)}=a, s_i^*\neq a\right)\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\| \\ &\leq \frac{16\kappa_0 K}{7\alpha n}\cdot\lambda\cdot h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)\end{aligned}$$

where we've used $h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)\leq\sum_{a\in[K]}h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)=h(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$ and the condition (10). An alternative bound for $\left\|\Delta_{\mathbf{M}}^{(t-1)}\right\|$:

$$\left\|\Delta_{\mathbf{M}}^{(t-1)}\right\| \leq \frac{8K}{7\alpha n}\sum_{i=1}^{n}\mathbb{I}\left(\widehat{s}_i^{(t-1)}=a, s_i^*\neq a\right)\left\|\mathbf{M}_{s_i^*}-\mathbf{M}_a\right\| \leq \frac{16\gamma K}{\alpha n}\cdot\Delta\cdot h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)$$

In other words, we have the following bound for $\Delta_{\mathbf{M}}^{(t-1)}$ that will be utilized repeatedly later:

$$\left\|\Delta_{\mathbf{M}}^{(t-1)}\right\| \leq \frac{16K}{\alpha n}h_a(\widehat{\mathbf{s}}^{(t-1)},\mathbf{s}^*)\cdot\min\{\kappa_0\lambda,\gamma\Delta\} \tag{60}$$

Moreover, under $\mathcal{Q}_1$ we have

$$\left\|\bar{\mathbf{E}}_a^*\right\| \lesssim \sqrt{\frac{d}{n_a^*}} \lesssim \sqrt{\frac{dK}{\alpha n}}$$

and it remains to bound $\left\|\Delta_{\mathbf{E}}^{(t-1)}\right\|$. Note that

$$\begin{aligned}
\left\|\Delta_{\mathbf{E}}^{(t-1)}\right\| &= \left\| \frac{1}{n_a^{(t-1)}} \sum_{i=1}^{n} \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) \mathbf{E}_i - \frac{1}{n_a^*} \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = a\right) \mathbf{E}_i \right\| \\
&\leq \left\| \frac{1}{n_a^{(t-1)}} \sum_{i=1}^{n} \left[\mathbb{I}\left(\widehat{s}_i^{(t-1)} = a\right) - \mathbb{I}\left(s_i^* = a\right) \right] \mathbf{E}_i \right\| + \left\| \frac{n_a^* - n_a^{(t-1)}}{n_a^{(t-1)} n_a^*} \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = a\right) \mathbf{E}_i \right\| \\
&\leq \left\| \frac{1}{n_a^{(t-1)}} \sum_{i=1}^{n} \mathbb{I}\left(\widehat{s}_i^{(t-1)} = a, s_i^* \neq a\right) \mathbf{E}_i \right\| + \left\| \frac{1}{n_a^{(t-1)}} \sum_{i=1}^{n} \mathbb{I}\left(\widehat{s}_i^{(t-1)} \neq a, s_i^* = a\right) \mathbf{E}_i \right\| \\
&+ \frac{1}{n_a^{(t-1)}} \cdot \left| \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = a, \widehat{s}_i^{(t-1)} \neq a\right) \right| \left\| \frac{1}{n_a^*} \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = a\right) \mathbf{E}_i \right\| \\
&+ \frac{1}{n_a^{(t-1)}} \cdot \left| \sum_{i=1}^{n} \mathbb{I}\left(s_i^* \neq a, \widehat{s}_i^{(t-1)} = a\right) \right| \left\| \frac{1}{n_a^*} \sum_{i=1}^{n} \mathbb{I}\left(s_i^* = a\right) \mathbf{E}_i \right\| \\
&\overset{(a)}{\lesssim} \frac{K\sqrt{(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}}{\alpha n} + \frac{K}{n} h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \sqrt{\frac{dK}{\alpha n}} \\
&\overset{(b)}{\lesssim} \frac{K\sqrt{(d+n)h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*)}}{\alpha n}
\end{aligned}$$

where in (a) we've used the fact that $\mathcal{Q}_2$ holds and (b) is due to that fact that $h_a(\widehat{\mathbf{s}}^{(t-1)}, \mathbf{s}^*) \lesssim \alpha n/K$.

11.4. Proof of Lemma 10.3

The conclusion directly follows from dilation, i.e., define

$$\mathbf{X}^* := \begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{0} \end{bmatrix}, \quad \mathbf{M}^* := \begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{M}^\top & \mathbf{0} \end{bmatrix}, \quad \Delta^* := \begin{bmatrix} \mathbf{0} & \Delta \\ \Delta^\top & \mathbf{0} \end{bmatrix}$$

and applying Theorem 1 in [60].

11.5. Proof of Lemma 10.4

To decouple the potential dependency of $\mathbf{E}_i$ and $\mathbf{\Xi}$, we consider an ϵ-net $\mathcal{N}_{d_1,d_2,r}^{\epsilon}$ for $\{\mathbf{\Xi} \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(\mathbf{\Xi}) \leq r, \|\mathbf{\Xi}\| \leq 1\}$. A standard ϵ-net argument, e.g., Lemma 7 in [64], would give the following cardinality bound:

$$|\mathcal{N}_{d_1,d_2,r}^{\epsilon}| \leq \left(\frac{4+\epsilon}{\epsilon}\right)^{(d_1+d_2)r}$$

Now we consider any fixed $\mathbf{X} \in \mathcal{N}^{\epsilon}_{d_1,d_2,r}$ and denote its compact SVD as $\mathbf{L}\mathbf{\Lambda}\mathbf{R}^\top$, then $\mathrm{vec}(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}) \overset{d}{=} \boldsymbol{\mathcal{N}}(0, \mathbf{I}_{r^{*2}})$ with $r^* := \mathrm{rank}(\mathbf{X}) \le r$. We have that

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \langle \mathbf{E}_i, \mathbf{X}\rangle^2 = \sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \left\langle \mathbf{L}^\top\mathbf{E}_i\mathbf{R}, \mathbf{\Lambda}\right\rangle^2 \\
&= \mathrm{vec}^\top(\mathbf{\Lambda}) \left(\sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \mathrm{vec}\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right) \mathrm{vec}^\top\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right)\right) \mathrm{vec}(\mathbf{\Lambda}) \\
&\le r \left\|\sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \mathrm{vec}\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right) \mathrm{vec}^\top\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right)\right\|
\end{aligned}$$

Using a standard argument via concentration inequality for χ^2, we can obtain for any $t > 0$ that

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \mathrm{vec}\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right) \mathrm{vec}^\top\left(\mathbf{L}^\top\mathbf{E}_i\mathbf{R}\right)\right\| \ge C(n_b^* + r^2 + t)\right) \le \exp(-ct)$$

for some absolute constants $c, C > 0$. As a result, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{\mathbf{\Xi}\in\mathbb{R}^{d_1\times d_2}, \mathrm{rank}(\mathbf{\Xi})\le r \\ \|\mathbf{\Xi}\|\le 1}} \sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \langle \mathbf{E}_i, \mathbf{\Xi}\rangle^2 \ge Cr(n_b^* + r^2 + t)\right) \\
&\le \mathbb{P}\left(\sup_{\mathbf{X}\in\mathcal{N}^{\epsilon}_{d_1,d_2,r}} \sum_{i=1}^n \mathbb{I}\left(s_i^* = b\right) \langle \mathbf{E}_i, \mathbf{X}\rangle^2 \ge C'(n_b^* + r^2 + t)\right) \\
&\le C^{dr} \exp(-ct)
\end{aligned}$$

where $C' > 0$ is some absolute constant depending only on ϵ. The proof is completed by choosing $t = Cdr$ for some large constant $C > 0$.

11.6. Proof of Lemma 10.5

Without loss of generality we only proof $j = 1$. It follows that

$$\begin{aligned}
\sigma^2_{\min}(\mathscr{M}_1(\boldsymbol{\mathcal{M}})) &\ge \kappa_1^{-2} \left\|\mathscr{M}_1(\boldsymbol{\mathcal{M}})\right\|^2 \ge \kappa_1^{-2} r_{\mathbf{U}}^{-1} \sum_{k=1}^K n_k \left\|\mathbf{M}_k\right\|_{\mathrm{F}}^2 \\
&\ge \kappa_1^{-2} r_{\mathbf{U}}^{-1} n\lambda^2 \ge \kappa_1^{-2}(Kr)^{-1} n\lambda^2
\end{aligned}$$

where the last inequality is due to $r_{\mathbf{U}} \le \sum_{k=1}^K r_k \le Kr$.

11.7. Proof of Lemma 10.6

We first prove (I). By definition of k-means

$$\left\|\mathfrak{M} - \mathbf{G}\right\|_{\mathrm{F}} \le \left\|\mathfrak{M} - \widehat{\mathbf{G}}\right\|_{\mathrm{F}} + \left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|_{\mathrm{F}} \le 2\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|_{\mathrm{F}} \le 2\sqrt{2K}\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\|$$

It suffices to notice that there exists some absolute constant $c, C > 0$ such that

$$\begin{aligned}\left\|\widehat{\mathbf{G}} - \mathbf{G}\right\| &= \left\|\mathscr{M}_3(\boldsymbol{\mathcal{X}} \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \boldsymbol{\mathcal{M}} \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top)\right\| \\ &= \left\|\mathscr{M}_3(\boldsymbol{\mathcal{E}})(\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top \otimes \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top)\right\| = \left\|\mathscr{M}_3(\boldsymbol{\mathcal{E}})(\widehat{\mathbf{V}} \otimes \widehat{\mathbf{U}})\right\| \\ &\leq C\left(\sqrt{d(r_{\mathbf{U}} + r_{\mathbf{V}})} + \sqrt{n}\right)\end{aligned}$$

where the last inequality holds with probability at least $1 - \exp(-cd)$ by Lemma 5 in [64]. Hence there exists some $C_0 > 0$, and with probability at least $1 - \exp(-cd)$ we have

$$\|\mathfrak{M} - \mathbf{G}\|_{\mathrm{F}} \leq C_0\sqrt{K}\left(\sqrt{dKr + n}\right)$$

for some absolute constant $C_0 > 0$. It remains to prove (II). By definition of $\boldsymbol{\mathcal{G}}$, we obtain

$$\begin{aligned}&\left\|\boldsymbol{\mathcal{G}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right\|_{\mathrm{F}} \\ &= \left\|\left[\boldsymbol{\mathcal{M}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 \widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\right\|_{\mathrm{F}} \\ &\geq \left\|\left[\boldsymbol{\mathcal{M}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \mathbf{U}\mathbf{U}^\top \times_2 \mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{F}} - \left\|\left[\boldsymbol{\mathcal{M}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 (\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top) \times_2 \mathbf{V}\mathbf{V}^\top\right\|_{\mathrm{F}} \\ &- \left\|\left[\boldsymbol{\mathcal{M}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \times_2 (\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top)\right\|_{\mathrm{F}} \\ &\geq \Delta - \frac{\Delta}{4} - \frac{\Delta}{4} \geq \frac{\Delta}{2}\end{aligned}$$

where we've used the fact that $\mathcal{Q}_0$ holds and the equivalence between $\sqrt{2}\left\|\sin\Theta(\mathbf{U}_1, \mathbf{U}_2)\right\|_{\mathrm{F}}$ and projection distance $\left\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\right\|_{\mathrm{F}}$.

11.8. Proof of Lemma 10.7

The proof of (I) is identical to that in the proof of Lemma 10.6 and hence we only show (II). By definition of $\boldsymbol{\mathcal{G}}$, we obtain

$$
\begin{aligned}
&\left\|\boldsymbol{\mathcal{G}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right\|_{\mathrm{F}} \\
&= \left\| \left[\boldsymbol{\mathcal{M}} \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{V}}_1 \widehat{\mathbf{V}}_1^\top \right\|_{\mathrm{F}} \\
&\geq \left\| \left[\boldsymbol{\mathcal{M}}_1 \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \mathbf{U}_1 \mathbf{U}_1^\top \times_2 \mathbf{V}_1 \mathbf{V}_1^\top \right\|_{\mathrm{F}} \\
&- \left\| \left[\boldsymbol{\mathcal{M}}_1 \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 (\widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top - \mathbf{U}_1 \mathbf{U}_1^\top) \times_2 \mathbf{V}_1 \mathbf{V}_1^\top \right\|_{\mathrm{F}} \\
&- \left\| \left[\boldsymbol{\mathcal{M}}_1 \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top \times_2 (\widehat{\mathbf{V}}_1 \widehat{\mathbf{V}}_1^\top - \mathbf{V}_1 \mathbf{V}_1^\top) \right\|_{\mathrm{F}} \\
&- \left\| \left[\boldsymbol{\mathcal{M}}_2 \times_3 (\mathbf{e}_i^\top - \mathbf{e}_j^\top)\right] \times_1 \widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1^\top \times_2 \widehat{\mathbf{V}}_1 \widehat{\mathbf{V}}_1^\top \right\|_{\mathrm{F}} \\
&\overset{(a)}{\geq} \|\mathbf{M}_1\|_{\mathrm{F}} - \frac{\|\mathbf{M}_1\|_{\mathrm{F}}}{6} - \frac{\|\mathbf{M}_1\|_{\mathrm{F}}}{6} - \|\mathbf{M}_2\|_{\mathrm{F}} \\
&\overset{(b)}{\geq} \frac{5}{9} \|\mathbf{M}_1\|_{\mathrm{F}} \geq \frac{\Delta}{2}
\end{aligned}
$$

where in (a) we've used (47), (b) and (c) are due to the facts that $\|\mathbf{M}_1\|_{\mathrm{F}} \geq 9\|\mathbf{M}_2\|_{\mathrm{F}}$ and $\Delta = \|\mathbf{M}_1 - \mathbf{M}_2\|_{\mathrm{F}} \leq 10/9 \cdot \|\mathbf{M}_1\|_{\mathrm{F}}$, by properly choosing the absolute constant C in Assumption 5.1 and the proof is therefore completed.