Cutoff for Exact Recovery of Gaussian Mixture Models

Xiaohui Chen and Yun Yang

***Abstract*— We determine the information-theoretic cutoff value on separation of cluster centers for exact recovery of cluster labels in a K-component Gaussian mixture model with equal cluster sizes. Moreover, we show that a semidefinite programming (SDP) relaxation of the K-means clustering method achieves such sharp threshold for exact recovery without assuming the symmetry of cluster centers.**

***Index Terms*— K-means, Gaussian mixture models, semidefinite relaxation, exact recovery, sharp threshold, optimality.**

I. Introduction

LET $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sequence of independent random vectors in $\mathbb{R}^p$ sampled from a K-component Gaussian mixture model with $K \leqslant n$. Specifically, we assume that there exists a partition $G_1^*, \ldots, G_K^*$ of the index set $[n] := \{1, \ldots, n\}$ such that if $i \in G_k^*$, then

$$\mathbf{X}_i = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \sigma^2 I_p), \tag{1}$$

where $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^p$ are the unknown cluster centers and $\sigma^2 > 0$ is the common noise variance. For simplicity, we assume that σ^2 is known. Our main focus of this paper is to investigate the problem of optimal exact recovery for the true partition (or clustering) structure $G_1^*, \ldots, G_K^*$.

For each partition $G_1, \ldots, G_K$ of $[n]$, let $H = (h_{ik}) \in \{0,1\}^{n \times K}$ be the binary *assignment* matrix of the observation $\mathbf{X}_i$ to the cluster k, i.e.,

$$h_{ik} = \begin{cases} 1, & \text{if } i \in G_k, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } i \in [n],\ k \in [K].$$

Since each row of H contains exactly one nonzero entry, there is one-to-one mapping (up to assignment labeling) between the partition and the assignment matrix. Thus recovery of the true clustering structure is equivalently to recovery of the associated assignment matrix.

Given the data matrix $X_{p \times n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, the optimal estimator that maximizes the probability of recovering the clustering labels correctly is the maximum a posteriori (MAP) estimator. If the label assignment is uniformly random, then the MAP estimator is equivalent to the maximum likelihood estimator (MLE), where the log-likelihood function is given by

$$\ell(H, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) = -\frac{np}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2^2.$$

Then the MLE corresponds to the solution of

$$\min_{H, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K} \sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2^2 \tag{2}$$

subject to the constraint that H is an assignment matrix.

Since we focus on the recovery of the true clustering structure $G_1^*, \ldots, G_K^*$, we may first profile the "nuisance parameters" $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, whose MLEs are given by

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^n h_{ik}\mathbf{X}_i}{\sum_{i=1}^n h_{ik}} = \frac{1}{|G_k|}\sum_{i \in G_k} \mathbf{X}_i,$$

where $|G_k| = \sum_{i=1}^n h_{ik}$ denotes the cardinality of the k-th cluster. Substituting $\hat{\boldsymbol{\mu}}_k$ into (2), we see that the MLE for H (and thus for $G_1, \ldots, G_K$) is the solution of the constrained combinatorial optimization problem:

$$\max_{G_1, \ldots, G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \langle \mathbf{X}_i, \mathbf{X}_j \rangle \quad \text{subject to} \quad \bigsqcup_{k=1}^{K} G_k = [n], \tag{3}$$

where $\sqcup$ denotes the disjoint union.

It is now clear that, under the Gaussian mixture model, the MLE in (3) is equivalent to the classical K-means clustering method [1], which minimizes the total intra-cluster squared Euclidean distances. Since the K-means clustering problem is known to be worst-case NP-hard [2], [3], one can expect that a polynomial-time algorithm for computing the MLE of the clustering structure with exact solutions only exists in certain cases. Because of this computational barrier of the original K-means problem, various computationally tractable approximation algorithms are proposed in literature.

A widely used algorithm for solving the K-means is Lloyd's algorithm [4], which is an iterative algorithm that sequentially refines the partition structure to ensure that the K-means objective function is monotonically decreasing. Lloyd's algorithm has a similar nature as the classical expectation-maximization (EM) algorithm [5] in that, while the EM implicitly performs soft clustering at every E-step, Lloyd's algorithm does hard clustering at each iteration via the Voronoi diagram.

Manuscript received January 5, 2020; revised November 30, 2020; accepted February 23, 2021. Date of publication March 2, 2021; date of current version May 20, 2021. The work of Xiaohui Chen was supported in part by NSF CAREER Award DMS-1752614, in part by UIUC Research Board Award RB18099, and in part by a Simons Fellowship. The work of Yun Yang was supported in part by NSF DMS-1907316. *(Corresponding author: Xiaohui Chen.)*

The authors are with the Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: xhchen@illinois.edu; yy84@illinois.edu).

Communicated by S. Boucheron, Associate Editor for Statistical Learning.
Digital Object Identifier 10.1109/TIT.2021.3063155

Given a suitable initialization (such as the spectral clustering method [6]), it is shown in [7] that the clustering error for Lloyd's algorithm converges to zero exponentially fast, provided that

$$\Delta^2 := \min_{1\leqslant k\neq l\leqslant K} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2^2 \geqslant C\sigma^2 \frac{Kn}{\underline{n}}\Big(1 \vee \frac{Kp}{n}\Big), \quad (4)$$

where $\underline{n} = \min_{k\in[K]} |G_k^*|$ is the minimal cluster size and $a \vee b = \max(a, b)$.

Separation lower bound in (4) is not sharp (in the high-dimensional setting when $p \gg n$). In the simplest symmetric two-component Gaussian mixture model:

$$\mathbf{X}_i = \boldsymbol{\mu}\eta_i + \boldsymbol{\varepsilon}_i,$$

where $\eta_i = 1, i \in G_1^*$ and $\eta_i = -1, i \in G_2^*$, [8] proposes a simple iterative thresholding algorithm that achieves the sharp threshold on $\|\boldsymbol{\mu}\|_2^2$ for exact recovery with high probability, which is given by

$$\sigma^2\left(1 + \sqrt{1 + \frac{2p}{n\log n}}\right)\log n. \quad (5)$$

It should be noted that the algorithm in [8] critically depends on the symmetry of the Gaussian centers (i.e., $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$) and it is structurally difficult to extend such algorithm with maintained statistical optimality to a general K-component Gaussian mixture model without assuming the centers are equally spaced.

Another active line of research focuses on various convex relaxed versions of the K-means problem that is solvable in polynomial-time [9]–[15]. The best known rate of convergence achieved by the semidefinite programming (SDP) relaxed K-means for the Gaussian mixture model (1) is given by [14]. Specifically, it is shown therein that misclassification errors of the SDP originally proposed in [9] for relaxing the K-means has the exponential rate of convergence $\exp(-C \cdot \mathsf{SNR}^2)$, where the signal-to-noise ratio is defined as

$$\mathsf{SNR}^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{\underline{n}\Delta^4}{p\sigma^4} \geqslant c\frac{n}{\underline{n}} \quad (6)$$

and $a \wedge b = \min(a, b)$. In particular, the exponential rate implies that exact recovery is achieved by the SDP relaxed K-means with high probability in the equal cluster size case $\underline{n} = n/K$ if minimal separation of cluster centers satisfies the lower bound

$$\Delta^2 \geqslant C\sigma^2\left(1 \vee \sqrt{\frac{Kp}{n\log n}}\right)\log n. \quad (7)$$

Now comparing (7) with the optimal exact results (5) in the special symmetric two-component Gaussian mixture model, it is natural to ask the following question:

> *does the SDP relaxed K-means clustering method achieve a sharp threshold for exact recovery of the general K-component Gaussian mixture model?*

To the best knowledge of ours, this is an open question in literature. In this paper, we provide an affirmative answer to this question: we show that there is an SDP relaxation of the K-means clustering method (given in (11) below) achieving the exact recovery with high probability if $\Delta^2 \geqslant (1+\alpha)\overline{\Delta}^2$, where

$$\overline{\Delta}^2 = 4\sigma^2\left(1 + \sqrt{1 + \frac{Kp}{n\log n}}\right)\log n. \quad (8)$$

In addition, if $\Delta^2 \leqslant (1-\alpha)\overline{\Delta}^2$, then the probability of exact recovery for any estimator vanishes to zero under the equal cluster size scenario. Thus $\overline{\Delta}^2$ yields the information-theoretic cutoff value on the minimal separation of cluster centers for exact recovery of the K-component Gaussian mixture model, and the SDP relaxation for the K-means is minimax-optimal in the sense that sharp phase transition of the probability of wrong recovery from zero to one occurs at the critical threshold given by the $\overline{\Delta}^2$.

A. Related Work

There is a vast literature studying the clustering problem on the Gaussian mixture model, or more generally finite mixture models. Regarding clustering labels as missing data, parameter estimation is often carried out by the EM algorithm [5], [16]. The EM algorithm has been extensively studied in the statistics and machine learning literature [17]–[25]. Optimal rate of convergence for estimating the mixing distribution in finite mixture models is derived in [17]. Consistency of the K-means estimation of the clustering centers is studied in [1], [26], without concerning the computational complexity. Computationally efficient algorithms for solving the K-means include Lloyd's algorithm [4], [7] and convex relaxations [9]–[15], [27]. Other popular clustering methods include the spectral clustering [28]–[35] and variants of the K-means [36]–[40]. Analysis under the mixture models has also been done under other clustering models such as the stochastic ball models [12], [27], [38].

Parallel to the (mixture) model-based clustering framework, there are many similar methods and algorithms proposed for community detection in network data based on the stochastic block model (SBM) [41], [42]. Successful algorithms for community detection, partial and exact recovery under the SBM have been extensively studied in literature – these include spectral algorithms [43]–[45], SDP relaxations [46]–[52], among others [53], [54].

B. Notation

Let $\mathbf{1}_n$ be the $n \times 1$ vector of all ones. For two matrices A and B of the same size, let $\langle A, B\rangle = \mathrm{tr}(A^T B)$ be the usual inner product. Throughout the rest of the paper, we fix the notation $n_k = |G_k^*|$, $m = \min_{1\leqslant k\neq l\leqslant K}\left\{\frac{2n_k n_l}{n_k+n_l}\right\}$, and $\underline{n} = \min_{k\in[K]} n_k$ as the minimal cluster size.

II. Main Result

In this section, we state our main result on the information-theoretic cutoff value of the exact recovery of the Gaussian mixture model in (1).

Theorem II.1 (Separation Upper Bound for Exact Recovery via SDP Relaxation): If there exist constants $\delta > 0$ and $\beta \in (0, 1)$ such that

$$\log n \geqslant \frac{(1-\beta)^2}{\beta^2}\frac{C_1 n}{m}, \quad \delta \leqslant \frac{\beta^2}{(1-\beta)^2}\frac{C_2}{K}, \quad m \geqslant \frac{4(1+\delta)^2}{\delta^2},$$

and

$$\Delta^2 \geqslant \frac{4\sigma^2(1+2\delta)}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\delta)}\frac{p}{m\log n}+C_3R_n}\right)\log n$$

with

$$R_n = \frac{(1-\beta)^2}{(1+\delta)\log n}\left(\frac{\sqrt{p\log n}}{\underline{n}}+\frac{\log n}{\underline{n}}\right),$$

then the SDP in (11) achieves exact recovery with probability at least $1 - C_4\, K^2\, n^{-\delta}$, where C_i, $i = 1, 2, 3, 4$, are universal constants.

The following corollary is a direct consequence (and a special case) of Theorem II.1 when the cluster sizes are equal.

Corollary II.2: Let $\alpha > 0$, $\Delta^2 = \min_{1\leqslant k\neq l\leqslant K} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2^2$, and $\overline{\Delta}^2$ be defined in (8). Suppose that the cluster sizes are equal and $K \leqslant C_1 \log(n)/\log\log(n)$ for some small constant $C_1 > 0$ depending only on α. If $\Delta^2 \geqslant (1+\alpha)\overline{\Delta}^2$, then the SDP in (11) achieves exact recovery with probability at least $1 - C_2(\log n)^{-c_3}$, where C_2, c_3 are constants depending only on α.

To derive a lower bound, we focus on the equal size case where clusters $\{G_k^*\}_{k=1}^K$ have roughly the same sizes. More precisely, recall that our unknown parameters are the cluster indicating variables $H = \{h_{ik} : i \in [n], k \in [K]\}$, and $\{n_k : k \in [K]\}$ are the unknown cluster sizes. Let $\delta_n = C\sqrt{K\log(n)/n}$ for some sufficiently large constant $C > 0$. Here, we consider $n_k \in [(1-\delta_n)\, n/K, (1+\delta_n)\, n/K]$ for $k \in [K]$ that allows a small fluctuation on the community size in establishing the lower bound. Particularly, we define the (localized) parameter space as

$$\Theta\big(n, K, \Delta\big) = \Big\{\big(\{h_{ik}\}, \{\boldsymbol{\mu}_k\}\big) : h_{ik} \in \{0,1\}, \boldsymbol{\mu}_k \in \mathbb{R}^p, \sum_{k=1}^K h_{ik} = 1,\ n_k := \sum_{i=1}^n h_{ik} \in \Big[(1-\delta_n)\,\frac{n}{K}, (1+\delta_n)\,\frac{n}{K}\Big], \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\| \geqslant \Delta,\ \forall i \in [n] \text{ and } \forall (k,l) \in [K]^2,\ k \neq l\Big\}.$$

Theorem II.3 (Separation Lower Bound for Exact Recovery: Equal Cluster Size Case): Let $\alpha \in (0,1)$. If $\Delta^2 \leqslant (1-\alpha)\overline{\Delta}^2$ and $K \leqslant \log n$, then we have

$$\inf_{\{\hat{h}_{ik}\}} \sup_{(H,\boldsymbol{\mu})\in\Theta(n,K,\Delta)} \mathbb{P}_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik} \neq h_{ik},\, i \in [n],\, k \in [K]\big) \geqslant 1 - cKn^{-1},$$

where $c > 0$ is a constant depending only on α and the infimum is over all possible estimators $\{\hat{h}_{ik}\}$ for $\{h_{ik}\}$.

Corollary II.2 and Theorem II.3 together imply that in the equal cluster size case when $n_1 = n_2 = \cdots = n_K = \frac{n}{K}$, the SDP relaxation (11) for the K-means is minimax-optimal in the sense that sharp phase transition of the probability of wrong recovery from zero to one occurs at the critical threshold given by the $\overline{\Delta}^2$ in (8).

III. Semidefinite Programming Relaxation: Primal and Dual

In this section, we describe the SDP relaxation of the K-means that achieves the cutoff value of the exact recovery and outline the strategy of showing that the SDP solution uniquely recovers the true clustering structure by a dual certificate argument via the primal-dual construction. We remark that similar primal-dual analyses are done in [11], [55].

Let $A = X^TX$ be the affinity matrix and $B = \mathrm{diag}(|G_1|^{-1}, \ldots, |G_K|^{-1})$. Then we can reparametrize (3) as

$$\max_H \langle A, HBH^T\rangle \quad \text{subject to } H \in \{0,1\}^{n\times K},\ H\mathbf{1}_K = \mathbf{1}_n, \tag{9}$$

which is a mixed integer program with a nonlinear objective function [9], [56]. If the cluster centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ are properly separated, then the affinity matrix A from the data has an approximate block diagonal structure (up to a permutation of the data index).

Changing variable $Z = HBH^T$, we observe that the $n \times n$ symmetric matrix Z satisfies the following properties:

(P1) positive semidefinite (psd) constraint: $Z \succeq 0$;
(P2) non-negative (entrywise) constraint: $Z \geqslant 0$, i.e., $Z_{ij} \geqslant 0$ for all $i, j \in [n]$;
(P3) unit row-sum constraint: $Z\mathbf{1}_n = \mathbf{1}_n$;
(P4) trace constraint: $\mathrm{tr}(Z) = K$.

Since Z is symmetric, properties (P2) and (P3) automatically ensure that Z is a stochastic matrix $Z\mathbf{1}_n = Z^T\mathbf{1}_n = \mathbf{1}_n$. Given any clustering structure $G_1, \ldots, G_K$, we may consider the associated cluster membership matrix:

$$Z_{ij} = \begin{cases} 1/|G_k| & \text{if } i,j \in G_k \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

Thus to recover the true clustering structure $G_1^*, \ldots, G_K^*$, it suffices to compare the estimated membership matrix and the true one Z^*.

After the change-of-variables, the objective function in (9) becomes linear in Z. Then we use the solution $\hat{Z}$ of the following (convex) SDP to estimate Z^*:

$$\hat{Z} = \mathrm{argmax}_{Z\in\mathscr{C}_K}\ \langle A, Z\rangle, \tag{11}$$

where

$$\mathscr{C}_K = \Big\{Z \in \mathbb{R}^{n\times n} \;\Big|\; Z\succeq 0,\ Z^T = Z,\ \mathrm{tr}(Z) = K,\ Z\mathbf{1}_n = \mathbf{1}_n,\ Z \geqslant 0\Big\}.$$

Note that the above SDP is first proposed in [9] and later studied in [14], [39], [40]. For spherical Gaussians (i.e., the noise covariance matrix is proportional to the identity matrix), since the SDP relaxation (11) does not require the knowledge of the noise variance σ^2 and the partition information other than the number of clusters K, it in fact can handle the more general case of unequal cluster sizes.

Remark III.1 (Adaptation to the Number of Clusters K): The SDP in (11) can be made adaptive to the unknown number of cluster K. When the number of clusters K is unknown, the constraint $\mathrm{tr}(Z) = K$ in the SDP (11) can be lifted to a penalization term in its objective function, i.e., we solve

$$\tilde{Z}_\lambda := \arg\max\{\langle A, Z\rangle - \lambda \mathrm{tr}(Z) : Z\succeq 0, Z^T = Z, Z\mathbf{1}_n = \mathbf{1}_n, Z \geqslant 0\}, \tag{12}$$

where $\lambda \geqslant 0$ is a regularization parameter. This is the *regularized K-means* proposed by [13], [15] and analyzed by [40] in the manifold clustering setting. Using the same

existing argument for proving the separation upper bound in Section IV, we see that with the λ choice being

$$\begin{aligned}&\sigma^2(\sqrt{n}+\sqrt{p}+\sqrt{2\log n})^2\\&\quad+C\beta^{-1}\sigma^2\left(n+K\log n+(1-\beta)K\delta\sqrt{pm\log n}\right)\\&\qquad\leqslant\lambda\leqslant p\sigma^2+\frac{\beta}{4}m\Delta^2,\end{aligned}\tag{13}$$

then under the same conditions in Theorem 2.1, $\tilde{Z}_\lambda = Z^*$ achieves exact recovery with probability at least $1-CK^2n^{-\delta}$. Note that a larger signal-to-noise ratio Δ^2/σ^2 permits a wider allowable range for λ to achieve exact recovery, and our conditions in Theorem II.1 ensures the existence of at least one such λ. The idea for $\tilde{Z}_\lambda$ to achieve the sharp threshold (i.e., the separation upper bound) is that the SDP giving $\hat{Z}$ in (11) and its regularized version in (12) have the same Lagrangian form and the dual problem. Thus we need only to extract the regime of the regularization parameter λ in (13) that ensures a successful dual certificate construction as characterized in $\hat{Z}$ (Section IV). In particular, the dual certificate constructed for $\hat{Z}$ is a convenient choice of $\lambda^\sharp$ that falls into the region (13) with high probability. In addition, [40] provides a practical method for adaptively tuning this regularization parameter λ. ■

Note that Z^* is a rank-K block diagonal matrix, and for any $Z\in\mathscr{C}_K$, due to the psd constraint, $\mathrm{tr}(Z)$ equals to the nuclear norm $\|Z\|_*$. Then the SDP in (11) can be effectively viewed as a low-rank matrix denoising procedure for the data affinity matrix A by finding its optimal matching from all feasible "rank-K" stochastic matrices proxied by the trace constraint.

On the other hand, the SDP solutions are not integral in general. If this is the scenario, then the standard relaxing-and-rounding paradigm [57] can be used to round the SDP solution back to a point in the feasible set of the original discrete optimization problem (3). In our case, we can apply the K-means clustering to the top K-eigenvectors of $\hat{Z}$ as a rounding procedure to extract the estimated partition structure $\hat{G}_1,\ldots,\hat{G}_K$.

However, it is observed that the rounding step is not always necessary and solution to the clustering problem (3) can be directly recovered from solving the relaxed SDP problems when the separation of cluster centers is large, which is sometimes referred to the *exact recovery* or *hidden integrality* phenomenon [12], [27]. This motivates the question we asked earlier in Section I that when and to what extend the SDP relaxation can in fact produce the exact recovery. The rest of the paper is devoted to characterize the precise cutoff value on the separation of cluster centers that yields the exact recovery.

A. Dual Problem

To analyze the exact recovery property of $\hat{Z}$, we first derive the dual problem for the (primal) SDP problem in (11). Let

$$\begin{aligned}&\mathcal{L}(Z,Q,\lambda,\boldsymbol{\alpha},B)\\&=\mathrm{tr}(AZ)+\mathrm{tr}(QZ)+\lambda(K-\mathrm{tr}(Z))\\&\quad+\boldsymbol{\alpha}^T\left(\mathbf{1}_n-\frac{Z+Z^T}{2}\mathbf{1}_n\right)+\mathrm{tr}(BZ)\\&=(\lambda K+\boldsymbol{\alpha}^T\mathbf{1}_n)+\mathrm{tr}\left\{\left[A+Q-\lambda\mathrm{Id}_n+B-\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)\right]Z\right\}\end{aligned}$$

be the Lagrangian function, where $Q_{n\times n}\succeq 0$, $\boldsymbol{\alpha}_{n\times 1}=(\alpha_1,\ldots,\alpha_n)^T$, $B_{n\times n}\geqslant 0$, and $\lambda\in\mathbb{R}$ are the Lagrangian multipliers. Consider the max-min problem:

$$\max_{Z\in\mathbb{R}^{n\times n}}\ \min_{Q\succeq 0,\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\geqslant 0}\mathcal{L}(Z,Q,\lambda,\alpha,B),$$

where the maximum over Z is unconstrained. If Z is not primal feasible for the SDP problem (11), then

$$\min_{Q\succeq 0,\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\geqslant 0}\mathcal{L}(Z,Q,\lambda,\alpha,B)=-\infty.$$

For example, consider $\mathrm{tr}(Z)\neq K$ and choose $\lambda=-\frac{c}{K-\mathrm{tr}(Z)}$ with an arbitrarily large $c>0$. On the other hand, if Z is feasible for (11), then

$$\mathrm{tr}(AZ)\leqslant\min_{Q\succeq 0,\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\geqslant 0}\mathcal{L}(Z,Q,\lambda,\alpha,B),$$

where the equality is attained if for example $Q=B=0$. Then,

$$\begin{aligned}\max_{Z\in\mathscr{C}_K}\mathrm{tr}(AZ)&\leqslant\max_{Z\in\mathbb{R}^{n\times n}}\ \min_{Q\succeq 0,\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\geqslant 0}\mathcal{L}(Z,Q,\lambda,\alpha,B)\\&\leqslant\min_{Q\succeq 0,\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\geqslant 0}\ \max_{Z\in\mathbb{R}^{n\times n}}\mathcal{L}(Z,Q,\lambda,\alpha,B).\end{aligned}$$

Similarly, if $A+Q-\lambda\mathrm{Id}_n+B-\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)\neq 0$, then

$$\max_{Z\in\mathbb{R}^{n\times n}}\mathrm{tr}\left\{\left[A+Q-\lambda\mathrm{Id}_n+B-\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)\right]Z\right\}=\infty,$$

which is avoided by the minimization over the Lagrangian multipliers. Thus with $Q=\lambda\mathrm{Id}_n+\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)-B-A$, we have

$$\begin{aligned}\max_{Z\in\mathscr{C}_K}\ \mathrm{tr}(AZ)\leqslant\ &\min_{\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\in\mathbb{R}^{n\times n}}\Big\{\lambda K+\boldsymbol{\alpha}^T\mathbf{1}_n:B\geqslant 0,\\&\lambda\mathrm{Id}_n+\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)-B-A\succeq 0\Big\},\end{aligned}$$

which is the weak duality between the primal SDP problem (11) and its dual problem:

$$\begin{aligned}&\min\nolimits_{\lambda\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,B\in\mathbb{R}^{n\times n}}\ \{\lambda K+\boldsymbol{\alpha}^T\mathbf{1}_n\}\\&\text{subject to}\ \ B\geqslant 0,\\&\qquad\lambda\mathrm{Id}_n+\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)-B-A\succeq 0.\end{aligned}\tag{14}$$

Moreover, the duality gap is given by

$$\begin{aligned}&\lambda K+\boldsymbol{\alpha}^T\mathbf{1}_n-\mathrm{tr}(AZ)\\&=\lambda\,\mathrm{tr}(Z)+\alpha^T\frac{Z+Z^T}{2}\mathbf{1}_n-\mathrm{tr}(AZ)\\&=\mathrm{tr}\left\{\left[\lambda\mathrm{Id}_n+\frac{1}{2}(\mathbf{1}_n\boldsymbol{\alpha}^T+\alpha\mathbf{1}_n^T)-A-B\right]Z\right\}+\mathrm{tr}(BZ)\\&\geqslant\mathrm{tr}(BZ)\geqslant 0.\end{aligned}\tag{15}$$

B. Optimality Conditions: Primal-Dual Construction

Let $\mathbf{1}_{G_k^*}$ be the $n\times 1$ vector such that it is equal to $\mathbf{1}_{n_k}$ on G_k^* and zero otherwise. To show that

$$Z^*=\begin{bmatrix}\frac{1}{n_1}J_{n_1}&0&\cdots&0\\0&\frac{1}{n_2}J_{n_2}&\cdots&0\\\vdots&\ddots&\cdots&\vdots\\0&\cdots&0&\frac{1}{n_K}J_{n_K}\end{bmatrix}=\sum_{k=1}^K\frac{1}{n_k}\mathbf{1}_{G_k^*}\mathbf{1}_{G_k^*}^T\tag{16}$$

is the solution of the primal SDP problem (11), we need the duality gap (15) is zero at $Z = Z^*$. To this end, we need to construct a *dual certificate* $(\lambda, \boldsymbol{\alpha}, B)$ such that:

(C1) $B \geqslant 0$;
(C2) $W_n := \lambda \mathrm{Id}_n + \frac{1}{2}(\mathbf{1}_n \boldsymbol{\alpha}^T + \alpha \mathbf{1}_n^T) - A - B \succeq 0$;
(C3) $\mathrm{tr}(W_n Z^*) = 0$;
(C4) $\mathrm{tr}(BZ^*) = 0$.

Note that (C1) and (C2) are dual feasibility constraints, while (C3) and (C4) are the optimality conditions (i.e., complementary slackness) corresponding to the zero duality gap in (15). In particular, (C4) implies that $B_{G_k^* G_k^*} = 0$ for all $k \in [K]$.

To ensure that Z^* is the *unique* solution of the SDP problem (11), we observe that Z^* is the only feasible matrix to the SDP (11) satisfying the block diagonal structure

$$\begin{bmatrix} Z^{(1)} & 0 & \cdots & 0 \\ 0 & Z^{(2)} & \cdots & 0 \\ \vdots & \ddots & \cdots & \vdots \\ 0 & \cdots & 0 & Z^{(K)} \end{bmatrix},$$

i.e., $Z_{G_k^* G_l^*} = 0$ for all distinct pair $(k,l) \in [K]^2$. Indeed, since each block $Z^{(k)}$ satisfies $Z^{(k)} \mathbf{1}_{n_k} = \mathbf{1}_{n_k}$ and is psd, $(1, {n_k}^{-1/2} \mathbf{1}_{n_k})$ is one eigenvalue-eigenvector pair of $Z^{(k)}$ and the trace of $Z^{(k)}$ is at least 1. On the other hand, due to the trace constraint $\sum_{k=1}^K \mathrm{tr}(Z^{(k)}) = \mathrm{tr}(Z) = k$, we then must have $\mathrm{tr}(Z^{(k)}) = 1$. In addition, 1 is its only nonzero eigenvalue with eigenvector ${n_k}^{-1/2} \mathbf{1}_{n_k}$. Consequently, $Z^{(k)}$ must take the form of $n_k^{-1} J_{n_k}$.

Given the above block diagonal structure and $\mathrm{tr}(BZ^*) = 0$, we conclude that Z^* is the unique solution to the SDP (11) if

(C5) $B_{G_k^* G_l^*} > 0$ for all distinct pair $(k,l) \in [K]^2$,

in addition to the optimality conditions (C1)-(C4).

IV. Proof of Theorem II.1

In this section, we show that a dual certificate described in Section III-B can be successfully constructed with high probability, thus proving Theorem II.1. First, observe that $W_n \succeq 0$ and $\mathrm{tr}(W_n Z^*) = 0$ imply that

$$W_n \mathbf{1}_{G_k^*} = 0 \quad \text{for all } k \in [K]. \tag{17}$$

The last display together with $B_{G_k^* G_k^*} = 0$ imply that for each distinct pair $(k,l) \in [K]^2$,

$$\lambda \mathbf{1}_{n_k} + \frac{1}{2}\mathbf{1}_{n_k}\Big(\sum_{i \in G_k^*} \alpha_i\Big) + \frac{1}{2}\boldsymbol{\alpha}_{G_k^*} n_k = A_{G_k^* G_k^*} \mathbf{1}_{n_k}, \tag{18}$$

$$\frac{1}{2}\mathbf{1}_{n_l}\Big(\sum_{i \in G_k^*} \alpha_i\Big) + \frac{1}{2}\boldsymbol{\alpha}_{G_l^*} n_k - A_{G_l^* G_k^*} \mathbf{1}_{n_k} = B_{G_l^* G_k^*} \mathbf{1}_{n_k}, \tag{19}$$

where $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_{G_1^*}^T, \ldots, \boldsymbol{\alpha}_{G_K^*}^T)$. From (18), we get

$$\sum_{i \in G_k^*} \alpha_i = \frac{1}{n_k} \mathbf{1}_{n_k}^T A_{G_k^* G_k^*} \mathbf{1}_{n_k} - \lambda.$$

Substituting the last equation back into (18), we get

$$\boldsymbol{\alpha}_{G_k^*} = \frac{2}{n_k} A_{G_k^* G_k^*} \mathbf{1}_{n_k} - \frac{\lambda}{n_k} \mathbf{1}_{n_k} - \frac{1}{n_k^2} \mathbf{1}_{n_k} (\mathbf{1}_{n_k}^T A_{G_k^* G_k^*} \mathbf{1}_{n_k}). \tag{20}$$

Next we construct a solution of B for (19). For $k \neq l$, we have

$$\begin{aligned} B_{G_l^* G_k^*} \mathbf{1}_{n_k} =& -\frac{n_l + n_k}{2n_l} \lambda \mathbf{1}_{n_l} + \frac{1}{2n_k} (\mathbf{1}_{n_k}^T A_{G_k^* G_k^*} \mathbf{1}_{n_k}) \mathbf{1}_{n_l} \\ &+ \frac{n_k}{n_l} A_{G_l^* G_l^*} \mathbf{1}_{n_l} - \frac{n_k}{2n_l^2} (\mathbf{1}_{n_l}^T A_{G_l^* G_l^*} \mathbf{1}_{n_l}) \mathbf{1}_{n_l} - A_{G_l^* G_k^*} \mathbf{1}_{n_k}. \end{aligned}$$

In particular, for $j \in G_l^*$,

$$\begin{aligned} & [B_{G_l^* G_k^*} \mathbf{1}_{n_k}]_j \\ =& -\frac{n_l + n_k}{2n_l} \lambda + \frac{1}{2n_k} \sum_{s,t \in G_k^*} \mathbf{X}_s^T \mathbf{X}_t + \frac{n_k}{n_l} \sum_{t \in G_l^*} \mathbf{X}_j^T \mathbf{X}_t \\ & - \frac{n_k}{2n_l^2} \sum_{s,t \in G_l^*} \mathbf{X}_s^T \mathbf{X}_t - \sum_{t \in G_k^*} \mathbf{X}_j^T \mathbf{X}_t \\ =& -\frac{n_l + n_k}{2n_l} \lambda + \frac{n_k}{2} (\overline{\mathbf{X}}_k^T \overline{\mathbf{X}}_k - \overline{\mathbf{X}}_l^T \overline{\mathbf{X}}_l) + n_k \mathbf{X}_j^T (\overline{\mathbf{X}}_l - \overline{\mathbf{X}}_k) \\ =& -\frac{n_l + n_k}{2n_l} \lambda + \frac{n_k}{2} (\|\overline{\mathbf{X}}_k - \mathbf{X}_j\|_2^2 - \|\overline{\mathbf{X}}_l - \mathbf{X}_j\|_2^2), \end{aligned} \tag{21}$$

where $\overline{\mathbf{X}}_k = n_k^{-1} \sum_{i \in G_k^*} \mathbf{X}_i$ is the empirical mean of data points in the k-th cluster. Without loss of generality, we may take a symmetric B (i.e., $B^T = B$) and then construct B as block-wise rank-one matrix satisfying the above row sum constraint (21):

$$B^\sharp_{G_l^* G_k^*} = \frac{B_{G_l^* G_k^*} \mathbf{1}_{G_k^*} \mathbf{1}_{G_l^*}^T B_{G_l^* G_k^*}}{\mathbf{1}_{G_l^*}^T B_{G_l^* G_k^*} \mathbf{1}_{G_k^*}} \tag{22}$$

for each distinct pair $(k,l) \in [K]^2$ and $B^\sharp_{G_k^* G_k^*} = 0$. For notational simplicity, let us denote the column sums and row sums of matrix $B_{G_k^* G_l^*}$ in (22) by $\mathbf{c}^{(k,l)} = \big(c_j^{(k,l)} : j \in G_l^*\big)$ and $\mathbf{r}^{(k,l)} = \big(r_i^{(k,l)} : i \in G_k^*\big)$, respectively. In addition, by letting $t^{(k,l)} = \sum_{j \in G_l^*} c_j^{(k,l)} = \sum_{i \in G_k^*} r_i^{(k,l)}$ be the total sum, then the construction in (22) becomes $[B_{G_l^* G_k^*}]_{ij} = r_i^{(k,l)} c_j^{(k,l)} / t^{(k,l)}$. For convenience, we also define $r_i^{(k,k)} = c_j^{(k,k)} = t^{(k,k)} = 0$ for all $i, j \in G_k^*$, so that $B_{G_k^* G_k^*} = 0$ for all $k \in [K]$ (define $0/0 = 0$).

Recall that to ensure uniqueness, we need to choose λ such that $B^\sharp_{G_k^* G_l^*} > 0$ for all distinct pair $(k,l) \in [K]^2$, which is, in view of (21), guaranteed whenever

$$\lambda < \min_{1 \leqslant k \neq l \leqslant K} \left\{ \frac{n_l n_k}{n_l + n_k} \min_{j \in G_l^*} (\|\overline{\mathbf{X}}_k - \mathbf{X}_j\|_2^2 - \|\overline{\mathbf{X}}_l - \mathbf{X}_j\|_2^2) \right\}. \tag{23}$$

On the other hand, we require that λ is not too small since $W_n = \lambda \mathrm{Id}_n + \frac{1}{2}(\mathbf{1}_n \boldsymbol{\alpha}^T + \boldsymbol{\alpha} \mathbf{1}_n^T) - A - B \succeq 0$. To identify the right λ, we will employ the following lemma that provides some high probability lower bounds that will be useful for bounding from below the column sums $\{c_j^{(k,l)} : j \in G_l^*\}$ and row sums $\{r_i^{(k,l)} : i \in G_k^*\}$ under proper separation conditions on the Gaussian centers. Recall that

$$\Delta = \min_{1 \leqslant k \neq l \leqslant K} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\| \quad \text{and} \quad m = \min_{1 \leqslant k \neq l \leqslant K} \left\{ \frac{2 n_k n_l}{n_k + n_l} \right\}.$$

Note that Δ is the minimum separation between the cluster centers and m quantifies the "minimum" cluster size in the pairwise sense.

Lemma IV.1 (Separation Bound on the Gaussian Centers): Let $\delta > 0$ and $1 > \beta > 0$. If there exists a sufficiently large universal constant $c_1 > 0$ such that

$$\Delta^2 \geqslant \frac{4\sigma^2(1+2\delta)}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\delta)}\frac{p}{m\log n}+c_1 R_n}\right)\log n \tag{24}$$

with

$$R_n = \frac{(1-\beta)^2}{(1+\delta)\log n}\left(\frac{\sqrt{p\log(nK)}}{\underline{n}}+\frac{\log(nK)}{\underline{n}}\right),$$

then as long as $m \geqslant 4(1+\delta^{-1})^2$,

$$\mathbb{P}\Big(\|\mathbf{X}_i-\overline{\mathbf{X}}_l\|^2-\|\mathbf{X}_i-\overline{\mathbf{X}}_k\|^2 \geqslant \frac{n_k+n_l}{n_k n_l}\sigma^2 p+\beta\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2-r_{kl},$$
$$\text{for all distinct pairs } (k,l)\in[K]^2 \text{ and } i\in G_k^*\Big) \leqslant \frac{K^2}{n^\delta}+\frac{8}{n},$$

where

$$r_{kl} = 2\sigma\sqrt{\frac{2\log(nK)}{n_l}}\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\| + 2\sigma^2\frac{n_k+n_l}{n_k n_l}\sqrt{2p\log(nK)}+\frac{4\sigma^2}{n_k}\log(nK).$$

If the conditions of Lemma IV.1 holds, then according to this lemma we may choose

$$\lambda^\sharp = p\sigma^2+\frac{\beta}{4}m\Delta^2, \tag{25}$$

so that it holds with probability at least $1-2n^{-\delta}-10n^{-1}$ that for all $k,l\in[K]$, $k\neq l$, $i\in G_k^*$, $j\in G_l^*$,

$$\begin{aligned} r_i^{(k,l)} &\geqslant \frac{\beta}{2}n_l\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2,\\ c_j^{(k,l)} &\geqslant \frac{\beta}{2}n_k\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2,\\ t^{(k,l)} &\geqslant \frac{\beta}{2}n_k n_l\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2, \end{aligned} \tag{26}$$

as long as $m \geqslant 4(1+\delta^{-1})^2$. This implies $B^\sharp_{G_k^* G_l^*} > 0$ for any distinct pair $(k,l)\in[K]^2$. We fix such a choice for λ in the rest of the proof.

Denote $\Gamma_K = \text{span}\{\mathbf{1}_{G_k^*} : k\in[K]\}^\perp$ be the orthogonal complement of the linear subspace of $\mathbb{R}^n$ spanned by the vectors $\mathbf{1}_{G_1^*},\ldots,\mathbf{1}_{G_K^*}$. In view of (17), we see that $\{\mathbf{1}_{G_k^*} : k\in[K]\}$ are eigenvectors of W_n associated to the zero eigenvalues. Thus to ensure $W_n \succeq 0$, we only need to check that: for any $\mathbf{v}=(v_1,\ldots,v_n)^T\in\Gamma_K$ such that $\|\mathbf{v}\|_2=1$,

$$\mathbf{v}^T W_n\mathbf{v} \geqslant 0.$$

Our next task is to derive a high probability lower bound for the quadratic form $\mathbf{v}^T W_n\mathbf{v}$. Plugging the definition of W_n, we write

$$\mathbf{v}^T W_n\mathbf{v} = \lambda\|\mathbf{v}\|^2+\frac{1}{2}(\mathbf{v}^T\mathbf{1}_n\boldsymbol{\alpha}^T\mathbf{v}+\mathbf{v}^T\boldsymbol{\alpha}\mathbf{1}^T\mathbf{v}) - \sum_{k,l=1}^K\sum_{i\in G_k^*}\sum_{j\in G_l^*}\mathbf{X}_i^T\mathbf{X}_j v_i v_j - \mathbf{v}^T B^\sharp\mathbf{v}.$$

Since $\mathbf{v}^T\mathbf{1}_{G_k^*}=0$ or $\sum_{i\in G_k^*}v_i=0$ for all $k\in[K]$ and $\mathbf{v}\in\Gamma_K$, we get

$$\mathbf{v}^T W_n\mathbf{v} = \lambda\|\mathbf{v}\|^2 - S(\mathbf{v}) - T(\mathbf{v}),$$

where $S(\mathbf{v}) := \|\sum_{k=1}^K\sum_{i\in G_k^*}\mathbf{X}_i v_i\|_2^2$ and $T(\mathbf{v}) = \mathbf{v}^T B^\sharp\mathbf{v}$. Recall the clustering model (1): $\mathbf{X}_i = \boldsymbol{\mu}_k+\boldsymbol{\varepsilon}_i$ for $i\in G_k^*$, we have

$$\sum_{i\in G_k^*}\mathbf{X}_i v_i = \boldsymbol{\mu}_k\sum_{i\in G_k^*}v_i+\sum_{i\in G_k^*}\boldsymbol{\varepsilon}_i v_i = \sum_{i\in G_k^*}\boldsymbol{\varepsilon}_i v_i.$$

so that

$$S(\mathbf{v}) = \sum_{i=1}^n\sum_{j=1}^n\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j^T v_i v_j$$

is a quadratic form in $\mathbf{v}$. Therefore, for each $\mathbf{v}\in\Gamma_K$ satisfying $\|\mathbf{v}\|=1$, $S(\mathbf{v})$ can be bounded by the largest singular value of the Gram matrix $G_n=\{\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j^T : i,j\in[n]\}$, so that $S(\mathbf{v}) = \mathbf{v}^T\mathcal{E}^T\mathcal{E}\mathbf{v} \leqslant \|\mathcal{E}^T\mathcal{E}\|_{\text{op}} = \|\mathcal{E}\|_{\text{op}}^2$, where matrix

$$\mathcal{E} = (\boldsymbol{\varepsilon}_1,\ \boldsymbol{\varepsilon}_2,\ \ldots,\ \boldsymbol{\varepsilon}_n)\in\mathbb{R}^{p\times n}$$

has i.i.d. $N(0,\sigma^2)$ entries. Applying Lemma VIII.2, we can reach

$$\mathbb{P}\Big(\max_{\mathbf{v}\in\Gamma_K,\,\|\mathbf{v}\|=1}S(\mathbf{v}) \geqslant \sigma^2(\sqrt{n}+\sqrt{p}+\sqrt{2t})^2\Big) \leqslant e^{-t},\quad \forall t>0.$$

Now we analyze the last term $T(\mathbf{v})$.

Lemma IV.2 (Bound on $T(\mathbf{v})$): Assume the separation condition (24) in Lemma IV.1 and consider the choice of λ as (25). We have for any $\delta>0$,

$$\mathbb{P}\Big(|T(\mathbf{v})| \geqslant C\beta^{-1}\sigma^2\big(n+K\log n+(1-\beta)K\delta\sqrt{mp\log n}\big)\|\mathbf{v}\|^2, \quad \forall\mathbf{v}\in\Gamma_K \;\Big|\; \{\bar{\boldsymbol{\varepsilon}}_k : k\in[K]\}\Big) \leqslant 4K^2 n^{-\delta}+10n^{-1}.$$

By combining previous bounds on $|S(\mathbf{v})|$ and $|T(\mathbf{v})|$ together, we obtain

$$\mathbb{P}\Big(\langle\mathbf{v},\,W_n\mathbf{v}\rangle \leqslant \lambda-\sigma^2(\sqrt{n}+\sqrt{p}+\sqrt{2\log n})^2 - C\beta^{-1}\sigma^2\big(n+K\log n+(1-\beta)K\delta\sqrt{mp\log n}\big), \quad \forall\mathbf{v}\in\Gamma_K,\,\|\mathbf{v}\|=1\Big) \leqslant (5K^2+1)\,n^{-\delta}.$$

Combining this with our constructions (25) for $\lambda^\sharp$, (20) for $\boldsymbol{\alpha}^\sharp$ and (22) for $B^\sharp$ and all previous analysis, we obtain that $(\lambda^\sharp,\boldsymbol{\alpha}^\sharp,B^\sharp)$ will be a dual certificate that satisfies (1)–(5) with probability at least $1-(5K^2+1)\,n^{-\delta}$ if

$$\begin{aligned} &\sigma^2(\sqrt{n}+\sqrt{p}+\sqrt{2\log n})^2 + C\beta^{-1}\sigma^2\,(n+K\log n+(1-\beta)K\delta\sqrt{pm\log n})\\ &\leqslant p\sigma^2+ \frac{\beta}{4}\,m\frac{4\sigma^2(1+2\delta)}{(1-\beta)^2}\Big(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\delta)}\frac{p}{m\log n}+c_1R_n}\Big)\log n. \end{aligned} \tag{27}$$

which is true if for some universal constants $C,c>0$,

$$\log n \geqslant \frac{(1-\beta)^2}{\beta^2}\frac{Cn}{m},\quad\text{and}\quad \delta\leqslant\frac{\beta^2}{(1-\beta)^2}\frac{c}{K}.$$

V. Proof of Theorem II.3

The first step is to reduce the worst-case misclassification risk to the average-case risk by putting a prior π^H over $H = \{h_{ik}\}$ with $(h_{i1}, \ldots, h_{iK})$ being i.i.d. following the multinomial distribution with one trial and probability vector $(n/K, \ldots, n/K)$. By the classical Chernoff bound we have

$$\mathbb{P}_{\pi^H}\Big(n_k := \sum_{i=1}^n h_{ik} \in \Big[(1-\delta_n)\,\frac{n}{K},\, (1+\delta_n)\,\frac{n}{K}\Big],\, k \in [K]\Big) \geqslant 1 - n^{-1}, \tag{28}$$

by choosing the constant C in δ_n large enough. As a consequence, we have (29), shown at the bottom of the next page. Conditioning on the event that $n_1 + n_2$ points belong to the first two clusters, the problem of correctly classifying all n samples into K clusters is always not easier than correctly classifying the n_1+n_2 points into the first and second clusters, that is,

$$\begin{aligned} &P_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik} \neq h_{ik},\, i \in [n],\, k \in [K]\big) \\ \geqslant &P_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik} \neq h_{ik},\, i \in G_1 \cup G_2,\, k \in [2]\big), \end{aligned}$$

where recall that $G_k = \{i \in [n] : h_{ik} = 1\}$ denote the k-th cluster. Now we apply the following minimax lower bound Lemma V.1 proved in Section V-A below for two clusters G_1 and G_2 conditioning on their total sizes n_1+n_2, we have (30), shown at the bottom of the next page, for some $c > 0$, where $\tilde{\pi}^{12}$ denote the conditional prior distribution of $\{\hat{h}_{ik}, i \in G_1 \cup G_2, k = 1, 2\}$ given the total sample size $n_1 + n_2$ of $G_1 \cup G_2$, which is uniform over $\{1, 2\}^{n_1+n_2}$. Here we have used the high probability bound (28) so that with probability at least $1 - n^{-1}$, the separation Δ satisfies

$$\Delta^2 \leqslant 4(1-\alpha/2)\sigma^2 \left(1 + \sqrt{1 + \frac{2p}{(n_1+n_2)\log(n_1+n_2)}}\right) \log n.$$

Note that the proof of Lemma V.1 also reduces the worst-case bound to the average-case bound, where the prior on the cluster label is uniform as the conditional distribution $\tilde{\pi}^{12}$ given the total size $n_1 + n_2$. Putting all pieces together and using $K \leqslant \log n$ give a proof of the claimed result.

A. Lower Bound for $K = 2$

Now we prove an information-theoretic limit for exact recovery of clusters labels in a symmetric two-component Gaussian mixture model,

$$\mathbf{X}_i = \eta_i \boldsymbol{\mu} + \sigma \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} N(0, I_p), \quad i = 1, \ldots, n, \tag{31}$$

where $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ are unknown centers of the two symmetric Gaussian components, and $\eta_i \in \{-1, 1\}$ is the label for the ith observation indicating which component it comes from.

Lemma V.1 (Separation Lower Bound for Exact Recovery: $K = 2$): Let $\alpha \in (0, 1)$. Consider the symmetric two-component Gaussian mixture model in (31) with an independent Rademacher prior distribution on η_i. If $\Delta^2 \leqslant (1-\alpha)\overline{\Delta}^2$, then

$$\inf_{\hat{\eta}} \sup_{\|\boldsymbol{\mu}\| \geqslant \Delta/2} \mathbb{P}(\hat{\eta} \neq \eta) \geqslant 1 - cn^{-1}, \tag{32}$$

where $c > 0$ is a constant depending only on α and the infimum is over all possible estimators $\hat{\eta}$ for $\eta = \{\eta_i\}_{i=1}^n \in \{\pm 1\}^n$.

Remark V.2: Our Lemma V.1 is stronger than the exact recovery notation in [8] and the probability of wrong recovery lower bound in (32) does not follows from the lower bound therein for the expected Hamming distance loss in the symmetric two-component Gaussian mixture model. Moreover, complementing the upper bound in Corollary II.2, the lower bound is sharply optimal in the sense that the probability of wrong recovery is arbitrarily close to one (rather than just bounded away from zero) if the separation signal size Δ^2 is below the cutoff value $\overline{\Delta}^2$. ■

Proof of Lemma V.1: To prove the lower bound, we follow the same setup as in the lower bound proof in [8] by placing a $N(0, \kappa_n^2 I_p)$ prior on $\boldsymbol{\mu}$ and an independent Rademacher prior on η. Note that algorithm that maximizing the probability of reconstructing labels correctly is the maximum a posterior (MAP) estimator $\widetilde{\eta} = \operatorname{argmax}_\eta p(\eta \,|\, \mathbf{X})$. Since the prior label assignment is uniform, MAP is in particular equivalent to maximum (integrated) likelihood estimator (MLE) after integrating out $\boldsymbol{\mu}$, i.e., $\widetilde{\eta} = \operatorname{argmax}_\eta L(\eta \,|\, \mathbf{X})$ where $L(\eta \,|\, \mathbf{X}) = p(\mathbf{X} \,|\, \eta)$ is viewed as a function of η. Specifically, the maximum (integrated) likelihood function can be computed as follows

$$\begin{aligned} &L(\eta \,|\, \mathbf{X}) \\ &= \int_{\mathbb{R}^p} \prod_{i=1}^n p(\mathbf{X}_i \,|\, \boldsymbol{\mu}, \eta_i)\, p(\boldsymbol{\mu})\, \mathrm{d}\boldsymbol{\mu} \\ &\propto \int_{\mathbb{R}^p} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \|\mathbf{X}_i - \eta_i \boldsymbol{\mu}\|^2 - \frac{1}{2\kappa_n^2} \|\boldsymbol{\mu}\|^2 \Big\}\, \mathrm{d}\boldsymbol{\mu} \\ &\propto \exp\Big\{ \frac{1}{2} \Big(\frac{n}{\sigma^2} + \frac{1}{\kappa_n^2}\Big)^{-1} \Big\| \frac{1}{\sigma^2} \sum_{i=1}^n \eta_i \mathbf{X}_i \Big\|^2 \Big\}. \end{aligned}$$

We can see from the last expression that the MLE fails if there exists some $i \in [n]$ such that $\big\|\eta_i \mathbf{X}_i + \sum_{j \neq i}^n \eta_j \mathbf{X}_j\big\|^2 < \big\| - \eta_i \mathbf{X}_i + \sum_{j \neq i}^n \eta_j \mathbf{X}_j\big\|^2$, or equivalently, $\big\langle \eta_i \mathbf{X}_i,\, \sum_{j \neq i}^n \eta_j \mathbf{X}_j \big\rangle < 0$. Therefore,

$$\begin{aligned} &\inf_{\widehat{\eta}} \mathbb{P}(\widehat{\eta} \neq \eta) = \mathbb{P}(\widetilde{\eta} \neq \eta) \\ &\geqslant \mathbb{P}\Big(\exists i \in [n],\ \text{such that } \big\langle \eta_i \mathbf{X}_i, \sum_{j \neq i}^n \eta_j \mathbf{X}_j \big\rangle < 0\Big). \end{aligned} \tag{33}$$

Without loss of generality, we assume $\sigma = 1$. Let $\overline{\boldsymbol{\varepsilon}}_n = n^{-1} \sum_{i=1}^n \boldsymbol{\varepsilon}_i$ be the sample average of the noise. Since $(\eta_1 \boldsymbol{\varepsilon}_1, \ldots, \eta_n \boldsymbol{\varepsilon}_n)$ has the same joint distribution as $(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$, we can write

$$\begin{aligned} &\frac{1}{n-1} \big\langle \eta_i \mathbf{X}_i, \sum_{j \neq i}^n \eta_j \mathbf{X}_j \big\rangle = \big\langle \boldsymbol{\mu} + \eta_i \boldsymbol{\varepsilon}_i,\, \mu + \frac{1}{n-1} \sum_{j \neq i} \eta_i \boldsymbol{\varepsilon}_i \big\rangle \\ &\overset{d}{=} \Big\langle \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i,\, \boldsymbol{\mu} + \frac{n}{n-1} \bar{\boldsymbol{\varepsilon}}_n - \frac{1}{n-1} \boldsymbol{\varepsilon}_i \Big\rangle \\ &= \underbrace{\Big\langle \boldsymbol{\varepsilon}_i - \overline{\boldsymbol{\varepsilon}}_n,\, \boldsymbol{\mu} + \frac{n}{n-1} \overline{\boldsymbol{\varepsilon}}_n \Big\rangle + \|\boldsymbol{\mu}\|^2}_{=: R_{i,1}} \\ &\quad + \underbrace{\frac{n}{n-1} \|\overline{\boldsymbol{\varepsilon}}_n\|^2 - \frac{1}{n-1} \|\boldsymbol{\varepsilon}_i\|^2}_{=: R_{i,2}} \end{aligned}$$

$$+\underbrace{\frac{2n-1}{n-1}\langle\boldsymbol{\mu},\overline{\boldsymbol{\varepsilon}}_n\rangle-\frac{1}{n-1}\langle\boldsymbol{\mu},\boldsymbol{\varepsilon}_i\rangle}_{=:R_{i,3}}.$$

Bound $R_{i,3}$**.** Let $\beta_n>0$ and

$$\begin{aligned}\mathcal{B}_1=\big\{&\sqrt{n}|\langle\boldsymbol{\mu},\overline{\boldsymbol{\varepsilon}}_n\rangle|\leqslant\sqrt{n-1}\beta_n\|\boldsymbol{\mu}\|^2,\\&\max_{i\in[n]}|\langle\boldsymbol{\mu},\boldsymbol{\varepsilon}_i\rangle|\leqslant\sqrt{n-1}\beta_n\|\boldsymbol{\mu}\|^2\big\},\\\widetilde{\mathcal{B}}_1=\big\{&\sqrt{n}|\langle\boldsymbol{\mu},\overline{\boldsymbol{\varepsilon}}_n\rangle|\leqslant\sqrt{n-1}\beta_n\|\boldsymbol{\mu}\|,\\&\max_{i\in[n]}|\langle\boldsymbol{\mu},\boldsymbol{\varepsilon}_i\rangle|\leqslant\sqrt{n-1}\beta_n\|\boldsymbol{\mu}\|\big\}.\end{aligned}$$

By the standard tail inequality of the Gaussian random variable and union bound, we have $\mathbb{P}(\mathcal{B}_1^c)\leqslant\min\{1,n\exp(-cn\beta_n^2\|\boldsymbol{\mu}\|^2)\}$ and $\mathbb{P}(\widetilde{\mathcal{B}}_1^c)\leqslant\min\{1,n\exp(-cn\beta_n^2)\}$ for some universal constant $c>0$. In addition, we have $\max_{i\in[n]}|R_{i,3}|\leqslant3\beta_n\|\boldsymbol{\mu}\|^2$ on the event $\mathcal{B}_1$ and $\max_{i\in[n]}|R_{i,3}|\leqslant3\beta_n\|\boldsymbol{\mu}\|$ on the event $\widetilde{\mathcal{B}}_1$.

Bound $R_{i,2}$**.** By tail inequalities of the chi-square random variable in Lemma VIII.1, we have for any $t>0$ and $i\in[n]$,

$$\begin{aligned}\mathbb{P}\big(\big|\|\boldsymbol{\varepsilon}_i\|^2-p\big|\geqslant2\sqrt{pt}+2t\big)&\leqslant2e^{-t},\\\mathbb{P}\big(\big|n\|\overline{\boldsymbol{\varepsilon}}\|^2-p\big|\geqslant2\sqrt{pt}+2t\big)&\leqslant2e^{-t}.\end{aligned}$$

Thus we have $\mathbb{P}(\mathcal{B}_2^c)\leqslant4n^{-1}$, where

$$\begin{aligned}\mathcal{B}_2=\big\{&\big|n\|\overline{\boldsymbol{\varepsilon}}_n\|^2-p\big|\leqslant2\sqrt{p\log n}+2\log n,\\&\max_{i\in[n]}\big|\|\boldsymbol{\varepsilon}_i\|^2-p\big|\leqslant2\sqrt{2p\log n}+4\log n\big\}.\end{aligned}$$

On the event $\mathcal{B}_2$, we have $\max_{i\in[n]}|R_{i,2}|\leqslant6(\sqrt{p\log n}+\log n)/(n-1)$.

Analyze $R_{i,1}$**.** From elementary calculations, we have that the conditional joint distribution of $U_i:=\big\langle\boldsymbol{\varepsilon}_i-\overline{\boldsymbol{\varepsilon}}_n,\,\boldsymbol{\mu}+\frac{n}{n-1}\overline{\boldsymbol{\varepsilon}}_n\big\rangle$, $i=1,\dots,n$, given $\overline{\boldsymbol{\varepsilon}}_n$ is

$$\begin{pmatrix}U_1\\U_2\\\vdots\\U_n\end{pmatrix}\Bigg|\,\overline{\boldsymbol{\varepsilon}}_n\sim N\left(\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix},\Big\|\boldsymbol{\mu}+\frac{n\overline{\boldsymbol{\varepsilon}}_n}{n-1}\Big\|^2\begin{pmatrix}1-n^{-1}&-n^{-1}&\cdots&-n^{-1}\\-n^{-1}&1-n^{-1}&\cdots&-n^{-1}\\\vdots&\vdots&\ddots&\vdots\\-n^{-1}&-n^{-1}&\cdots&1-n^{-1}\end{pmatrix}\right).$$

Conditioning on $\overline{\boldsymbol{\varepsilon}}_n$, let $\{Z_i\}_{i=1}^n$ be i.i.d. $N\big(0,(1-n^{-1})\big\|\boldsymbol{\mu}+\frac{n}{n-1}\overline{\boldsymbol{\varepsilon}}_n\big\|^2\big)$ random variables. Since $\mathbb{E}(U_i^2\,|\,\overline{\boldsymbol{\varepsilon}}_n)=\mathbb{E}(Z_i^2\,|\,\overline{\boldsymbol{\varepsilon}}_n)$ and $\mathbb{E}((U_i-U_j)^2\,|\,\overline{\boldsymbol{\varepsilon}}_n)\geqslant\mathbb{E}((Z_i-Z_j)^2\,|\,\overline{\boldsymbol{\varepsilon}}_n)$ for $i,j\in[n]$, by Slepian's inequality (cf. Theorem 7.2.9 in [58]) we have

$$\begin{aligned}\mathbb{P}\Big(\max_{i\in[n]}U_i>t\,\Big|\,\overline{\boldsymbol{\varepsilon}}_n\Big)\geqslant&\mathbb{P}\Big(\max_{i\in[n]}Z_i>t\,\Big|\,\overline{\boldsymbol{\varepsilon}}_n\Big)\\=&1-\Big(1-\mathbb{P}\big(Z_1>t\,\big|\,\overline{\boldsymbol{\varepsilon}}_n\big)\Big)^n,\quad t\in\mathbb{R}.\end{aligned}$$

Combining the previous three terms with (33), we obtain (34), shown at the bottom of the next page. For $\gamma_n>0$, we define

$$\mathcal{B}_3=\big\{\big|\|\boldsymbol{\mu}\|^2-\Delta^2/4\big|\leqslant\Delta^2\gamma_n/4\big\}.$$

With the prior distribution $\boldsymbol{\mu}\sim N(0,4^{-1}\kappa_n^2I_p)$ where $\kappa_n^2=\frac{\Delta^2}{4p(1-\nu_n)}$ and $\nu_n=\sqrt{\frac{n\Delta^2}{4p\log^2n}}$, it follows from the proof of Theorem 5 in [8] (cf. equation (28)) that $\mathbb{P}(\mathcal{B}_3^c)\leqslant2\exp(-p\gamma_n^2/32)$, provided $4\nu_n\leqslant\gamma_n\leqslant1$. Moreover, using the lower tail bound of the chi-square random variable in Lemma VIII.1, we have $\mathbb{P}(\mathcal{B}_4^c)\leqslant\exp(-p\theta_n^2/4)$, where

$$\mathcal{B}_4=\big\{\|\overline{\boldsymbol{\varepsilon}}_n\|^2\geqslant\frac{p}{n}(1-\theta_n)\big\}.$$

To analyze the right-hand side of (34), we first consider the higher dimensional case where $p\geqslant\log^2n$. In such regime, we divide further into three cases depending on the separation signal size as following.

Medium signal size case: $\frac{2\log^{3/2}n}{\sqrt{n}}<\Delta<2\sqrt{\frac{p\log n}{n}}$**.** Since $-(U_1,\dots,U_n)$ has the same joint distribution of $(U_1,\dots,U_n)$ given $\overline{\boldsymbol{\varepsilon}}_n$, we can bound on $\mathcal{B}_1\cap\mathcal{B}_2$,

$$\begin{aligned}&\mathbb{P}\Big(-U_i\leqslant(1+3\beta_n)\|\boldsymbol{\mu}\|^2+\frac{6(\sqrt{p\log n}+\log n)}{n-1},\ \forall i\in[n]\Big)\\=&\mathbb{P}\Big(\max_{i\in[n]}U_i\leqslant(1+3\beta_n)\|\boldsymbol{\mu}\|^2+\frac{6(\sqrt{p\log n}+\log n)}{n-1}\Big)\\\leqslant&\mathbb{E}\Big[1-\mathbb{P}\Big(Z_1>(1+3\beta_n)\|\boldsymbol{\mu}\|^2+\frac{6(\sqrt{p\log n}+\log n)}{n-1}\Big|\,\overline{\boldsymbol{\varepsilon}}_n\Big)\Big]^n.\end{aligned}$$

Let $Z\sim N(0,1)$ be the standard Gaussian random variable. Thus, on the event $\bigcap_{i=1}^4\mathcal{B}_i$, we have

$$\begin{aligned}&\mathbb{P}\Big(Z_1>(1+3\beta_n)\|\boldsymbol{\mu}\|^2+\frac{6(\sqrt{p\log n}+\log n)}{n-1}\Big|\,\overline{\boldsymbol{\varepsilon}}_n\Big)\\=&\mathbb{P}\left(Z>\frac{(1+3\beta_n)\|\boldsymbol{\mu}\|^2+\frac{6(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{1-\frac{1}{n}}\sqrt{\|\boldsymbol{\mu}\|^2+\frac{n^2}{(n-1)^2}\|\overline{\boldsymbol{\varepsilon}}_n\|^2+\frac{2n}{n-1}\langle\boldsymbol{\mu},\overline{\boldsymbol{\varepsilon}}_n\rangle}}\Big|\,\overline{\boldsymbol{\varepsilon}}_n\right)\\\geqslant&\mathbb{P}\left(Z>\frac{(1+3\beta_n)(1+\gamma_n)\Delta^2+\frac{24(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{4(1-\frac{1}{n}-2\beta_n)(1-\gamma_n)\Delta^2+\frac{16p}{n-1}(1-\theta_n)}}\right)\\=&\Phi^c(V_n),\end{aligned}$$

$$\begin{aligned}&\inf_{\{\hat{h}_{ik}\}}\sup_{(H,\boldsymbol{\mu})\in\Theta(n,K,\Delta)}\mathbb{P}_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik}\neq h_{ik},\,i\in[n],\,k\in[K]\big)\\\geqslant&\inf_{\{\hat{h}_{ik}\}}\sup_{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|\geqslant\Delta,\forall(k,l)\in[K]^2,k\neq l}\mathbb{E}_{\pi^H}\mathbb{P}_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik}\neq h_{ik},\,i\in[n],\,k\in[K]\big)-n^{-1}.\end{aligned}\tag{29}$$

$$\inf_{\{\hat{h}_{ik},i\in G_1\cup G_2,k=1,2\}}\sup_{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|\geqslant\Delta}\mathbb{E}_{\tilde{\pi}^{12}}P_{(H,\boldsymbol{\mu})}\big(\hat{h}_{ik}\neq h_{ik},\,i\in G_1\cup G_2,\,k\in[2]\big)\geqslant1-cK/n.\tag{30}$$

where

$$V_n = \frac{(1+3\beta_n)(1+\gamma_n)\Delta^2 + \frac{24(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{4(1-\frac{1}{n}-2\beta_n)(1-\gamma_n)\Delta^2 + \frac{16p}{n-1}(1-\theta_n)}}.$$

and $\Phi^c(t) = \mathbb{P}(Z \geqslant t)$. Combining all pieces together, we obtain that

$$\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - [1-(1-r_n)\Phi^c(V_n)]^n - r_n \tag{35}$$

provided $4\nu_n \leqslant \gamma_n \leqslant 1$, where

$$r_n = \min\{1, n\exp(-cn\beta_n^2(1-\gamma_n^2)\Delta^2)\} + 4n^{-1} + 2\exp(-p\gamma_n^2/32) + \exp(-p\theta_n^2/4).$$

Note that $\Delta < 2\sqrt{\frac{p\log n}{n}}$, implying $\nu_n^2 \leqslant \frac{1}{\log n}$. We choose

$$\beta_n^2 = \frac{1}{\log n}, \quad \gamma_n^2 = \frac{16}{\log n}, \quad \theta_n^2 = \frac{1}{\log n}.$$

Since $p \geqslant \log^2 n$ and $\Delta > \frac{2\log^{3/2} n}{\sqrt{n}}$, we have

$$\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - [1-(1-cn^{-1})\Phi^c(V_n)]^n - cn^{-1}.$$

By using the fact that $\log\Phi^c(t) \sim -t^2/2$ as $t \to \infty$, we can conclude that as long as $V_n \leqslant \sqrt{2(1-\delta)\log n}$ for any $\delta \in (0,1)$, then for $n \geqslant 2c$, we have

$$\begin{aligned}\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant& 1 - \Big(1 - \frac{1-cn^{-1}}{n^{1-\delta}}\Big)^n - cn^{-1} \\ \geqslant& 1 - e^{-n^{\delta}/2} - cn^{-1} \geqslant 1 - c'n^{-1},\end{aligned}$$

where c' is a constant depending on δ. The condition $V_n \leqslant \sqrt{2(1-\delta)\log n}$ is implied by $\Delta^2 \leqslant (1-\alpha)\overline{\Delta}^2$ for some $\delta := \delta(\alpha)$ and by inverting the function $x \mapsto x/\sqrt{4x+16p/n}$ for $x > 0$.

Low signal size case: $\Delta \leqslant \frac{2\log^{3/2} n}{\sqrt{n}}$. The argument is similar to the medium signal size case, so we only sketch the proof. On the event $\widetilde{\mathcal{B}}_1 \bigcap \mathcal{B}_2 \bigcap \mathcal{B}_3 \bigcap \mathcal{B}_4$, we have

$$\begin{aligned}&\mathbb{P}\Big(\max_{i\in[n]} U_i \leqslant \|\boldsymbol{\mu}\|^2 + 3\beta_n\|\boldsymbol{\mu}\| + \frac{6(\sqrt{p\log n}+\log n)}{n-1}\Big) \\ &\leqslant \mathbb{E}\Big[1 - \mathbb{P}\Big(Z_1 > \|\boldsymbol{\mu}\|^2 + 3\beta_n\|\boldsymbol{\mu}\| + \frac{6(\sqrt{p\log n}+\log n)}{n-1} \big| \overline{\boldsymbol{\varepsilon}}_n\Big)\Big]^n\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\Big(Z_1 > \|\boldsymbol{\mu}\|^2 + 3\beta_n\|\boldsymbol{\mu}\| + \frac{6(\sqrt{p\log n}+\log n)}{n-1} \big| \overline{\boldsymbol{\varepsilon}}_n\Big) \\ &\geqslant \mathbb{P}\left(Z > \frac{\|\boldsymbol{\mu}\|^2 + 3\beta_n\|\boldsymbol{\mu}\| + \frac{6(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{1-\frac{1}{n}}\sqrt{\|\boldsymbol{\mu}\|^2 + \frac{n^2}{(n-1)^2}\|\overline{\boldsymbol{\varepsilon}}_n\|^2 - 2\beta_n\|\boldsymbol{\mu}\|}} \big| \overline{\boldsymbol{\varepsilon}}_n\right) \\ &\geqslant \mathbb{P}\left(Z > \frac{(1+\gamma_n)\Delta^2 + 6\beta_n\sqrt{1+\gamma_n}\Delta + \frac{24(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{4(1-\frac{1}{n})(1-\gamma_n)\Delta^2 + \frac{16p}{n-1}(1-\theta_n) - 16\beta_n\sqrt{1-\gamma_n}\Delta}}\right) \\ &= \Phi^c(V_n),\end{aligned}$$

where

$$V_n = \frac{(1+\gamma_n)\Delta^2 + 6\beta_n\sqrt{1+\gamma_n}\Delta + \frac{24(\sqrt{p\log n}+\log n)}{n-1}}{\sqrt{4(1-\frac{1}{n})(1-\gamma_n)\Delta^2 + \frac{16p}{n-1}(1-\theta_n) - 16\beta_n\sqrt{1-\gamma_n}\Delta}}.$$

Combining all pieces together, we obtain that

$$\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - [1-(1-r_n)\Phi^c(V_n)]^n - r_n,$$

provided $4\nu_n \leqslant \gamma_n \leqslant 1$, where

$$r_n = \min\{1, n\exp(-cn\beta_n^2)\} + 4n^{-1} + 2\exp(-p\gamma_n^2/32) + \exp(-p\theta_n^2/4).$$

Now we choose

$$\beta_n^2 = \frac{\log^2 n}{n}, \quad \gamma_n^2 = \frac{16}{\log n}, \quad \theta_n^2 = \frac{1}{\log n}.$$

If $\frac{2\log^2 n}{n} < \Delta \leqslant \frac{2\log^{3/2} n}{\sqrt{n}}$, then $\beta_n\Delta \leqslant p/n$ (recall $p \geqslant \log^2 n$) and there exists a sequence $\xi_n \to 0$ as $n \to \infty$ such that

$$V_n \leqslant (1+\xi_n)(\Delta/2 + 3\beta_n + \xi_n) = o(1),$$

which implies that $\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - cn^{-1}$. If $\Delta \leqslant \frac{2\log^2 n}{n}$, then

$$V_n \leqslant (1+\xi_n)\frac{3\Delta\beta_n}{2\sqrt{p/n}} = o(1)$$

and $\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - cn^{-1}$.

High signal size case: $2\sqrt{\frac{p\log n}{n}} \leqslant \Delta \leqslant \sqrt{1-\alpha}\ \overline{\Delta}$. Note that in this regime, we have $p/n = o(\Delta^2)$ and $p = O(n)$. Then the sharp threshold $\overline{\Delta}^2 = 8(1+o(1))\log n$, which is asymptotically independent of p. Thus we place an (essentially one-dimensional) point mass prior on $\boldsymbol{\mu}$ at $(\Delta/2, 0, \ldots, 0)^T \in \mathbb{R}^p$. A similar calculation yields

$$\begin{aligned}&L(\eta \,|\, \mathbf{X}) \propto \exp\Big\{\frac{1}{\sigma^2}\Big\langle \boldsymbol{\mu}, \sum_{i=1}^n \eta_i \mathbf{X}_i\Big\rangle\Big\}, \quad \text{and} \\ &\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant \mathbb{P}\big(\exists i \in [n], \text{ such that } \langle \boldsymbol{\mu}, \eta_i\mathbf{X}_i\rangle < 0\big).\end{aligned}$$

Since $\{\langle \boldsymbol{\mu}, \eta_i\mathbf{X}_i\rangle\}_{i=1}^n$ are i.i.d. random variables with

$$\begin{aligned}\mathbb{P}\big(\langle \boldsymbol{\mu}, \eta_i\mathbf{X}_i\rangle \geqslant 0\big) =& \mathbb{P}\big(\|\boldsymbol{\mu}\|^2 + \langle \boldsymbol{\mu}, \eta_i\boldsymbol{\varepsilon}_i\rangle \geqslant 0\big) \\ =& \mathbb{P}(Z \geqslant -\|\boldsymbol{\mu}\|) = 1 - \Phi^c(\Delta/2),\end{aligned}$$

we have

$$\begin{aligned}\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant& 1 - \big(1 - \Phi^c(\Delta/2)\big)^n \\ \geqslant& 1 - \Big(1 - \frac{1}{n^{1-\delta}}\Big)^n \geqslant 1 - e^{-n^{\delta}} \geqslant 1 - cn^{-1},\end{aligned}$$

when $\Delta^2 \leqslant 4(2-\delta)\log n \leqslant (1-\alpha)\overline{\Delta}^2$ for some δ depending only on α. Here the constant c depending only on δ (and thus only on α).

$$\begin{aligned}\mathbb{P}(\tilde{\eta} \neq \eta) \geqslant& 1 - \mathbb{P}\Big(R_{i,1} + R_{i,2} + R_{i,3} \geqslant 0,\ \forall i \in [n]\Big) = 1 - \mathbb{P}\Big(-U_i \leqslant \|\boldsymbol{\mu}\|^2 + R_{i,2} + R_{i,3},\ \forall i \in [n]\Big) \\ \geqslant& \begin{cases} 1 - \mathbb{P}(-U_i \leqslant (1+3\beta_n)\|\boldsymbol{\mu}\|^2 + \frac{6(\sqrt{p\log n}+\log n)}{n-1}, \forall i \in [n]) & \text{on } \mathcal{B}_1 \cap \mathcal{B}_2 \\ 1 - \mathbb{P}(-U_i \leqslant \|\boldsymbol{\mu}\|^2 + 3\beta_n\|\boldsymbol{\mu}\| + \frac{6(\sqrt{p\log n}+\log n)}{n-1}, \forall i \in [n]) & \text{on } \widetilde{\mathcal{B}}_1 \cap \mathcal{B}_2 \end{cases}.\end{aligned} \tag{34}$$

Finally, we deal with the lower dimensional case where $p < \log^2 n$. In such regime, we also have $\overline{\Delta}^2 = 8(1 + o(1)) \log n$. Following the same argument as in the **high signal size case** under $p \geqslant \log^2 n$, we conclude that $\mathbb{P}(\widetilde{\eta} \neq \eta) \geqslant 1 - cn^{-1}$. ■

VI. Discussions

In this paper, we characterized the information-theoretic sharp threshold for exact recovery of Gaussian mixture models. There are still some interesting open questions, which we list below.

General noise covariance matrix. The SDP relaxation in (11) does not require to know the noise variance σ^2 only in the spherical Gaussian case, i.e., the noise $\boldsymbol{\varepsilon}_i$ has i.i.d. $N(0, \sigma^2 I_p)$ distribution. Consider the general covariance matrix case $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma)$ when Σ is not necessarily spherical. If Σ is known, then we can first apply the transform $\Sigma^{-1/2} X_i$ to make the noise spherical and the sharp threshold in (8) holds and reads in terms of the minimal Mahalanobis distance

$$\widetilde{\Delta}^2 = \min_{1 \leqslant j < k \leqslant K} d_\Sigma^2(\mu_j, \mu_k) = 4 \left(1 + \sqrt{1 + \frac{Kp}{n \log n}} \right) \log n,$$

where $d_\Sigma^2(\mu_j, \mu_k) = (\mu_j - \mu_k)^T \Sigma^{-1} (\mu_j - \mu_k)$. If Σ is unknown, [59] showed that in the $K = 2$ case the misclassification probability of the Bayes classifier decays exponentially fast in the Mahalanobis distance $d_\Sigma^2(\mu_1, \mu_2) = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$ rather than $\frac{\Delta^2}{\sigma^2} = \frac{\Delta^2}{\|\Sigma\|_{op}}$. Thus we conjecture that:

> *there is a sharp threshold for exact recovery under the general unknown covariance matrix given by $\widetilde{\Delta}^2$ above.*

Average-case algorithmic hardness in multiple clusters. Both our upper and lower bounds for exact recovery in Corollary II.2 and Theorem II.3 require the number of clusters $K = O(\log n)$. We argue that this condition is likely to be necessary for achieving the sharp threshold of exact recovery. Consider the balanced spherical Gaussian mixture model with common noise variance and multiple communities for $K \geqslant 3$. It is shown in [60] that: (i) detection and partial (i.e., correlated) recovery are information-theoretically possible if $\rho > 2\sqrt{\frac{pK \log K}{n}} + 2 \log K$; (ii) detection and partial recovery are impossible if $\rho < \sqrt{\frac{2p(K-1)\log(K-1)}{n}}$, where ρ is the squared signal-to-noise ratio in the Gaussian mixture model (an equivalent quantity of Δ^2/σ^2 in our notation). In contrast, it is also known from [60], [61] that spectral methods have correlated recovery with the true community labels if and only if $\rho > \sqrt{\frac{p}{n}}(K-1)$. The phase transition of spectral methods is a direct consequence of the BBP phase transition in the random matrix theory [62], [63]. Thus for fixed K, there is no gap (modulo constants) between computation and information theoretic thresholds. In addition, a sufficient condition for partial recovery of the same SDP as in our paper is given by $\frac{\Delta^2}{\sigma^2} \gtrsim (1 + \sqrt{p/n})K$ in [14]. Based on evidence from statistical physics, it is conjectured by [64] (and remains as an open problem) that the computational threshold coincides with the spectral methods for partial recovery for large K, thus suggesting there is a computationally hard regime where no polynomial time algorithm can attain the information-theoretic threshold when $K \to \infty$.

Now turning into exact recovery. Recall that our result shows that the information-theoretical threshold is

$$\frac{\overline{\Delta}^2}{\sigma^2} = 4 \left(\log n + \sqrt{\log^2 n + \frac{Kp \log n}{n}} \right),$$

which is achieved by an SDP when $K \lesssim \log(n)/\log\log(n)$. Thus, in such growth region of K, there is no computational hardness for exact recovery, which is a similar scenario in the partial recovery case when $K = O(1)$. Note that the threshold $\overline{\Delta}^2/\sigma^2$ is larger (modulo constants) than the partial recovery sufficient condition for the spectral methods and the SDP, which in turn is strictly larger than its necessary condition (i.e., information-theoretic threshold) as $K \to \infty$. Hence, we propose the following conjecture:

> *for $K \gg \log n$, there is no polynomial time algorithm can achieve the average-case exact recovery information-theoretic threshold.*

If this conjecture is true, then our current regime $K \lesssim \log(n)/\log\log(n)$ where the SDP achieves the information-theoretic limit is sharp, i.e., $K \asymp \log(n)$ would be an algorithmic hardness for exact recovery. The conjecture also implies that transition of hardness of clustering Gaussian mixture models for partial recovery and exact recovery is from $O(1)$ and $O(\log n)$, respectively.

Unbalanced communities. Corollary II.2 and Theorem II.3 together imply that in the equal cluster size case when $n_1 = n_2 = \cdots = n_K = \frac{n}{K}$, the SDP relaxation (11) for the K-means is minimax-optimal in the sense that sharp phase transition of the probability of wrong recovery from zero to one occurs at the critical threshold given by the $\overline{\Delta}^2$ in (8). It remains an interesting open question whether the separation gap $\overline{\Delta}$ is sharp when cluster sizes are unbalanced.

VII. Proof of Key Lemmas

A. Proof of Lemma IV.1

Without loss of generality, we may assume $\sigma = 1$. Denote $\boldsymbol{\theta} = \boldsymbol{\mu}_k - \boldsymbol{\mu}_l$ and define the event $\mathcal{A} = \bigcap_{k,l,i} \mathcal{A}_{kl}^{(i)}$, where

$$\mathcal{A}_{kl}^{(i)} = \Big\{ \|\mathbf{X}_i - \overline{\mathbf{X}}_l\|^2 - \|\mathbf{X}_i - \overline{\mathbf{X}}_k\|^2 \geqslant \frac{n_k + n_l}{n_k n_l} p + \beta \|\boldsymbol{\theta}\|^2 - r_{kl} \Big\},$$

with the index (k, l, i) ranging over all distinct pairs $(k, l) \in [K]^2$ and all $i \in G_k^*$ and

$$r_{kl} = 2\sqrt{\frac{2\log(nK)}{n_l}} \|\boldsymbol{\theta}\| + 2\frac{n_k + n_l}{n_k n_l} \sqrt{2p \log(nK)} + \frac{4}{n_k} \log(nK).$$

Recall that $\mathbf{X}_i = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_i$ for each $i \in G_k^*$ and $k \in [K]$. We can write

$$\begin{aligned} \|\mathbf{X}_i - \overline{\mathbf{X}}_l\|^2 - \|\mathbf{X}_i - \overline{\mathbf{X}}_k\|^2 &= \|\boldsymbol{\theta} + \boldsymbol{\varepsilon}_i - \overline{\boldsymbol{\varepsilon}}_l\|^2 - \|\boldsymbol{\varepsilon}_i - \overline{\boldsymbol{\varepsilon}}_k\|^2 \\ &= \langle \boldsymbol{\theta} - \overline{\boldsymbol{\varepsilon}}_l + \overline{\boldsymbol{\varepsilon}}_k, \boldsymbol{\theta} - \overline{\boldsymbol{\varepsilon}}_l + 2\boldsymbol{\varepsilon}_i - \overline{\boldsymbol{\varepsilon}}_k \rangle \end{aligned}$$

$$= \|\boldsymbol{\theta}\|^2 + \|\overline{\boldsymbol{\varepsilon}}_l\|^2 + \frac{1}{n_k}\left(2 - \frac{1}{n_k}\right)\|\boldsymbol{\varepsilon}_i\|^2 - \left(\frac{n_k-1}{n_k}\right)^2 \|\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\|^2 - 2\langle\boldsymbol{\theta}, \overline{\boldsymbol{\varepsilon}}_l\rangle + 2\left\langle \boldsymbol{\theta} - \overline{\boldsymbol{\varepsilon}}_l + \left(\frac{n_k-1}{n_k}\right)^2 \overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}, \boldsymbol{\varepsilon}_i \right\rangle,$$

where $\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}} = (n_k-1)^{-1}\sum_{j\in G_k^*\setminus\{i\}} \boldsymbol{\varepsilon}_j$. Set $\zeta_n = 2\log(nK)$ and define

$$\mathcal{B}_{kl}^{(i)} = \Big\{ \|\overline{\boldsymbol{\varepsilon}}_l\|^2 \geqslant n_l^{-1}(p - 2\sqrt{p\zeta_n}),\ \|\boldsymbol{\varepsilon}_i\|^2 \geqslant p - 2\sqrt{p\zeta_n},$$
$$\|\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\|^2 \leqslant (n_k-1)^{-1}(p + 2\sqrt{p\zeta_n} + 2\zeta_n),$$
$$\langle\boldsymbol{\theta}, \overline{\boldsymbol{\varepsilon}}_l\rangle \leqslant \sqrt{2n_l^{-1}\zeta_n}\|\boldsymbol{\theta}\|,$$
$$\|\overline{\boldsymbol{\varepsilon}}_l - (1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\|^2 \leqslant (n_l^{-1}+n_k^{-1})(p+2\sqrt{p\zeta_n}+2\zeta_n), \tag{36}$$
$$\langle\boldsymbol{\theta}, (1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}} - \overline{\boldsymbol{\varepsilon}}_l\rangle \leqslant \sqrt{2(n_l^{-1}+n_k^{-1})\zeta_n}\|\boldsymbol{\theta}\| \Big\}. \tag{37}$$

Note that $\boldsymbol{\varepsilon}_i, \overline{\boldsymbol{\varepsilon}}_l$, and $\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}$ are mutually independent. Thus conditional on $\overline{\boldsymbol{\varepsilon}}_l$ and $\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}$, we have

$$\langle \boldsymbol{\theta} - \overline{\boldsymbol{\varepsilon}}_l + (1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}, -\boldsymbol{\varepsilon}_i\rangle \sim N\Big(0, \Big\|\boldsymbol{\theta} - \overline{\boldsymbol{\varepsilon}}_l + (1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\Big\|^2\Big).$$

Then, on the event where (36) and (37) hold, we can bound U^* in (38), shown at the bottom of the next page, where $\Phi^c(t)$ denotes the tail probability $\mathbb{P}(Z \geqslant t)$ for a standard normal random variable Z. Note that $n_l^{-1} + n_k^{-1} \leqslant 2m^{-1}$. Under the separation condition (24) on the Gaussian centers, we have $\|\boldsymbol{\theta}\|^2 \geqslant 8\log n$ and

$$2\sqrt{2(n_l^{-1}+n_k^{-1})\zeta_n}\|\boldsymbol{\theta}\| \leqslant \frac{2}{\sqrt{m}}\|\boldsymbol{\theta}\|^2.$$

Thus, on events (36) and (37), we have

$$U^* \leqslant \Phi^c\left(\frac{(1-\beta)\|\boldsymbol{\theta}\|^2}{2\sqrt{(1+\frac{2}{\sqrt{m}})\|\boldsymbol{\theta}\|^2 + (n_l^{-1}+n_k^{-1})(p+2\sqrt{p\zeta_n}+2\zeta_n)}}\right).$$

Now under the separation condition (24) and noting that $(1+\delta)(1+\frac{2}{\sqrt{m}}) \leqslant 1+2\delta$, we see that

$$\frac{(1-\beta)^2}{8(1+\delta)\log n}\|\boldsymbol{\theta}\|^4 - \left(1 + \frac{2}{\sqrt{m}}\right)\|\boldsymbol{\theta}\|^2 - r_1 \geqslant 0,$$

where

$$r_1 = p(n_l^{-1}+n_k^{-1}) + 2(\sqrt{p\zeta_n} + \zeta_n)(n_l^{-1}+n_k^{-1}).$$

Hence we get

$$U^* \leqslant \Phi^c(\sqrt{2(1+\delta)\log n}) \leqslant n^{-(1+\delta)},$$

where the second inequality follows from the standard Gaussian tail bound $\Phi^c(x) \leqslant e^{-x^2/2}$ for $x \geqslant 0$. In addition, applying the probability tail bounds for χ^2 distributions in Lemma VIII.1, we have $\mathbb{P}(\mathcal{B}_{kl}^{(i)^c}) \leqslant 6/(n^2K^2)$. Now putting pieces together, we have

$$\mathbb{P}(\mathcal{A}^c) \leqslant \sum_{1\leqslant k\neq l\leqslant K}\sum_{i\in G_k^*} \mathbb{P}(\mathcal{A}_{kl}^{(i)^c} \cap \mathcal{B}_{kl}^{(i)}) + \mathbb{P}(\mathcal{B}_{kl}^{(i)^c}) \leqslant \sum_{1\leqslant k\neq l\leqslant K}\sum_{i\in G_k^*} \mathbb{E}\left[U^*\mathbf{1}((36),\ (37)\text{ hold})\right] + \frac{6}{n} \leqslant \frac{K^2}{n^\delta} + \frac{8}{n}.$$

B. Proof of Lemma IV.2

Without loss of generality, we may assume $\sigma = 1$. Recall that the column sums and row sums of matrix $B_{G_k^*G_l^*}$ are denoted by $\mathbf{c}^{(k,l)} = \big(c_j^{(k,l)} : j \in G_l^*\big)$ and $\mathbf{r}^{(k,l)} = \big(r_i^{(k,l)} : i \in G_k^*\big)$, respectively. In addition, $t^{(k,l)} = \sum_{j\in G_l^*} c_j^{(k,l)} = \sum_{i\in G_k^*} r_i^{(k,l)}$ is the total sum, and the construction of $B^\sharp$ in (22) can be written as $[B^\sharp_{G_l^*G_k^*}]_{ij} = r_i^{(k,l)}c_j^{(k,l)}/t^{(k,l)}$, for any distinct pair $(k,l) \in [K]^2$. Under this notation, for each $\mathbf{v} \in \Gamma_K$, we may write

$$T(\mathbf{v}) = \sum_{k=1}^K \sum_{l\neq k}\sum_{i\in G_k^*}\sum_{j\in G_l^*} \frac{r_i^{(k,l)}c_j^{(k,l)}}{t^{(k,l)}} v_i v_j = \sum_{k=1}^K \sum_{l\neq k}\left\{\frac{1}{t^{(k,l)}}\Big(\sum_{i\in G_k^*} v_i r_i^{(k,l)}\Big)\Big(\sum_{j\in G_l^*} v_j c_j^{(k,l)}\Big)\right\}.$$

Using once again the property $\sum_{i\in G_k^*} v_i = 0$ for all $k \in [K]$, we can simplify

$$\begin{aligned}
&\sum_{j\in G_l^*} v_j c_j^{(k,l)} \\
&= \sum_{j\in G_l^*} v_j\Big[-\frac{n_l+n_k}{2n_l}\lambda + \frac{n_k}{2}(\|\overline{\mathbf{X}}_k\|^2 - \|\overline{\mathbf{X}}_l\|^2) + n_k\langle \mathbf{X}_j, \overline{\mathbf{X}}_l - \overline{\mathbf{X}}_k\rangle\Big] \\
&= n_k\langle \overline{\mathbf{X}}_l - \overline{\mathbf{X}}_k, \sum_{j\in G_l^*} v_j\mathbf{X}_j\rangle \\
&= n_k\langle \overline{\mathbf{X}}_l - \overline{\mathbf{X}}_k, \sum_{j\in G_l^*} v_j\boldsymbol{\varepsilon}_j\rangle \\
&= n_k\langle \boldsymbol{\mu}_l - \boldsymbol{\mu}_k + \overline{\boldsymbol{\varepsilon}}_l - \overline{\boldsymbol{\varepsilon}}_k, \sum_{j\in G_l^*} v_j\boldsymbol{\varepsilon}_j\rangle.
\end{aligned}$$

Similarly, $\sum_{i\in G_k^*} v_i r_i^{(k,l)} = n_l\langle \boldsymbol{\mu}_k - \boldsymbol{\mu}_l + \overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_l, \sum_{i\in G_k^*} v_i\boldsymbol{\varepsilon}_i\rangle$. Then

$$\sum_{i\in G_k^*}\sum_{j\in G_l^*} v_i v_j r_i^{(k,l)}c_j^{(k,l)} = -n_k n_l(T_{1,kl} + T_{2,kl} + T_{3,kl}),$$

and

$$T(\mathbf{v}) = -\sum_{k=1}^K\sum_{l\neq k}\left\{\frac{n_k n_l}{t^{(k,l)}}(T_{1,kl} + T_{2,kl} + T_{3,kl})\right\}, \tag{39}$$

where

$$T_{1,kl}(\mathbf{v}) = \Big[\sum_{i\in G_k^*} v_i\langle \boldsymbol{\mu}_k - \boldsymbol{\mu}_l, \boldsymbol{\varepsilon}_i\rangle\Big]\cdot\Big[\sum_{j\in G_l^*} v_j\langle \boldsymbol{\mu}_k - \boldsymbol{\mu}_l, \boldsymbol{\varepsilon}_j\rangle\Big],$$
$$T_{2,kl}(\mathbf{v}) = \Big[\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_l, \boldsymbol{\varepsilon}_i\rangle\Big]\cdot\Big[\sum_{j\in G_l^*} v_j\langle \overline{\boldsymbol{\varepsilon}}_k - \overline{\boldsymbol{\varepsilon}}_l, \boldsymbol{\varepsilon}_j\rangle\Big],$$

$$T_{3,kl}(\mathbf{v}) = \Big[\sum_{i\in G_k^*} v_i\langle \boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_i\rangle\Big]\cdot\Big[\sum_{j\in G_l^*} v_j\langle \overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_j\rangle\Big] + \Big[\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle\Big]\cdot\Big[\sum_{j\in G_l^*} v_j\langle \boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_j\rangle\Big].$$

To bound these three terms, we will use the following lemma, whose proof is deferred to the end of this section.

Lemma VII.1 (Uniform High Probability Bounds for Random Fluctuation Terms): For any $\delta > 0$, it holds with probability at least $1-4K^2n^{-\delta}$ that for any $\mathbf{v}\in\Gamma_K$ and any distinct pair $(k,l)\in[K]^2$,

$$\Big|\sum_{i\in G_k^*} v_i\langle \boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_i\rangle\Big| \leqslant \|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|\big(n_k+\sqrt{2n_k\log n}+2\log n\big)^{1/2}\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}, \tag{40}$$

$$\Big|\sum_{i\in G_k^*} v_i\,\|\boldsymbol{\varepsilon}_i\|^2\Big| \leqslant Cp^{1/2}[n_k^{1/2}+\log^2(n)\,]\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}, \tag{41}$$

$$\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle \leqslant C\sqrt{\frac{(p+\log n)\,n_k}{n_l}}\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}, \tag{42}$$

$$\Big|\sum_{i\in G_k^*} v_i\Big\langle\frac{n_k-1}{n_k}\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}},\boldsymbol{\varepsilon}_i\Big\rangle\Big| \leqslant Cp^{1/2}(\delta\log n)^{1/2}\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}, \tag{43}$$

for some universal constant $C>0$.

Bound $T_{1,kl}$: By applying the Cauchy-Schwarz inequality and inequality (40), we can bound

$$\begin{aligned}|T_{1,kl}(\mathbf{v})| &= \Big|\sum_{i\in G_k^*} v_i\langle \boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_i\rangle\sum_{j\in G_l^*} v_j\langle \boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_j\rangle\Big|\\ &\leqslant \|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\Big(\sum_{j\in G_l^*} v_j^2\Big)^{1/2}\\ &\quad\cdot\big(n_k+\sqrt{2n_k\log n}+2\log n\big)^{1/2}\big(n_l+\sqrt{2n_l\log n}+2\log n\big)^{1/2}.\end{aligned}$$

Throughout the proof, we can always work under the event

$$\{t^{(k,l)}\geqslant \beta n_k n_l\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2/2 \text{ for all distinct pairs } (k,l)\in[K]^2 \text{ and } i\in G_k^*\}, \tag{44}$$

which according to the choice of $\lambda^\sharp$ in (25) after Lemma IV.1, holds with probability at least $1-K^2n^{-\delta}-8n^{-1}$. Under this event, we get a uniform bound for first sum of $T_{1,kl}$'s in the decomposition (39) of $T(\mathbf{v})$ for all $\mathbf{v}\in\Gamma_K$:

$$\begin{aligned}&\Big|\sum_{k,l=1}^K \frac{n_k n_k}{t^{(k,l)}}T_{1,kl}(\mathbf{v})\Big|\\ &\leqslant \frac{2}{\beta}\Big\{\sum_{k=1}^K\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\big(n_k+\sqrt{2n_k\log n}+2\log n\big)^{1/2}\Big\}\\ &\quad\cdot\Big\{\sum_{l=1}^K\Big(\sum_{j\in G_l^*} v_j^2\Big)^{1/2}\big(n_l+\sqrt{2n_l\log n}+2\log n\big)^{1/2}\Big\}\\ &\overset{(a)}{\leqslant} \frac{2}{\beta}\Big(\sum_{k=1}^K\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\cdot\Big(\sum_{k=1}^K\big(n_k+\sqrt{2n_k\log n}+2\log n\big)\Big)^{1/2}\\ &\quad\cdot\Big(\sum_{l=1}^K\sum_{j\in G_l^*} v_j^2\Big)^{1/2}\cdot\Big(\sum_{l=1}^K\big(n_l+\sqrt{2n_l\log n}+2\log n\big)\Big)^{1/2}.\\ &\leqslant \frac{2}{\beta}\big(n+\sqrt{2nK\log n}+2K\log n\big)\,\|\mathbf{v}\|^2,\end{aligned}$$

where step (a) is due to the Cauchy-Schwarz inequality, and the last step uses the identity $\sum_{k=1}^K n_k = n$ and inequality $\sum_{k=1}^K\sqrt{n_k}\leqslant\sqrt{K\sum_{k=1}^K n_k}$.

Bound $T_{2,kl}$: Due to the symmetry, we only need to analyze the first sum in $T_{2,kl}$, which can be further decomposed as

$$\begin{aligned}&\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle\\ &=\sum_{i\in G_k^*} v_i\Big\langle\frac{1}{n_k}\boldsymbol{\varepsilon}_i+\frac{n_k-1}{n_k}\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\Big\rangle\\ &=\sum_{i\in G_k^*}\frac{v_i}{n_k}\|\boldsymbol{\varepsilon}_j\|^2-\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle+\sum_{i\in G_k^*} v_i\Big\langle\frac{n_k-1}{n_k}\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}},\boldsymbol{\varepsilon}_i\Big\rangle\\ &=:G_1(\mathbf{v})+G_2(\mathbf{v})+G_3(\mathbf{v}),\end{aligned}$$

where the three terms $G_1(\mathbf{v})$, $G_2(\mathbf{v})$ and $G_3(\mathbf{v})$ are respectively bounded by using inequalities (41), (42) and (43) in Lemma VII.1. Therefore, we can reach

$$\begin{aligned}&\sum_{i\in G_k^*} v_i\langle \overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle\\ &\leqslant C\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\\ &\quad\cdot\Big(\sqrt{\frac{p}{n_k}}+\frac{\log^2(n)\sqrt{p}}{n_k}+\sqrt{\frac{(p+\log n)\log n_k}{n_l}}+\sqrt{\delta p\log n}\Big)\end{aligned}$$

$$\begin{aligned}U^* &:= \mathbb{P}\Big(2\langle\boldsymbol{\theta}-\overline{\boldsymbol{\varepsilon}}_l+(1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}},\,-\boldsymbol{\varepsilon}_i\rangle\geqslant(1-\beta)\|\boldsymbol{\theta}\|^2\,\Big|\,\overline{\boldsymbol{\varepsilon}}_l,\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\Big)\\ &= \Phi^c\left(\frac{(1-\beta)\|\boldsymbol{\theta}\|^2}{2\sqrt{\|\boldsymbol{\theta}\|^2+2\langle\boldsymbol{\theta},(1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}-\overline{\boldsymbol{\varepsilon}}_l\rangle+\|\overline{\boldsymbol{\varepsilon}}_l-(1-n_k^{-1})^2\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}\|^2}}\right)\\ &\leqslant \Phi^c\left(\frac{(1-\beta)\|\boldsymbol{\theta}\|^2}{2\sqrt{\|\boldsymbol{\theta}\|^2+2\sqrt{2(n_l^{-1}+n_k^{-1})\zeta_n}\|\boldsymbol{\theta}\|+(n_l^{-1}+n_k^{-1})(p+2\sqrt{p\zeta_n}+2\zeta_n)}}\right).\end{aligned} \tag{38}$$

$$\leqslant C'\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\bigg(\sqrt{\delta p\log n}+\log^2(n)\sqrt{\frac{p}{\underline{n}}}\bigg).$$

This implies the following bound on $T_{2,kl}$ due to the symmetry,

$$|T_{2,kl}(\mathbf{v})| \leqslant C''\bigg(\delta p\log n+\frac{p\log^4(n)}{\underline{n}}\bigg)\cdot\Big(\sum_{i\in G_k^*} v_i^2\Big)^{1/2}\Big(\sum_{j\in G_l^*} v_j^2\Big)^{1/2}.$$

Then we may obtain by using the lower bound condition in Lemma IV.1 as $\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^2\geqslant C_1(1-\beta)^{-1}\sqrt{(1+\delta)\,p\log n/m}$ that under the event (44),

$$\begin{aligned}&\bigg|\sum_{k,l=1}^K\frac{n_kn_k}{t^{(k,l)}}T_{2,kl}(\mathbf{v})\bigg|\\ &\leqslant\frac{C_2(1-\beta)}{\beta}\Big(\delta\sqrt{mp\log n}+\frac{\sqrt{mp\log^7 n}}{\underline{n}}\Big)\\ &\quad\cdot\bigg(\sum_{k=1}^K\Big(\sum_{i\in G_k^*}v_i^2\Big)^{1/2}\bigg)\bigg(\sum_{l=1}^K\Big(\sum_{j\in G_l^*}v_j^2\Big)^{1/2}\bigg)\\ &\leqslant\frac{C_2(1-\beta)K}{\beta}\Big(\delta\sqrt{mp\log n}+\frac{\sqrt{mp\log^7 n}}{\underline{n}}\Big)\|\mathbf{v}\|^2,\end{aligned}$$

where the last step is due to the Cauchy-Schwarz inequality.

Bound $T_{3,kl}$: Note that term $|T_{3,kl}(\mathbf{v})|$ satisfies

$$\begin{aligned}&|T_{3,kl}(\mathbf{v})|\\ &\leqslant\frac{1}{2}\Big(\sum_{i\in G_k^*}v_i\langle\boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_i\rangle\Big)^2+\frac{1}{2}\Big(\sum_{j\in G_l^*}v_j\langle\boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_j\rangle\Big)^2\\ &\quad+\frac{1}{2}\Big(\sum_{i\in G_k^*}v_i\langle\overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_i\rangle\Big)^2+\frac{1}{2}\Big(\sum_{j\in G_l^*}v_j\langle\overline{\boldsymbol{\varepsilon}}_k-\overline{\boldsymbol{\varepsilon}}_l,\boldsymbol{\varepsilon}_j\rangle\Big)^2.\end{aligned}$$

Therefore, $|T_{3,kl}(\mathbf{v})|$ can be bounded by the sum of the upper bounds for $|T_{1,kl}(\mathbf{v})|$ and $|T_{2,kl}(\mathbf{v})|$.

Putting all pieces together, we can finally reach

$$\begin{aligned}&|T(\mathbf{v})|\\ &\leqslant\Big|\sum_{k=1}^K\sum_{l\neq k}\frac{n_kn_l}{t^{(k,l)}}T_{1,kl}\Big|+\Big|\sum_{k=1}^K\sum_{l\neq k}\frac{n_kn_l}{t^{(k,l)}}T_{2,kl}\Big|+\Big|\sum_{k=1}^K\sum_{l\neq k}\frac{n_kn_l}{t^{(k,l)}}T_{3,kl}\Big|\\ &\leqslant\frac{C_3}{\beta}\|\mathbf{v}\|^2\Big(n+K\log n+(1-\beta)K\delta\sqrt{mp\log n}+\frac{\sqrt{mp\log^7 n}}{\underline{n}}\Big).\end{aligned}$$

C. Proof of Lemma VII.1

We can apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}&\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|^{-1}\Big|\sum_{i\in G_k^*}v_i\langle\boldsymbol{\mu}_k-\boldsymbol{\mu}_l,\boldsymbol{\varepsilon}_i\rangle\Big|\\ &\leqslant\Big(\sum_{i\in G_k^*}v_i^2\Big)^{1/2}\Big(\sum_{i\in G_k^*}\Big\langle\frac{\boldsymbol{\mu}_k-\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|},\boldsymbol{\varepsilon}_i\Big\rangle^2\Big)^{1/2}.\end{aligned}$$

Since

$$\Big\langle\frac{\boldsymbol{\mu}_k-\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|},\boldsymbol{\varepsilon}_i\Big\rangle\overset{\text{i.i.d.}}{\sim}N(0,1),\quad i=1,\ldots,n,$$

we obtain by Lemma VIII.1 and a union bound argument that with probability at least $1-K^2n^{-1}$,

$$\sum_{i\in G_k^*}\Big\langle\frac{\boldsymbol{\mu}_k-\boldsymbol{\mu}_l}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_l\|},\boldsymbol{\varepsilon}_i\Big\rangle^2\leqslant n_k+\sqrt{2n_k\log n}+2\log n$$

for all $k,l\in[K]$. A combination of the preceding two displays yields the first claimed inequality (40).

Since $\sum_{i\in G_k^*}v_i=0$ for any $\mathbf{v}\in\Gamma_K$, we can also write the left hand side of inequality (41) as $G_1(\mathbf{v}_k)=\sum_{i\in G_k^*}v_i(\|\boldsymbol{\varepsilon}_i\|^2-p)$, which can be viewed as a centered empirical process indexed by $\mathbf{v}_k\in\mathbb{R}^{n_k}$, the restriction $\mathbf{v}_{\restriction G_k^*}$ of $\mathbf{v}\in\Gamma_K$ onto G_k^*. We may assume without loss of generality that $\mathbf{v}_k\in\mathbb{V}_k:=\{\mathbf{v}_{\restriction G_k^*}:\mathbf{v}\in\Gamma_K,\|\mathbf{v}_k\|=1\}$. By Theorem 4 in [65], there exists a universal constant C such that for any $t>0$,

$$\begin{aligned}&\mathbb{P}\Big(\|G_1\|_{\mathbb{V}_k}\geqslant 2\,\mathbb{E}[\|G_1\|_{\mathbb{V}_k}]+t\Big)\\ &\qquad\leqslant\exp\Big(-\frac{t^2}{3\tau_1^2}\Big)+3\exp\Big(-\frac{t}{C\|\mathsf{M}_1\|_{\psi_1}}\Big).\end{aligned}$$

where $\|G_1\|_{\mathbb{V}_k}=\sup_{\mathbf{v}_k\in\mathbb{V}_k}|G_1(\mathbf{v}_k)|$ and $\tau_1^2=\sup_{\mathbf{v}_k\in\mathbb{V}_k}\sum_{i\in G_k^*}v_i^2\,\mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^2-p]^2\leqslant 2p$, and $\mathsf{M}_1=\max_{i\in G_k^*}\max_{\mathbf{v}_k\in\mathbb{V}_k}|v_i(\|\boldsymbol{\varepsilon}_i\|^2-p)|\leqslant\max_{i\in G_k^*}|\|\boldsymbol{\varepsilon}_i\|^2-p|$. By the maximal inequality in Lemma 2.2.2 in [66] and Lemma VIII.3, we have

$$\|\mathsf{M}_1\|_{\psi_1}\leqslant C\log(n_k)\max_{i\in G_k^*}\|\|\boldsymbol{\varepsilon}_i\|^2-p\|_{\psi_1}\leqslant Cp^{1/2}\log(n_k).$$

By the Cauchy-Schwarz inequality, we have for all $\mathbf{v}\in\mathbb{V}_k$,

$$\begin{aligned}\Big|\sum_{i\in G_k^*}v_i(\|\boldsymbol{\varepsilon}_i\|^2-p)\Big|&\leqslant\Big(\sum_{i\in G_k^*}v_i\Big)^{1/2}\Big(\sum_{i\in G_k^*}(\|\boldsymbol{\varepsilon}_i\|^2-p)^2\Big)^{1/2}\\ &\leqslant\Big(\sum_{i\in G_k^*}(\|\boldsymbol{\varepsilon}_i\|^2-p)^2\Big)^{1/2}.\end{aligned}$$

Then Jensen's inequality implies that

$$\mathbb{E}[\|G_1\|_{\mathbb{V}_k}]\leqslant\Big[\sum_{i\in G_k^*}\mathbb{E}(\|\boldsymbol{\varepsilon}_i\|^2-p)^2\Big]^{1/2}=(2n_kp)^{1/2}.$$

Thus with probability at least $1-4n^{-1}$, we have

$$\|G_1\|_{\mathbb{V}_k}\leqslant Cp^{1/2}[n_k^{1/2}+\log^2(n)],\tag{45}$$

which entails the second claimed inequality (41).

Next we prove the third claimed inequality. Note that conditional on $\overline{\boldsymbol{\varepsilon}}_l$, $G_2(\mathbf{v}_k):=\langle\overline{\boldsymbol{\varepsilon}}_l,\sum_{i\in G_k^*}v_i\boldsymbol{\varepsilon}_i\rangle\sim N(0,\|\overline{\boldsymbol{\varepsilon}}_l\|^2\sum_{i\in G_k^*}v_i^2)$ is a centered Gaussian process indexed by $\mathbf{v}_k\in\mathbb{V}_k$. By the Borell-Sudakov-Tsirel'son inequality (cf. Theorem 2.5.8 in [67]), we have

$$\mathbb{P}\Big(\|G_2\|_{\mathbb{V}_k}\geqslant\mathbb{E}[\|G_2\|_{\mathbb{V}_k}\,|\,\overline{\boldsymbol{\varepsilon}}_l]+\tau_2\sqrt{2\log n}\;\Big|\;\overline{\boldsymbol{\varepsilon}}_l\Big)\leqslant n^{-1},$$

where $\tau_2^2=\|\overline{\boldsymbol{\varepsilon}}_l\|^2\sup_{\mathbf{v}_k\in\mathbb{V}_k}\sum_{i\in G_k^*}v_i^2\leqslant\|\overline{\boldsymbol{\varepsilon}}_l\|^2$. Then Dudley's entropy integral bound (cf. Corollary 2.2.8 in [66]) yields that

$$\mathbb{E}[\|G_2\|_{\mathbb{V}_k}\,|\,\overline{\boldsymbol{\varepsilon}}_l]\leqslant C\|\overline{\boldsymbol{\varepsilon}}_l\|n_k^{1/2},$$

where we have used the fact that the ε-covering entropy of the unit sphere in $\mathbb{R}^{n_k}$ is at most $Cn_k \log(1/\varepsilon)$ for any $\varepsilon \in (0,1)$. Combining the last two displays with the inequality

$$\mathbb{P}(\|\overline{\boldsymbol{\varepsilon}}_l\|^2 \geqslant n_l^{-1}(p + 2\sqrt{p\log n} + 2\log n)) \leqslant n^{-1},$$

and a union bound argument, we get with probability at least $1 - K^2 n^{-1}$,

$$\|G_2\|_{\mathbb{V}_k} \leqslant C\sqrt{\frac{(p+\log n)n_k}{n_l}}, \tag{46}$$

implying the third inequality (42).

Now we prove the last inequality. Note that

$$\begin{aligned}\sum_{i\in G_k^*} v_i \Big\langle \frac{n_k - 1}{n_k}\overline{\boldsymbol{\varepsilon}}_{k\setminus\{i\}}, \boldsymbol{\varepsilon}_i\Big\rangle &= \frac{1}{n_k}\sum_{\{(i,j)\in G_k^*: i\neq j\}} v_i\langle \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j\rangle \\ &=: \frac{1}{n_k} U_1(\mathbf{v}_k),\end{aligned}$$

where

$$U_1(\mathbf{v}_k) = \sum_{\{(i,j)\in G_k^*: i\neq j\}} \frac{1}{2}(v_i + v_j)\langle \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j\rangle$$

is a degenerate U-process of order two. To simplify the notation, we may assume $G_k^* = \{1, \ldots, n_k\}$ in the rest of this proof. Applying Lemma VIII.4 with

$$\mathcal{A} = \{A \otimes \mathrm{Id}_p \,|\, A = \{a_{ij}\}_{i,j\in[n_k]},\, a_{ij} = (v_i + v_j)/2,\, \mathbf{v}_k \in \mathbb{V}_k\},$$

we get

$$\begin{aligned}&\mathbb{P}\Big(\Big|\|U_1\|_{\mathbb{V}_k} - \mathbb{E}[\|U_1\|_{\mathbb{V}_k}]\Big| \geqslant t\Big) \\ &\quad\leqslant 2\exp\Big[-C\min\Big(\frac{t^2}{\|\boldsymbol{\varepsilon}\|_{\mathcal{A}}^2}, \frac{t}{\sup_{A\in\mathcal{A}}\|A\otimes \mathrm{Id}_p\|_{\mathrm{op}}}\Big)\Big]\end{aligned}$$

where $\|U_1\|_{\mathbb{V}_k} = \sup_{\mathbf{v}_k\in\mathbb{V}_k} U_1(\mathbf{v}_k)$, $\boldsymbol{\varepsilon}^T = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_{n_k}^T)$, and

$$\|\boldsymbol{\varepsilon}\|_{\mathcal{A}} = \mathbb{E}\Big[\sup_A \Big\|(A\otimes \mathrm{Id}_p + A^T \otimes \mathrm{Id}_p)\boldsymbol{\varepsilon}\Big\|\Big].$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\|A\|_{\mathrm{op}} &= \max_{\|\mathbf{u}\|=1} \mathbf{u}^T A\mathbf{u} = \max_{\|\mathbf{u}\|=1}\sum_{i,j=1}^{n_l} u_i u_j \frac{v_i + v_j}{2} \\ &= \max_{\|\mathbf{u}\|=1}\Big(\sum_{i=1}^{n_k} u_i v_i\Big)\Big(\sum_{j=1}^{n_k} u_j\Big) \\ &\leqslant \max_{\|\mathbf{u}\|=1}\Big(\sum_{i=1}^{n_k} u_i^2\Big)^{1/2}\Big(\sum_{i=1}^{n_k} v_i^2\Big)^{1/2}\Big(\sum_{j=1}^{n_k} u_j^2\Big)^{1/2} n_k^{1/2} \leqslant n_k^{1/2}.\end{aligned}$$

Since $\|A\otimes \mathrm{Id}_p\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}$, we have

$$\sup_A \|(A\otimes \mathrm{Id}_p)\boldsymbol{\varepsilon}\| \leqslant \sup_A \|A\otimes \mathrm{Id}_p\|_{\mathrm{op}}\|\boldsymbol{\varepsilon}\| \leqslant n_l^{1/2}\|\boldsymbol{\varepsilon}\|.$$

Then Jensen's inequality yields that

$$\|\boldsymbol{\varepsilon}\|_{\mathcal{A}}^2 \leqslant 4n_k\,\mathbb{E}[\|\boldsymbol{\varepsilon}\|^2] = 4n_k^2 p.$$

To bound $\mathbb{E}[\|U_1\|_{\mathbb{V}_k}]$, we note that

$$\begin{aligned}|U_1(\mathbf{v}_k)| &= \Big|\sum_{j=1}^{n_k} v_j \Big\langle \boldsymbol{\varepsilon}_j, \sum_{i\neq j, i\in[n_k]} \boldsymbol{\varepsilon}_j\Big\rangle\Big| \\ &\leqslant \Big(\sum_{j=1}^{n_k} v_j^2\Big)^{1/2}\Big(\sum_{j=1}^{n_k}\Big\langle \boldsymbol{\varepsilon}_j, \sum_{i\neq j, i\in[n_k]}\boldsymbol{\varepsilon}_j\Big\rangle^2\Big)^{1/2} \\ &\leqslant \Big(\sum_{j=1}^{n_k}\Big\langle \boldsymbol{\varepsilon}_j, \sum_{i\neq j, i\in[n_k]}\boldsymbol{\varepsilon}_j\Big\rangle^2\Big)^{1/2}.\end{aligned}$$

From Jensen's inequality and the independence between ε_j and $\sum_{i\neq j, i\in[n_k]}\boldsymbol{\varepsilon}_j$, we have

$$\begin{aligned}\mathbb{E}[\|U_1\|_{\mathbb{V}_k}] &\leqslant \Big(\sum_{j=1}^{n_k}\mathbb{E}\Big[\Big\langle \boldsymbol{\varepsilon}_j, \sum_{i\neq j, i\in[n_k]}\boldsymbol{\varepsilon}_j\Big\rangle^2\Big]\Big)^{1/2} \\ &= \Big(\sum_{j=1}^{n_k}\mathbb{E}\Big[\Big\|\sum_{i\neq j, i\in[n_k]}\boldsymbol{\varepsilon}_j\Big\|^2\Big]\Big)^{1/2} \leqslant n_k p^{1/2}.\end{aligned}$$

Thus we see that with probability at least $1 - 2n^{-\delta}$,

$$\|U_1\|_{\mathbb{V}_k} \leqslant Cn_k p^{1/2}(\delta\log n)^{1/2}, \tag{47}$$

which implies the last claimed inequality (43).

VIII. Supporting Lemmas

Lemma VIII.1 (Tail Bound for χ^2 Distributions): If $\mathbf{Z} \sim N(\mathbf{0}, \mathrm{Id}_p)$, then for all $t > 0$,

$$\begin{aligned}\mathbb{P}(\|\mathbf{Z}\|^2 \geqslant p + 2\sqrt{pt} + 2t) &\leqslant e^{-t}, \\ \mathbb{P}(\|\mathbf{Z}\|^2 \leqslant p - 2\sqrt{pt}) &\leqslant e^{-t}.\end{aligned}$$

Proof of Lemma VIII.1: See Lemma 1 in [68]. ■

Lemma VIII.2 (Deviation of Gaussian Random Matrices): If $\mathcal{E} \in \mathbb{R}^{p\times n}$ has i.i.d. $N(0,\sigma^2)$ entries, then

$$\mathbb{P}\big(\|\mathcal{E}\|_{\mathrm{op}} \geqslant \sigma(\sqrt{n} + \sqrt{p} + \sqrt{2t})\big) \leqslant e^{-t}, \quad \forall t > 0.$$

Proof of Lemma VIII.2: See Corollary 5.35 in [69]. ■

Lemma VIII.3: Let $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$ be i.i.d. $N(\mathbf{0}, \mathrm{Id}_p)$. Then there exists a universal constant C such that

$$\|\|\boldsymbol{\varepsilon}_1\|^2 - p\|_{\psi_1} + \|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle\|_{\psi_1} \leqslant Cp^{1/2}.$$

Proof of Lemma VIII.3: Note that $\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle = \sum_{j=1}^p \varepsilon_{1j}\varepsilon_{2j}$, and each additive component $\varepsilon_{1j}\varepsilon_{2j}$ is sub-exponential with $\|\varepsilon_{1j}\varepsilon_{2j}\|_{\psi_1} \leqslant \|\varepsilon_{1j}\|_{\psi_2}\|\varepsilon_{2j}\|_{\psi_2} = 1$ (cf. Lemma 2.7.7 in [58]). By Bernstein's inequality (cf. Theorem 2.8.2 in [58]), there exists a universal constant C_1 such that for any $t > 0$,

$$\mathbb{P}(|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle| \geqslant t) \leqslant 2\exp[-C_1\min(t^2/p, t)].$$

Let C be a large positive real number. By integration-by-parts and change-of-variables, we have

$$\begin{aligned}&\mathbb{E}\Big[\exp\Big(\frac{|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle|}{C}\Big)\Big] \\ &= \int_1^\infty \mathbb{P}\Big(\exp\Big(\frac{|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle|}{C}\Big) > t\Big)\,\mathrm{d}t \\ &= \int_1^\infty \mathbb{P}\Big(|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle| > C\log t\Big)\,\mathrm{d}t \\ &= \int_0^\infty \mathbb{P}\Big(|\langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\rangle| > Cx\Big)\mathrm{e}^x\,\mathrm{d}x \\ &\leqslant 2\int_0^{p/C} \mathrm{e}^{-\frac{C_1C^2}{p}x^2 + x}\,\mathrm{d}x + 2\int_{p/C}^\infty \mathrm{e}^{-(C_1C-1)x}\,\mathrm{d}x\end{aligned}$$

$$\leqslant 2\mathrm{e}^{\frac{p}{4C_1C^2}}\sqrt{\frac{\pi p}{C_1C^2}}+\frac{2}{C_1C-2}\mathrm{e}^{-(C_1C-2)\frac{p}{C}}.$$

Thus if we take $C=Kp^{1/2}$ for some large enough universal constant $K>0$, then

$$\mathbb{E}\left[\exp\left(\frac{|\langle\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2\rangle|}{C}\right)\right]\leqslant 2,$$

which implies that $\|\langle\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2\rangle\|_{\psi_1}\leqslant Kp^{1/2}$. The ψ_1 norm bound for $\|\boldsymbol{\varepsilon}_1\|^2-p$ follows from similar lines. ■

Lemma VIII.4 (Uniform Hanson-Wright Inequality for Gaussian Quadratic Forms): Let $\boldsymbol{\varepsilon}\sim N(0,\mathrm{Id}_p)$ and $\mathcal{A}$ be a bounded class of $p\times p$ matrices. Consider the random variable

$$Z=\sup_{A\in\mathcal{A}}(\boldsymbol{\varepsilon}^TA\boldsymbol{\varepsilon}-\mathbb{E}[\boldsymbol{\varepsilon}^TA\boldsymbol{\varepsilon}]).$$

Then there exists a universal constant C such that for any $t>0$,

$$\mathbb{P}(|Z-\mathbb{E}[Z]|\geqslant t)\leqslant 2\exp\Big[-C\min\Big(\frac{t^2}{\|\boldsymbol{\varepsilon}\|_{\mathcal{A}}^2},\frac{t}{\sup_{A\in\mathcal{A}}\|A\|_{\mathrm{op}}}\Big)\Big],$$

where $\|\boldsymbol{\varepsilon}\|_{\mathcal{A}}=\mathbb{E}[\sup_{A\in\mathcal{A}}\|(A+A^T)\boldsymbol{\varepsilon}\|]$.

Proof of Lemma VIII.4: Note that the standard Gaussian random vector $\boldsymbol{\varepsilon}$ satisfies the concentration inequality

$$\mathbb{P}(|\varphi(\boldsymbol{\varepsilon})-\mathbb{E}[\varphi(\boldsymbol{\varepsilon})]|\geqslant t)\leqslant 2\exp(-t^2/2)$$

for any $t>0$ and every 1-Lipschitz function $\varphi:\mathbb{R}^p\to\mathbb{R}$ such that $\mathbb{E}[|\varphi(\boldsymbol{\varepsilon})|]<\infty$ (cf. Theorem 2.5.7 in [67]). Then the lemma follows from Theorem 2.10 in [70]. ■

ACKNOWLEDGMENT

The authors would like to thank two anonymous referees and the Associate Editor Prof. Stephane Boucheron for their many constructive comments. The author Xiaohui Chen acknowledges that the part of this work was carried out in the Institute for Data, System, and Society (IDSS) at Massachusetts Institute of Technology.

REFERENCES

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Sympos. Math. Statist. Probab.*, 1967, pp. 281–297.

[2] S. Dasgupta, "The hardness of K-means clustering," Univ. California, San Diego, CA, USA, Tech. Rep. CS2007-0890, 2007.

[3] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar K-means problem is np-hard," in *WALCOM: Algorithms Computing*, S. Das and R. Uehara, Eds. Berlin, Germany: Springer, 2009, pp. 274–285.

[4] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[6] R. Kannan and S. Vempala, "Spectral algorithms," *Found. Trends Theor. Comput. Sci.*, vol. 4, nos. 3–4, pp. 157–288, 2008.

[7] Y. Lu and H. H. Zhou, "Statistical and computational guarantees of Lloyd's algorithm and its variants," 2016, *arXiv:1612.02099*. [Online]. Available: http://arxiv.org/abs/1612.02099

[8] M. Ndaoud, "Sharp optimal recovery in the two-component Gaussian mixture model," 2018, *arXiv:1812.08078*. [Online]. Available: http://arxiv.org/abs/1812.08078

[9] J. Peng and Y. Wei, "Approximating K-means-type clustering via semidefinite programming," *SIAM J. Optim.*, vol. 18, no. 1, pp. 186–205, Jan. 2007.

[10] D. G. Mixon, S. Villar, and R. Ward, "Clustering sub Gaussian mixtures by semidefinite programming," *Inf. Inference, A, J. IMA*, vol. 6, no. 4, pp. 389–415, Dec. 2017.

[11] X. Li, Y. Li, S. Ling, T. Strohmer, and K. Wei, "When do birds of a feather flock together? K-means, proximity, and conic programming," *Math. Program.*, vol. 179, nos. 1–2, pp. 295–341, Jan. 2020.

[12] Y. Fei and Y. Chen, "Hidden integrality of SDP relaxations for sub-Gaussian mixture models," in *Proc. 31st Conf. On Learn. Theory*, vol. 75, S. Bubeck, V. Perchet, and P. Rigollet, Eds. Stockholm, Sweden: PMLR, Jul. 2018, pp. 1931–1965. [Online]. Available: http://proceedings.mlr.press/v75/fei18a.html

[13] M. Royer, "Adaptive clustering through semidefinite programming," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1795–1803. [Online]. Available: http://papers.nips.cc/paper/6776-adaptive-clustering-through-semidefinite-programming.pdf

[14] C. Giraud and N. Verzelen, "Partial recovery bounds for clustering with the relaxed K-means," *Math. Statist. Learn.*, vol. 1, no. 3, pp. 317–374, 2018.

[15] F. Bunea, C. Giraud, M. Royer, and N. Verzelen, "PECOK: A convex optimization approach to variable clustering," 2016, *arXiv:1606.05100*. [Online]. Available: http://arxiv.org/abs/1606.05100

[16] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, Jun. 2002.

[17] J. Chen, "Optimal rate of convergence for finite mixture models," *Ann. Statist.*, vol. 23, no. 1, pp. 221–233, Feb. 1995.

[18] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, no. 1, pp. 77–120, Feb. 2017.

[19] J. Xu, D. Hsu, and A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 2684–2692. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157382.3157398

[20] J. M. Klusowski and W. D. Brinda, "Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM," 2016, *arXiv:1608.02280*. [Online]. Available: http://arxiv.org/abs/1608.02280

[21] B. Yan, M. Yin, and P. Sarkar, "Convergence of gradient em on multi-component mixture of Gaussians," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6959–6969. [Online]. Available: http://dl.acm.org/citation.cfm?id=3295222.3295439

[22] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of em suffice for mixtures of two Gaussians," in *Proc. Conf. Learn. Theory*, vol. 65, S. Kale and O. Shamir, Eds. Amsterdam, The Netherlands: PMLR, 2017, pp. 704–710. [Online]. Available: http://proceedings.mlr.press/v65/daskalakis17b.html

[23] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, "Singularity, misspecification and the convergence rate of EM," *Ann. Statist.*, vol. 48, no. 6, pp. 3161–3182, Dec. 2020.

[24] Y. Wu and H. H. Zhou, "Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations," 2019, *arXiv:1908.10935*. [Online]. Available: http://arxiv.org/abs/1908.10935

[25] R. Dwivedi, N. Ho, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu, "Sharp analysis of expectation-maximization for weakly identifiable models," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 108, S. Chiappa and R. Calandra, Eds. Palermo, Italy: PMLR, Aug. 2020, pp. 1866–1876. [Online]. Available: http://proceedings.mlr.press/v108/dwivedi20a.html

[26] D. Pollard, "Strong consistency of K-means clustering," *Ann. Statist.*, vol. 9, no. 1, pp. 135–140, Jan. 1981.

[27] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward, "Relax, no need to round: Integrality of clustering formulations," in *Proc. Conf. Innov. Theor. Comput. Sci.*, New York, NY, USA, Jan. 2015, pp. 191–200.

[28] M. Meila and J. Shi, "Learning segmentation by random walks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 873–879.

[29] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856.

[30] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *J. Comput. Syst. Sci*, vol. 68, p. 2004, Jun. 2004.

[31] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *Learning Theory*, P. Auer and R. Meir, Eds. Berlin, Germany: Springer, 2005, pp. 458–469.

[32] A. Kumar and R. Kannan, "Clustering with spectral norm and the K-means algorithm," in *Proc. IEEE 51st Annu. Symp. Found. Comput.*, Oct. 2010, pp. 299–308.

[33] P. Awasthi and O. Sheffet, "Improved spectral-norm bounds for clustering," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, A. Gupta, K. Jansen, J. Rolim, and R. Servedio, Eds. Berlin, Germany: Springer, 2012, pp. 37–49.

[34] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[35] U. von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *Ann. Statist.*, vol. 36, no. 2, pp. 555–586, Apr. 2008.

[36] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=1283383.1283494

[37] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of Lloyd-type methods for the k-means problem," *J. ACM*, vol. 59, no. 6, p. 28:1–28:22, Jan. 2013.

[38] A. Nellore and R. Ward, "Recovery guarantees for exemplar-based clustering," *Inf. Comput.*, vol. 245, pp. 165–180, Dec. 2015.

[39] X. Chen and Y. Yang, "Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data," *Bernoulli*, vol. 27, no. 1, pp. 586–614, Feb. 2021.

[40] X. Chen and Y. Yang, "Diffusion K-means clustering on manifolds: Provable exact recovery via semidefinite relaxations," *Appl. Comput. Harmon. Anal.*, vol. 52, pp. 303–347, May 2021.

[41] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Netw.*, vol. 5, no. 2, pp. 109–137, Jun. 1983.

[42] M. E. Dyer and A. M. Frieze, "The solution of some random NP-hard problems in polynomial expected time," *J. Algorithms*, vol. 10, no. 4, pp. 451–489, Dec. 1989.

[43] F. Krzakala *et al.*, "Spectral redemption in clustering sparse networks," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 52, pp. 20935–20940, Dec. 2013. [Online]. Available: https://www.pnas.org/content/110/52/20935

[44] L. Massoulié, "Community detection thresholds and the weak ramanujan property," in *Proc. 46th Annu. ACM Symp. Theory Comput.* New York, NY, USA: ACM, 2014, pp. 694–703.

[45] J. Lei and A. Rinaldo, "Consistency of spectral clustering in stochastic block models," *Ann. Statist.*, vol. 43, no. 1, pp. 215–237, Feb. 2015.

[46] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 471–487, Jan. 2016.

[47] A. A. Amini and E. Levina, "On semidefinite relaxations for the block model," *Ann. Statist.*, vol. 46, no. 1, pp. 149–179, Feb. 2018.

[48] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2788–2797, May 2016.

[49] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming: Extensions," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5918–5937, Oct. 2016.

[50] O. Guédon and R. Vershynin, "Community detection in sparse networks via Grothendieck's inequality," *Probab. Theory Rel. Fields*, vol. 165, nos. 3–4, pp. 1025–1049, Aug. 2016.

[51] A. S. Bandeira, "Random Laplacian matrices and convex relaxations," *Found. Comput. Math.*, vol. 18, no. 2, pp. 345–379, Apr. 2018.

[52] X. Li, Y. Chen, and J. Xu, "Convex relaxation methods for community detection," *Stat. Sci.*, vol. 36, no. 1, pp. 2–15, Feb. 2021.

[53] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model," *Electron. J. Probab.*, vol. 21, p. 24, Jun. 2016.

[54] E. Mossel, J. Neeman, and A. Sly, "Belief propagation, robust reconstruction and optimal recovery of block models," *Ann. Appl. Probab.*, vol. 26, no. 4, pp. 2211–2256, Aug. 2016.

[55] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar, "On the tightness of an SDP relaxation of K-means," 2015, *arXiv:1505.04778*. [Online]. Available: http://arxiv.org/abs/1505.04778

[56] P. Hansen and B. Jaumard, "Cluster analysis and mathematical programming," *Math. Program.*, vol. 79, nos. 1–3, pp. 191–215, Oct. 1997.

[57] D. P. Williamson and D. B. Shmoys, *The Design Approximation Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2011.

[58] R. Vershynin, *High-Dimensional Probability: An Introduction with Application Data Science* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2018.

[59] C. Giraud, *Introduction to High-Dimensional Statistics* (Monographs on Statistics and Applied Probability). vol. 139. Boca Raton, FL, USA: CRC Press, 2015.

[60] J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu, "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 4872–4894, Jul. 2018.

[61] Y. Wu and J. Xu, "Statistical problems with planted structures: Information-theoretical and computational limits," 2018, *arXiv:1806.00118*. [Online]. Available: http://arxiv.org/abs/1806.00118

[62] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Ann. Probab.*, vol. 33, no. 5, pp. 1643–1697, Sep. 2005.

[63] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statist. Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[64] T. Lesieur, C. de Bacco, J. Banks, F. Krzakala, C. Moore, and L. Zdeborova, "Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2016, pp. 601–608. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7852287/

[65] R. Adamczak, "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains," *Electron. J. Probab.*, vol. 13, no. 0, pp. 1000–1034, 2008.

[66] A. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York, NY, USA: Springer-Verlag, 1996.

[67] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[68] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.

[69] R. Vershynin, *Introduction to Non-Asymptotic Analysis Random Matrices*. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 210–268.

[70] R. Adamczak, "A note on the Hanson-Wright inequality for random vectors with dependencies," *Electron. Commun. Probab.*, vol. 20, p. 13, Jun. 2015.

Xiaohui Chen received the Ph.D. degree in electrical and computer engineering from The University of British Columba (UBC), Vancouver, Canada, in 2013. He was a Post-Doctoral Fellow with the Toyota Technological Institute at Chicago (TTIC), a philanthropically endowed academic computer science institute located on the University of Chicago Campus. He held a Visiting Faculty position with the Institute for Data, Systems, and Society (IDSS), Massachusetts Institute of Technology (MIT), from 2019 to 2020. He is currently an Associate Professor of statistics with the University of Illinois at Urbana-Champaign. He received an NSF CAREER Award in 2018, an Arnold O. Beckman Award at UIUC in 2018, an Outstanding Young Researcher Award from the International Chinese Statistical Association (ICSA) in 2019, an Associate Appointment in the Center for Advanced Study at UIUC from 2020 to 2021, and a Simons Fellowship in mathematics from the Simons Foundation from 2020 to 2021.

Yun Yang received the B.S. degree in mathematics from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in statistics from Duke University in 2014. From 2014 to 2016, he was a Post-Doctoral Researcher with the University of California, Berkeley, CA, USA, working on problems in machine learning, optimization, and high-dimensional statistics. From 2016 to 2018, he was an Assistant Professor in statistics with the Florida State University. He is currently an Assistant Professor in statistics with the University of Illinois Urbana-Champaign. His current research interests include scalable statistical computation, statistical learning theory, and machine learning.