Proceedings of Machine Learning Research vol 75:1–34, 2018 31st Annual Conference on Learning Theory

Hidden Integrality of SDP Relaxations for Sub-Gaussian Mixture Models

Yingjie Fei YF275@CORNELL.EDU
Yudong Chen YUDONG.CHEN@CORNELL.EDU
Cornell University

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of finding discrete clustering structures under Sub-Gaussian Mixture Models. We establish a hidden integrality property of a semidefinite programming (SDP) relaxation for this problem: while the optimal solutions to the SDP are not integer-valued in general, their estimation errors can be upper bounded by the error of an idealized integer program. The error of the integer program, and hence that of the SDP, are further shown to decay exponentially in the signal-to-noise ratio. To the best of our knowledge, this is the first exponentially decaying error bound for convex relaxations of mixture models. A special case of this result shows that in certain regimes the SDP solutions are in fact integral and exact, improving on existing exact recovery results for convex relaxations. More generally, our result establishes sufficient conditions for the SDP to correctly recover the cluster memberships of at least $(1-\delta)$ fraction of the points for any $\delta \in (0,1)$. Error bounds for estimating cluster centers can also be derived directly from our results.
Keywords: Sub-Gaussian Mixture Models, semidefinite programming, integer programming.

1. Introduction

We consider the Sub-Gaussian Mixture Models (SGMMs), where one is given n random points drawn from a mixture of k sub-Gaussian distributions with different means. SGMMs, particularly its special case Gaussian Mixture Models (GMMs), are widely used in a broad range of applications including speaker identification, background modeling and online recommendations systems. In these applications, one is typically interested in two inference problems under SGMMs:

- **Clustering:** (approximately) identify the cluster membership of each point, that is, which of the k mixture components generates a given point;
- **Center estimation:** estimate the k centers of an mixture, that is, the means of the k components.

Standard approaches to these problems, such as k-means clustering, typically lead to integer programming problems that are non-convex and NP-hard to optimize (Aloise et al., 2009; Jain et al., 2002; Mahajan et al., 2009). Consequently, much work has been done in developing computationally tractable algorithms for SGMMs, including expectation maximization (Dempster et al., 1977), Lloyd's algorithm (Lloyd, 1982), spectral methods (Vempala and Wang, 2004), the method of moments (Pearson, 1936), and many more. Among them, convex relaxation, including those based on linear programming (LP) and semidefinite programming (SDP), have emerged as an important approach for clustering SGMMs. This approach has several attractive properties: (a) it is solvable in

polynomial time, and does not require a good initial solution to be provided; (b) it has the flexibility to incorporate different quality metrics and additional constraints; (c) it is not restricted to specific forms of SGMMs (such as Gaussian distributions), and is robust against model misspecification (Peng and Xia, 2005; Peng and Wei, 2007; Nellore and Ward, 2015).

Theoretical performance guarantees for convex relaxation methods have been studied in a body of classical and recent work. As will be discussed in the related work section (Section 2), these existing results often have one of the two forms:

1. How well the (rounded) solution of a relaxation optimizes a particular objective function (e.g., the k-means or k-medians objective) compared to the original integer program, as captured by an *approximation factor* (Charikar et al., 1999; Kanungo et al., 2004; Peng and Wei, 2007; Li and Svensson, 2016);

2. When the solution of a relaxation corresponds exactly to the ground-truth clustering, a phenomenon known as *exact recovery*, which is considered in a more recent line of work (Nellore and Ward, 2015; Awasthi et al., 2015; Mixon et al., 2017; Iguchi et al., 2017; Li et al., 2017).

In many practical scenarios, optimizing a particular objective function, and designing approximation algorithms for doing so, is often only a means to the true goal of the problem, namely learning the true underlying model that generates the observed data. Establishing exact recovery guarantees is more directly relevant to this goal. However, such results often require very stringent conditions on the separation or signal-to-noise ratio (SNR) of the model. In practice, convex relaxation solutions are rarely exact, even when the data are generated from the assumed model. On the other hand, it is observed that the solutions, while not exact or integer-valued, are often a good approximation to the desired solution that represents the ground truth. Such a phenomenon is not captured by results on exact recovery.

In this paper, we aim to significantly strengthen our understanding of convex relaxation approaches to SGMMs. In particular, we study the regime where their solutions are not integral in general, and seek to directly characterize the *estimation errors* of the solutions—namely, their distance to desired integer solution corresponding to the true underlying model. For a specific class of SDP relaxations for SGMMs, our results reveal a perhaps surprising property of them: While the SDP solutions are not integer-valued in general, their errors can be controlled by that of the solutions of an idealized integer program (IP), in which one tries to estimate cluster memberships when an oracle reveals the *true centers* of the SGMM. In particular, we show that, in a precise sense to be formalized later, the estimation errors of the SDP and IP satisfy the following relationship (Theorem 1):

$$\text{error(SDP)} \lesssim \text{error(IP)}.$$

We refer to this property as *hidden integrality* of the SDP relaxations; its proof in fact involves showing that the optimal solutions of certain intermediate linear optimization problems are integral. We then further upper bound the error of the IP and show that it decays *exponentially* in terms of the SNR (Theorem 2):

$$\text{error(IP)} \lesssim \exp\left[-\Omega(\text{SNR}^2)\right],$$

where the SNR is defined as the ratio of the separation and standard deviation of the sub-Gaussian components. Combining these two results immediately leads to explicit bounds on the error of the SDP solution (Corollary 1).

When the SNR is sufficiently large, the above results imply that the SDP solution is integral and exact up to numerical errors, hence recovering (sometimes improving) existing results on exact recovery as a special case. Moreover, when the SNR is lower and the SDP solution is fractional, an explicit clustering can be obtained from the SDP solution via a simple, optimization-free rounding procedure. We show that the error of this explicit clustering (in terms of the fraction of points misclassified) also decays exponentially in the SNR (Theorem 3). As a consequence, we obtain sufficient conditions for misclassifying at most δ fraction of the points for any given $\delta \in [0, 1]$. Finally, we show that the SDP solutions also lead to an efficient estimator of the cluster *centers*, for which estimation error bounds are established (Theorem 4). Significantly, our results often match and sometimes improve upon state-of-the-art performance guarantees in settings for which known results exist, and lead to new guarantees in other less studied settings of SGMMs. Detailed discussion of these implications of our results and comparison with existing ones will be provided after we state our main theorems.

Paper Organization The remainder of the paper is organized as follows. In Section 2, we discuss related work on SGMMs and its special cases. In Section 3, we describe the problem setup for SGMMs and provide a summary of our clustering algorithms. In Section 4, we present our main results, discuss some of their consequences and compare them with existing results.

2. Related work

The study of SGMMs has a long history and is still an active area of research. Here we review the most relevant results with theoretical guarantees, with a focus on SDP relaxation methods.

Dasgupta (1999) is among the first to obtain performance guarantees for GMMs. Subsequent work has obtained improved guarantees, achieved by various algorithms including spectral methods. These results often establish sufficient conditions, in terms of the separation between the cluster centers (or equivalently the SNR), for achieving (near)-exact recovery of the cluster memberships. Vempala and Wang (2004) obtain one of the best results and require SNR $\gtrsim (k \ln n)^{1/4}$, which is later improved and extended by Achlioptas and McSherry (2005); Kumar and Kannan (2010); Awasthi and Sheffet (2012). We compare these results with ours in Section 4.

Expectation-Maximization (EM) and Lloyd's algorithms are among the most popular methods for GMMs. Despite their empirical effectiveness, non-asymptotic statistical guarantees are established only recently. In particular, convergence and center estimation error bounds for EM under GMMs with two components are derived in Balakrishnan et al. (2017); Klusowski and Brinda (2016), with extension to multiple components given in Yan et al. (2017). The work of Lu and Zhou (2016) provides a general convergence analysis for Lloyd's algorithm, which implies clustering and center estimation guarantees for random models including SGMMs. All these results assume that one has access to a sufficiently good initial solution, typically obtained by spectral methods. Recent breakthrough has been made by Daskalakis et al. (2016); Xu et al. (2016), who establish global convergence of randomly-initialized EM for GMMs with two symmetric components. Complementarily, Jin et al. (2016) show that EM may fail to converge under GMMs with $k \geq 3$ components due to the existence of bad local minima.

Most relevant to us are work on convex relaxation methods for GMMs and k-means/median problems, with SDP relaxations first considered in Peng and Xia (2005); Peng and Wei (2007). Thanks to convexity, these methods do not suffer from the issues of bad local minima faced by EM and Lloyd's, though it is far from trivial to round their (typically fractional) solutions into valid

clustering solutions with provable and quality guarantees. In this direction, Awasthi et al. (2015); Li et al. (2017) establish conditions for LP/SDP relaxations to achieve exact recovery. The work of Mixon et al. (2017) consider SDP relaxations as a denoising method, and prove error bounds for a form of approximate recovery. Robustness of SDP relaxations under a semi-random GMM is studied in Awasthi and Vijayaraghavan (2017). Most of these results are directly comparable to ours, and we discuss them in more details in Section 4 after presenting our main theorems.

Clustering problems under Stochastic Block Models (SBMs) have also witnessed fruitful progress on convex relaxation methods; see Abbe (2017) for a survey. Much work has been done on exact recovery guarantees for SDP relaxations of SBMs (Krivelevich and Vilenchik, 2006; Oymak and Hassibi, 2011; Amini and Levina, 2014; Ames and Vavasis, 2014; Chen et al., 2014). A more recent line of work establishes approximate recovery guarantees of the SDPs (Guédon and Vershynin, 2016; Montanari and Sen, 2016). Particularly relevant to us is the work by Fei and Chen (2017), who also establish exponentially decaying error bounds. Despite the apparent similarity in the forms of the error bounds, our results require very different analytical techniques, due to the fundamental difference between the geometric and probabilistic structures of SBMs and SGMMs; moreover, our results reveal the more subtle hidden integrality property of SDP relaxations, which we believe holds more broadly beyond specific models like SBMs and SGMMs.

3. Models and algorithms

In this section, we formally set up the clustering problem under SGMMs and describe our SDP relaxation approach.

3.1. Notations

We first introduce some notations. Vectors and matrices are denoted by bold letters such as $\mathbf{u}$ and $\mathbf{M}$. For a vector $\mathbf{u}$, we denote by u_i its i-th entry. For a matrix $\mathbf{M}$, $\mathrm{Tr}(\mathbf{M})$ denotes its trace, M_{ij} its (i,j)-th entry, $\mathrm{diag}\,(\mathbf{M})$ the vector of its diagonal entries, $\|\mathbf{M}\|_1 \coloneqq \sum_{i,j} M_{ij}$ its entry-wise ℓ_1 norm, $\mathbf{M}_{i\bullet}$ its i-th row and $\mathbf{M}_{\bullet j}$ its j-th column. We write $\mathbf{M} \succeq 0$ if $\mathbf{M}$ is symmetric positive semidefinite. The trace inner product between two matrices $\mathbf{M}$ and $\mathbf{Q}$ of the same dimension is denoted by $\langle \mathbf{M}, \mathbf{Q} \rangle \coloneqq \mathrm{Tr}(\mathbf{M}^\top \mathbf{Q})$. For a number a, $\mathbf{M} \geq a$ means $M_{ij} \geq a, \forall i, j$. We denote by $\mathbf{1}_m$ the all-one column vector of dimension m. For a positive integer i, let $[i] \coloneqq \{1, 2, \ldots, i\}$. For two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists a universal constant $C > 0$ such that $a_n \leq C b_n$ for all n, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Finally, $\|X\|_{\psi_2} \coloneqq \inf\left\{t > 0 : \mathbb{E}\exp\left(X^2/t^2\right) \leq 2\right\}$ denotes the sub-Gaussian norm of a random variable X, and X is called sub-Gaussian if $\|X\|_{\psi_2} < \infty$. Note that Normal and bounded random variables are sub-Gaussian.

3.2. Sub-Gaussian Mixture Model

We focus on Sub-Gaussian Mixture Models (SGMMs) with balanced clusters and isotropic components.

Model 1 (Sub-Gaussian Mixture Model) *Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ be k unknown cluster centers. We observe n random points in $\mathbb{R}^d$ of the form*

$$\mathbf{h}_i \coloneqq \boldsymbol{\mu}_{\sigma^*(i)} + \mathbf{g}_i, \qquad i \in [n]$$

where $\sigma^*(i) \in [k]$ is the unknown cluster label of the i-th point, and $\{\mathbf{g}_i\}$ are i.i.d. zero-mean random vectors such that each g_{ij} are i.i.d. with $\|g_{ij}\|_{\psi_2} = \tau$. We assume that the ground-truth clusters have equal sizes, that is, $|\{i \in [n] : \sigma^*(i) = a\}| = \frac{n}{k}$ for each $a \in [k]$.

Throughout the paper we assume $n \geq 4$ and $k \geq 2$ to avoid degeneracy. Let $\boldsymbol{\sigma}^* \in [k]^n$ be the vector of the true cluster labels, that is, its i-th coordinate is $\sigma_i^* \equiv \sigma^*(i)$ (we use them interchangeably throughout the paper.) This unknown true underlying clustering can be encoded by *cluster matrix* $\mathbf{Y}^* \in \{0,1\}^{n\times n}$ such that for each $i, j \in [n]$,

$$Y_{ij}^* = \begin{cases} 1 & \text{if } \sigma^*(i) = \sigma^*(j), \text{ i.e., points } i \text{ and } j \text{ are in the same cluster,} \\ 0 & \text{if } \sigma^*(i) = \sigma^*(j), \text{ i.e., points } i \text{ and } j \text{ are in different clusters,} \end{cases}$$

with the convention $Y_{ii}^* = 1, \forall i \in [n]$. The task is to estimate the underlying clustering $\mathbf{Y}^*$ given the observed data $\{\mathbf{h}_i : i \in [n]\}$. From the data one may compute the pairwise squared distance matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, defined as

$$A_{ij} = \|\mathbf{h}_i - \mathbf{h}_j\|_2^2, \qquad (i,j) \in [n] \times [n].$$

The separation of the centers of clusters a and b is denoted by $\Delta_{ab} := \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2$, and $\Delta := \min_{a\neq b\in[k]} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2$ is the minimum separation of the centers. Playing a crucial role in our results is the quantity

$$s := \frac{\Delta}{\tau}, \tag{1}$$

which is a measure of the SNR of an SGMM.

3.3. Semidefinite programming relaxation

We now describe our SDP relaxation for clustering SGMMs. To begin, note that any candidate clustering of n points into k clusters can be represented using an *assignment matrix* $\mathbf{F} \in \{0,1\}^{n\times k}$ where

$$F_{ia} = \begin{cases} 1 & \text{if point } i \text{ is assigned to cluster } a \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{F} := \left\{\mathbf{F} \in \{0,1\}^{n\times k} : \mathbf{F}\mathbf{1}_k = \mathbf{1}_n\right\}$ be the set of all possible assignment matrices. Given the points $\{\mathbf{h}_i\}$ to be clustered, a natural approach is to find a assignment $\mathbf{F}$ that minimizes the total within-cluster pairwise distance. This objective can be expressed as

$$\sum_{i,j} A_{ij}\, \mathbb{I}\{i \text{ and } j \text{ are assigned to the same cluster}\} = \sum_{i,j} A_{ij}(\mathbf{F}\mathbf{F}^\top)_{ij} = \left\langle \mathbf{F}\mathbf{F}^\top, \mathbf{A} \right\rangle.$$

Therefore, the approach described above is equivalent to solving the integer program (2) below:

$$\begin{aligned} \min_{\mathbf{F}} \ & \left\langle \mathbf{F}\mathbf{F}^\top, \mathbf{A} \right\rangle \\ \text{s.t. } & \mathbf{F} \in \mathcal{F} \\ & \mathbf{1}_n^\top \mathbf{F} = \frac{n}{k}\mathbf{1}_k^\top \end{aligned} \tag{2}$$

$$\begin{aligned} \min_{\mathbf{Y}} \ & \langle \mathbf{Y}, \mathbf{A} \rangle \\ \text{s.t. } & \mathbf{Y}\mathbf{1}_n = \frac{n}{k}\mathbf{1}_n \\ & \mathbf{Y} \succeq 0 \\ & \operatorname{diag}(\mathbf{Y}) = \mathbf{1}_n \\ & \mathbf{Y} \in \{0,1\}^{n\times n}; \operatorname{rank}(\mathbf{Y}) = k. \end{aligned} \tag{3}$$

In (2) the additional constraint $\mathbf{1}_n^\top \mathbf{F} = \frac{n}{k}\mathbf{1}_k^\top$ enforces that all k clusters have the same size $\frac{n}{k}$, as we are working with an SGMM whose true clusters are balanced. Under this balanced model, it is not hard to see that the program (2) is equivalent to the classical k-means formulation. With a change of variable $\mathbf{Y} = \mathbf{F}\mathbf{F}^\top$, we may lift the program (2) to the space of $n \times n$ matrices and obtain the equivalent formulation (3). Both programs (2) and (3) involve non-convex combinatorial constraints and are computationally hard to solve. To obtain a tractable formulation, we drop the non-convex rank constraint in (3) and replace the integer constraint with a linear constrain $0 \leq \mathbf{Y} \leq 1$ (the constraint $\mathbf{Y} \leq 1$ is redundant). This leads to the following SDP relaxation:

$$\begin{aligned} \widehat{\mathbf{Y}} \in \underset{\mathbf{Y}\in\mathbb{R}^{n\times n}}{\arg\min}\ & \langle \mathbf{Y}, \mathbf{A}\rangle \\ \text{s.t. } & \mathbf{Y}\mathbf{1}_n = \frac{n}{k}\mathbf{1}_n \\ & \mathbf{Y} \succeq 0 \\ & \operatorname{diag}(\mathbf{Y}) = \mathbf{1}_n \\ & \mathbf{Y} \geq 0. \end{aligned} \tag{4}$$

It is not hard to see that the true cluster matrix $\mathbf{Y}^*$ is feasible to program (4). We view any optimal solution $\widehat{\mathbf{Y}}$ to (4) as an estimate of the true clustering $\mathbf{Y}^*$. Our goal is to characterize the cluster recovery/estimation error $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$ in terms of the number of points n, number of clusters k, data dimension d and SNR s defined above. Note that here we measure the error of $\widehat{\mathbf{Y}}$ in ℓ_1 metric; as we shall see later, this metric is directly related to the clustering error (i.e., the fraction of misclassified points).

We remark that the SDP (4) is somewhat different from the more classical and well-known SDP relaxation of k-means proposed by Peng and Wei (2007). This SDP (4) is closely related to the one considered by Amini and Levina (2014) in the context of the Stochastic Block Model, though it seems to be much less studied under SGMMs with the notable exception of Li et al. (2017).

3.4. Explicit clustering

Our main results directly concern the SDP solution $\widehat{\mathbf{Y}}$, which is not integral in general and hence does not directly correspond to an explicit clustering. In case an explicit clustering is desired, we may easily extract cluster memberships from the solution $\widehat{\mathbf{Y}}$ using a simple procedure.

The procedure consists of two steps given as Algorithms 1 and 2, respectively. In the first step, we treat the rows of $\widehat{\mathbf{Y}}$ as elements of $\mathbb{R}^n$, and consider the ℓ_1 balls centered at each row with a certain radius. The ball that contains the most rows is identified, and the indices of the rows in this ball are output and removed. The process continues iteratively with the remaining rows of $\widehat{\mathbf{Y}}$. This step outputs a number of sets whose sizes are no larger than $\frac{n}{k}$ but may not equal to each other.

Algorithm 1 First step

Input: data matrix $\widehat{\mathbf{Y}} \in \mathbb{R}^{n \times n}$, size of each cluster $\frac{n}{k}$.

1. $B_0 \leftarrow \emptyset$, $t \leftarrow 0$, $V \leftarrow [n]$
2. While $V \setminus \bigcup_{i=0}^{t} B_i \neq \emptyset$:
   (a) $t \leftarrow t + 1$
   (b) $V_t \leftarrow V \setminus \bigcup_{i=0}^{t-1} B_i$
   (c) For each $u \in V_t$: $B(u) \leftarrow \left\{ w \in V_t : \|\widehat{\mathbf{Y}}_{u\bullet} - \widehat{\mathbf{Y}}_{w\bullet}\|_1 \leq \frac{n}{4k} \right\}$.
   (d) $B_t \leftarrow \arg\max_{B(u): u \in V_t} |B(u)|$
   (e) If $|B_t| > \frac{n}{k}$, then remove arbitrary elements in B_t so that $|B_t| = \frac{n}{k}$

Output: sets $\{B_t\}_{t \geq 1}$.

In the second step, we convert the sets output by Algorithm 1 above into k equal-size clusters. This is done by picking the k largest sets among them, and distributing points in the remaining sets across the chosen k sets so that each of the k sets contains exactly $\frac{n}{k}$ points.

Algorithm 2 Second step

Input: approximate clustering sets $\{B_t\}_{t \geq 1}$, number of points n, number of clusters to extract k.

1. $k' \leftarrow \left| \{B_t\}_{t \geq 1} \right|$
2. Choose k largest sets among $\{B_t\}_{t \geq 1}$ and rename the chosen sets as $\{U_t\}_{t \in [k]}$
3. Arbitrarily distribute elements of $\{B_t\}_{t \geq 1} \setminus \{U_t\}_{t \in [k]}$ among $\{U_t\}_{t \in [k]}$ so that each U_t has exactly $\frac{n}{k}$ elements
4. For each $i \in [n]$: $\widehat{\sigma}_i \leftarrow t$, where t is the unique index in $[k]$ such that $i \in U_t$

Output: clustering assignment vector $\widehat{\boldsymbol{\sigma}} \in [k]^n$.

Our final clustering algorithm, `cluster`, is a combination of the above two algorithms.

Algorithm 3 `cluster`

Input: data matrix $\widehat{\mathbf{Y}} \in \mathbb{R}^{n \times n}$, number of points n, number of clusters to extract k.

1. Run Algorithm 1 with $\widehat{\mathbf{Y}}$ and $\frac{n}{k}$ as input and get $\{B_t\}_{t \geq 1}$
2. Run Algorithm 2 with $\{B_t\}_{t \geq 1}$, n and k as input and get $\widehat{\boldsymbol{\sigma}}$

Output: clustering assignment $\widehat{\boldsymbol{\sigma}} \in [k]^n$.

The output

$$\widehat{\boldsymbol{\sigma}} := \texttt{cluster}(\widehat{\mathbf{Y}}, n, k)$$

is a vector in $[k]^n$ such that point i is assigned to the $\widehat{\sigma}_i$-th cluster. We are interested in controlling the clustering error of $\widehat{\boldsymbol{\sigma}}$ relative to the ground-truth clustering $\boldsymbol{\sigma}^*$. Let $\mathcal{S}_k$ denote the symmetric group consisting of all permutations of $[k]$. The clustering error is defined by

$$\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) := \min_{\pi \in \mathcal{S}_k} \frac{1}{n} \left| \{ i \in [n] : \widehat{\sigma}_i \neq \pi(\sigma_i^*) \} \right|, \tag{5}$$

which is the proportion of points that are misclassified, modulo permutations of the cluster labels.

Variants of the above `cluster` procedure have been considered before by Makarychev et al. (2016) and Mixon et al. (2017). In our main results, we show that the clustering error $\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$ is always upper bounded by the ℓ_1 error $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$ of the SDP solution $\widehat{\mathbf{Y}}$.

4. Main results

In this section, we establish the connection between the estimation error of the SDP relaxation (4) and that of what we call the Oracle Integer Program. Using this connection, we derive explicit bounds on the error of the SDP, and explore their implications for clustering and center estimation.

4.1. Oracle Integer Program

Consider an idealized setting where an oracle reveals the true cluster centers $\{\boldsymbol{\mu}_a\}_{a \in [k]}$. Moreover, we are given the data points $\{\bar{\mathbf{h}}_i\}_{i \in [n]}$, where $\bar{\mathbf{h}}_i := \boldsymbol{\mu}_{\sigma^*(i)} + (2c)^{-1}\mathbf{g}_i$ for $c := \frac{1}{8}$ and $\{\mathbf{g}_i\}$ are the same (realizations of the) random variables in the original SGMM. In other words, $\{\bar{\mathbf{h}}_i\}$ are the same as the original data points $\{\mathbf{h}_i\}$ generated by the SGMM, except that the standard deviation (or more generally, the sub-Gaussian norm) of noise $\{\mathbf{g}_i\}$ is scaled by $(2c)^{-1} = 4$. To cluster $\{\bar{\mathbf{h}}_i\}$ in this idealized setting, a natural approach is to simply assign each point to the closest cluster center, so that the total distance of the points to their assigned centers are minimized. We may formulate this procedure as an integer program, by representing each candidate clustering assignment using an assignment matrix $\mathbf{F} \in \mathcal{F}$ as before. Then, for each assignment matrix $\mathbf{F}$, the quantity

$$\eta(\mathbf{F}) := \sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja}$$

is exactly the sum of the distances of each point to its assigned cluster center. The clustering procedure above thus amounts to solving the following "Oracle Integer Program (IP)":

$$\min_{\mathbf{F}} \eta(\mathbf{F}), \qquad \text{s.t. } \mathbf{F} \in \mathcal{F}. \tag{6}$$

Let $\mathbf{F}^* \in \mathcal{F}$ be the assignment matrix associated with the true underlying clustering of the SGMM; that is, $F^*_{ja} = \mathbb{I}\{\sigma^*(j) = a\}$ for each $j \in [n], a \in [k]$. For each assignment $\mathbf{F} \in \mathcal{F}$, it is easy to see that the quantity $\frac{1}{2}\|\mathbf{F} - \mathbf{F}^*\|_1$ is exactly the number of nodes that are assigned differently in $\mathbf{F}$ and $\mathbf{F}^*$, and hence measures the clustering error of $\mathbf{F}$ with respect to the ground truth $\mathbf{F}^*$.

A priori, there is no obvious connection between the estimation error of a solution to the above Oracle IP and that of a solution to the SDP. In particular, the latter involves a continuous relaxation whose solutions are fractional in general, and the true centers are unknown therein. Surprisingly, we are able to establish a formal connection between the two, and in particular show that the error of the SDP is bounded by the error of the IP in an appropriate sense.

4.2. Errors of SDP relaxation and Oracle IP

To establish the connection, we begin with the following observation: for a solution $\mathbf{F} \in \mathcal{F}$ to potentially be an optimal solution of the Oracle IP (6), it must satisfy $\eta(\mathbf{F}) \leq \eta(\mathbf{F}^*)$ since $\mathbf{F}^*$ is feasible to (6). Consequently, the quantity

$$\max \left\{ \frac{1}{2}\|\mathbf{F} - \mathbf{F}^*\|_1 : \mathbf{F} \in \mathcal{F}, \eta(\mathbf{F}) \leq \eta(\mathbf{F}^*) \right\} \tag{7}$$

is the worst-case error of a potentially optimal solution to the Oracle IP. This quantity turns out to be an *upper* bound of the error of any optimal solution $\widehat{\mathbf{Y}}$ to the SDP relaxation, as is shown in the theorem below.

Theorem 1 (IP bounds SDP) *Under Model 1, there exist some universal constants $C_s > 0, C \geq 1$ for which the following holds. If the SNR satisfies*

$$s^2 \geq C_s \left(\sqrt{\frac{kd}{n} \log n} + k\sqrt{\frac{d}{n}} + k \right), \tag{8}$$

then we have

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \leq 2 \cdot \max \left\{ \frac{\|\mathbf{F} - \mathbf{F}^*\|_1}{\|\mathbf{F}^*\|_1} : \eta(\mathbf{F}) \leq \eta(\mathbf{F}^*), \mathbf{F} \in \mathcal{F} \right\}$$

with probability at least $1 - n^{-C} - 2e^{-n}$.

The proof is given in Section B, and consists of two main steps: (i) showing that with high probability the SDP error is upper bounded by the objective value of a linear program (LP), and (ii) showing that the LP admits an *integral* optimal solution and relating this solution to the quantity (7). We note that the key step (ii), which involves establishing certain hidden integrality properties, is completely deterministic. The SNR condition (8) is required only in the probabilistic step (i); therefore, sharper analysis in step (i) will lead to potentially more relaxed conditions on the SNR.

To obtain an explicit bound on the SDP error, it suffices to upper bound the error of the Oracle IP. This turns out to be a relatively easy task compared to directly controlling the error of the SDP. The reason is that the Oracle IP has only finitely many feasible solutions, allowing one to use a union-bound-like argument. In particular, our analysis establishes that the error of Oracle IP decays *exponentially* in the SNR, as summarized in the theorem below.

Theorem 2 (Exponential rates of IP) *Under Model 1, there exist universal constants $C_s, C_g, C_e > 0$ for which the following holds. If $s^2 \geq C_s k$, then we have*

$$\max \left\{ \frac{\|\mathbf{F} - \mathbf{F}^*\|_1}{\|\mathbf{F}^*\|_1} : \eta(\mathbf{F}) \leq \eta(\mathbf{F}^*), \mathbf{F} \in \mathcal{F} \right\} \leq C_g \exp\left[-\frac{s^2}{C_e} \right]$$

with probability at least $1 - \frac{3}{2}n^{-1}$.

The proof is given in Section C. An immediate consequence of Theorems 1 and 2 is that the SDP (4) also achieves an exponentially decaying error rate.

Corollary 1 (Exponential rates of SDP) *Under the SNR condition (8), there exist universal constants $C_m, C_e > 0$ such that*

$$\frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1} \leq C_m \exp\left[-\frac{s^2}{C_e}\right]$$

with probability at least $1 - 2n^{-1}$.

Our next result concerns the explicit clustering $\widehat{\boldsymbol{\sigma}}$ extracted from $\widehat{\mathbf{Y}}$ using the procedure described in Section 3.4. In particular, we show that the number of misclassified points is upper bounded by the error in $\widehat{\mathbf{Y}}$ and hence also exhibits an exponential decay.

Theorem 3 (Clustering error) *The error rate in $\widehat{\boldsymbol{\sigma}}$ is always upper bounded by the error in $\widehat{\mathbf{Y}}$:*

$$\mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) \lesssim \frac{\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1}{\|\mathbf{Y}^*\|_1}.$$

Consequently, under the SNR condition (8), there exist universal constants $C_m, C_e > 0$ such that

$$\mathtt{err}\left(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*\right) \leq C_m \exp\left[-\frac{s^2}{C_e}\right]$$

with probability at least $1 - 2n^{-1}$.

The proof is given in Section E. Note that the above bound in terms of the *clustering error* is optimal (up to a constant in the exponent) in view of the minimax results in Lu and Zhou (2016).

4.3. Consequences

We explore the consequences of our error bounds in Corollary 1 and Theorem 3.

- **Exact recovery:** If the SNR s^2 satisfies the condition (8) and moreover $s^2 \gtrsim \log n$, then Theorem 3 guarantees that $\mathtt{err}\left(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*\right) < \frac{1}{n}$, which means that $\mathtt{err}\left(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*\right) = 0$ and the true underlying clustering is recovered exactly. Note that these conditions can be simplified to $s^2 \gtrsim k + \log n$ when $n \gtrsim d$. In fact, by Corollary 1 we know that the SDP solution satisfies the bound $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 < \frac{1}{4}$ in this case, so simply rounding $\widehat{\mathbf{Y}}$ *element-wise* produces the ground-truth cluster matrix $\mathbf{Y}^*$. Therefore, the SDP relaxation is able to achieve exact recovery (sometimes called *strong consistency* in the literature on SBM (Abbe, 2017)) of the underlying clusters when the SNR is sufficiently large.

In fact, our results apply even in regimes with a lower SNR, for which exact recovery of the clusters is impossible due to potential overlap between points from different clusters. In such regimes, Corollary 1 and Theorem 3 imply *approximate recovery* guarantees for the SDP relaxation:

- **Almost exact recovery:** If s^2 satisfies the condition (8) and $s^2 = \omega(1)$, then Theorem 3 implies that $\mathtt{err}\left(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*\right) = o(1)$. That is, the SDP recovers asymptotically the cluster memberships of almost all points, which is sometimes called *weak consistency.*

- **Recovery with δ-error:** More generally, for any number $\delta \in (0, 1)$, Theorem 3 implies the following non-asymptotic recovery guarantee: If s^2 satisfies the condition (8) and $s^2 \gtrsim \log \frac{1}{\delta}$, then $\mathtt{err}\left(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*\right) \leq \delta$. That is, the SDP correctly recovers the cluster memberships of at least $(1 - \delta)$ fraction of the points.

We compare the above results with existing ones in Section 4.4 to follow.

Cluster center estimation: We may obtain an estimate of the cluster *centers* using estimated cluster labels $\widehat{\boldsymbol{\sigma}}$ produced by the SDP relaxation. In particular, we simply compute the empirical means of the points within each estimated clusters; that is,

$$\hat{\boldsymbol{\mu}}_a := \frac{k}{n} \sum_{i:\widehat{\sigma}_i = a} \mathbf{h}_i$$

for each $a \in [k]$. As a corollary of our bounds on clustering errors, we obtain the following guarantee on center estimation.

Theorem 4 (Cluster center estimation error) *Suppose that* $\max_{a,b\in[k]} \Delta_{ab} \le C_q \Delta$ *for some universal constant* $C_q > 0$. *Under the same conditions of Theorem 1, there exist universal constants* $C_m, C_e > 0$ *such that*

$$\max_{a\in[k]} \min_{\pi\in\mathcal{S}_k} \|\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_{\pi(a)}\|_2 \le C_m \tau \left(\sqrt{\frac{kd + \log n}{n}} + \left(\sqrt{d + \log n}\right) \cdot \exp\left[-\frac{s^2}{C_e}\right] \right)$$

with probability at least $1 - 3n^{-1}$.

The proof is given in Section F. Note that the error is measured again up to permutation of the cluster labels. Our error bound consists of two terms. The first term, $\tau\sqrt{\frac{kd+\log n}{n}}$, is the error of estimating a d-dimensional cluster center vector using the $\frac{n}{k}$ data points (with standard deviation τ) from that cluster. This term is unavoidable even when the true cluster labels are known. On the other hand, the second term captures the error due to incorrect cluster labels for some of the points. When $s^2 \gtrsim \log n$ and $d \gtrsim \log n$, we achieve the minimax optimal rate $\tau\sqrt{\frac{d}{n/k}}$ for center estimation.

4.4. Comparison with existing results

Table 1 summarizes several most representative results in the literature on clustering SGMM/GMM. Most of them are in terms of SNR conditions required to achieve exact recovery of the underlying clusters. Note that our results imply sufficient conditions for *both* exact and approximate recovery.

Most relevant to us is the work of Li et al. (2017), which considers similar SDP relaxation formulations. They show that exact recovery is achieved when $s^2 \gtrsim k + \log n$ and $n \gg d^2 k^3 \log k$. In comparison, a special case of our Corollary 1 guarantees exact recovery whenever $s^2 \gtrsim k + \log n$ and $n \gtrsim d$, which is milder then the condition in Li et al. (2017).

The work in Lu and Zhou (2016) also proves an exponentially decaying clustering error rate, but for a different algorithm (Lloyd's algorithm). To achieve non-trivial approximate recovery of the clusters, they require $s^2 \gg k^2 + k^3 \frac{d}{n}$ and $k^3 \ll \frac{n}{\log n}$ as $n \to \infty$. Our SNR condition in (8) has milder dependency on k, though dependency on n and d are a bit more subtle. We do note that under their more restricted SNR condition, Lu and Zhou (2016) are able to obtain tight constants in the exponent of the error rate.

Finally, the work of Mixon et al. (2017) considers the SDP relaxation introduced by Peng and Wei (2007) and provides bounds on center estimation when $s^2 \gtrsim k^2$. An intermediate result of theirs concerns errors of the SDP solutions; under the setting of balanced clusters, their error bound can be compared with ours after appropriate rescaling. In particular, their result implies the error

bound $\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \lesssim \frac{n^2}{s^2}$ when n is sufficiently large. This bound is non-trivial when $s^2 \gtrsim k$ since $\|\mathbf{Y}^*\|_F^2 = \frac{n^2}{k}$. Under the same conditions on s^2 and n, our results imply the exponential error bound

$$\|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \leq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \lesssim \frac{n^2}{k} e^{-s^2},$$

which is strictly better.

To sum up, corollaries of our results provide more relaxed conditions for exact or approximate recovery compared to most of the existing results listed in Table 1. Our results are weaker by a $\sqrt{k}$ factor than the one in Vempala and Wang (2004), which focuses on exact recovery under spherical Gaussian mixtures; on the other hand, our results apply to the more general sub-Gaussian setting, and imply approximate recovery guarantees under more general SNR conditions.

Paper	Condition on SNR^2	Recovery Type	Algorithm
Vempala and Wang (2004)	$\Omega\left(\sqrt{k \log n}\right)$	Exact	Spectral
Achlioptas and McSherry (2005)	$\Omega\left(k \log n + k^2\right)$	Exact	Spectral
Kumar and Kannan (2010)	$\Omega\left(k^2 \cdot \text{polylog}\left(n\right)\right)$	Exact	Spectral
Awasthi and Sheffet (2012)	$\Omega\left(k \cdot \text{polylog}\left(n\right)\right)$	Exact	Spectral
Lu and Zhou (2016)	$\Omega\left(k^2\right)$	Approximate	Spectral +Lloyd's
	$\Omega\left(k^2 + \log n\right)$	Exact	
Mixon et al. (2017)	$\Omega\left(k^2\right)$	For center estimation	SDP
Li et al. (2017)	$\Omega\left(k + \log n\right)$	Exact	SDP
Ours	$\Omega\left(k\right)$	Approximate	SDP
	$\Omega\left(k + \log n\right)$	Exact	

Table 1: Summary of existing results on cluster recovery for GMM. Here "approximate" means correct recovery of the memberships of at least $(1 - \delta)$ fraction of the points for a fixed constant $\delta \in (0, 1)$. Some of the results listed assume that $n \gg \text{poly}(k, d)$; see Section 4.4 for details.

5. Conclusion

In this paper, we have considered clustering problems under SGMMs using an SDP relaxation. We have shown that the SDP performs at least as well as an idealized IP, which achieves an exponentially decaying error rate. As a by-product of our analysis, we have obtained an error bound for estimating mixture centers via the SDP.

Our work points to several interesting future directions. An immediate problem is extending our results to the case of imbalanced clusters and non-isotropic distributions. It is also of interest to study the robustness of SDP relaxations for SGMMs by considering adversarial attacks or arbitrary outliers in the generated data under various semi-random models (Awasthi and Vijayaraghavan, 2017). Other directions that are worth exploring include obtaining better constants in error bounds, identifying sharp thresholds for different types of recovery, and obtaining tight localized proximity conditions in the lines of Li et al. (2017).

Acknowledgments

Y. Fei and Y. Chen were partially supported by the National Science Foundation CRII award 1657420 and grant 1704828.

References

Emmanuel Abbe. Community detection and the stochastic block model: recent developments. *Journal of Machine Learning Research, to appear*, 2017. URL http://www.princeton.edu/~eabbe/publications/sbm_jmlr_4.pdf.

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Brendan P. W. Ames and Stephen A. Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1-2):299–337, 2014.

Arash A. Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*, 2014.

Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. *arXiv preprint arXiv:1206.3204*, 2012.

Pranjal Awasthi and Aravindan Vijayaraghavan. Clustering semi-random mixtures of gaussians. *arXiv preprint arXiv:1711.08841*, 2017.

Pranjal Awasthi, Afonso S. Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200. ACM, 2015.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 1–10. ACM, 1999.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science*, pages 634–644. IEEE, 1999.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. *arXiv preprint arXiv:1609.00368*, 2016.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *arXiv preprint arXiv:1705.08391*, 2017.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

Takayuki Iguchi, Dustin G. Mixon, Jesse Peterson, and Soledad Villar. Probably certifiably correct k-means clustering. *Mathematical Programming*, 165(2):605–642, 2017.

Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, pages 731–740. ACM, 2002.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

Jason M. Klusowski and W. D. Brinda. Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. *arXiv preprint arXiv:1608.02280*, 2016.

Michael Krivelevich and Dan Vilenchik. Semirandom models as benchmarks for coloring algorithms. In *Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 211–221, 2006.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE Computer Society, 2010.

Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM Journal on Computing*, 45(2):530–547, 2016.

Xiaodong Li, Yang Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. When do birds of a feather flock together? K-means, proximity, and conic programming. *arXiv preprint arXiv:1710.06008*, 2017.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Yu Lu and Harrison H. Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. In *International Workshop on Algorithms and Computation*, pages 274–285. Springer, 2009.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Learning communities in the presence of errors. In *29th Annual Conference on Learning Theory*, pages 1258–1291, 2016.

Dustin G. Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, page iax001, 2017.

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 814–827, Cambridge, MA, USA, June 2016. doi: 10.1145/2897518.2897548. URL http://doi.acm.org/10.1145/2897518.2897548.

Abhinav Nellore and Rachel Ward. Recovery guarantees for exemplar-based clustering. *Information and Computation*, 245:165–180, 2015.

Samet Oymak and Babak Hassibi. Finding dense clusters via "low rank+ sparse" decomposition. *arXiv preprint arXiv:1104.5186*, 2011.

Karl Pearson. Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2):34–59, 1936.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

Jiming Peng and Yu Xia. A new theoretical framework for k-means-type clustering. In *Foundations and Advances in Data Mining*, pages 79–96. Springer, 2005.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

Roman Vershynin. High-dimensional probability. 2017.

Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence analysis of gradient EM for multi-component gaussian mixture. *arXiv preprint arXiv:1705.08530*, 2017.

Appendix A. Additional notations

We define the shorthand $\gamma := \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1$. For a matrix $\mathbf{M}$, we write $\|\mathbf{M}\|_\infty := \max_{i,j} |M_{ij}|$ as its entry-wise ℓ_∞ norm, and $\|\mathbf{M}\|_{\text{op}}$ as its spectral norm (maximum singular value). We let $\mathbf{I}$ and $\mathbf{J}$ be the $n \times n$ identity matrix and all-one matrix, respectively. For a real number x, $\lceil x \rceil$ denotes its ceiling. We denote by $C_a^* := \{i \in [n] : \sigma^*(i) = a\}$ the set of indices of points in cluster a, and we define $\ell := |C_a^*| = \frac{n}{k}$.

Appendix B. Proof of Theorem 1

B.1. Initial steps

We assume $\gamma > 0$ since otherwise we are done. We can write $\mathbf{A} = \mathbf{C} + \mathbf{C}^\top - 2\mathbf{H}\mathbf{H}^\top$, where $\mathbf{H}$ is a matrix whose i-th row is the point $\mathbf{h}_i$ and $\mathbf{C}$ is a matrix where the entries in the i-th row are identical and equal to $\|\mathbf{h}_i\|_2^2$. Since the row-sum constraint in the program (4) ensures that the matrix $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ has zero row sum, we have $\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{C} \right\rangle = \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{C}^\top \right\rangle = 0$ which implies $\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{C} + \mathbf{C}^\top \right\rangle = 0$.

Let $\mathbf{G} \coloneqq \mathbf{H} - \mathbb{E}\mathbf{H}$ be a matrix of entries in $\mathbf{H}$ with their means removed. We can compute

$$\begin{aligned}\mathbf{H}\mathbf{H}^\top &= (\mathbf{G} + \mathbb{E}\mathbf{H})(\mathbf{G} + \mathbb{E}\mathbf{H})^\top \\ &= \mathbf{G}\mathbf{G}^\top + \mathbf{G}(\mathbb{E}\mathbf{H})^\top + (\mathbb{E}\mathbf{H})\mathbf{G}^\top + (\mathbb{E}\mathbf{H})(\mathbb{E}\mathbf{H})^\top\end{aligned}$$

and

$$\mathbb{E}\mathbf{H}\mathbf{H}^\top = \mathbb{E}\mathbf{G}\mathbf{G}^\top + (\mathbb{E}\mathbf{H})(\mathbb{E}\mathbf{H})^\top.$$

Therefore

$$\mathbf{H}\mathbf{H}^\top - \mathbb{E}\mathbf{H}\mathbf{H}^\top = \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right) + \mathbf{G}(\mathbb{E}\mathbf{H})^\top + (\mathbb{E}\mathbf{H})\mathbf{G}^\top.$$

Let $\mathbf{U} \in \mathbb{R}^{n\times k}$ be the matrix of the left singular vectors of $\mathbf{Y}^*$. For any $\mathbf{M} \in \mathbb{R}^{n\times n}$, define the projection $\mathcal{P}_T(\mathbf{M}) \coloneqq \mathbf{U}\mathbf{U}^\top\mathbf{M} + \mathbf{M}\mathbf{U}\mathbf{U}^\top - \mathbf{U}\mathbf{U}^\top\mathbf{M}\mathbf{U}\mathbf{U}^\top$ and its orthogonal complement $\mathcal{P}_{T^\perp}(\mathbf{M}) \coloneqq \mathbf{M} - \mathcal{P}_T(\mathbf{M})$. The fact that $\widehat{\mathbf{Y}}$ is optimal and $\mathbf{Y}^*$ is feasible to the program (4) implies

$$\begin{aligned}0 &\le -\frac{1}{2}\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{A} \right\rangle \\ &= \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{H}\mathbf{H}^\top - \mathbb{E}\mathbf{H}\mathbf{H}^\top \right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbb{E}\mathbf{H}\mathbf{H}^\top \right\rangle \\ &= \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top + \mathbf{G}(\mathbb{E}\mathbf{H})^\top + (\mathbb{E}\mathbf{H})\mathbf{G}^\top \right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbb{E}\mathbf{H}\mathbf{H}^\top \right\rangle \\ &= \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathcal{P}_T\left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right) \right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathcal{P}_{T^\perp}\left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right) \right\rangle \\ &\quad + 2\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{G}(\mathbb{E}\mathbf{H})^\top \right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbb{E}\mathbf{H}\mathbf{H}^\top \right\rangle \\ &=: S_1 + S_2 + 2S_3 + S_4.\end{aligned}$$

We may control S_1, S_2 and S_4 using the following.

Proposition 1 *If* $s^2 \ge C\left(\sqrt{\frac{kd}{n}\log(nk)} + \sqrt{\frac{k}{n}}\log(nk)\right)$ *for some universal constant* $C > 0$, *then* $S_1 \le \frac{1}{100}\Delta^2\gamma$ *with probability at least* $1 - (2n)^{-2}$.

Proposition 2 *If* $s^2 \ge Ck\left(\sqrt{\frac{d}{n}} + 1\right)$ *for some universal constant* $C > 0$, *then* $S_2 \le \frac{1}{100}\Delta^2\gamma$ *with probability at least* $1 - 2e^{-n}$.

Proposition 3 *We have* $S_4 = -\frac{1}{2}\sum_{a\neq b} T_{ab}\Delta_{ab}^2 \le -\frac{1}{4}\Delta^2\gamma$ *where* $T_{ab} \coloneqq \sum_{i\in C_a^*, j\in C_b^*}\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)_{ij}$.

The proofs are given in Sections B.4, B.5 and B.6, respectively. Combining the above propositions, we have $S_1 + S_2 \le -\frac{1}{2}S_4$ and therefore

$$0 \le S_3 + \frac{1}{4}S_4 =: S_0 \tag{9}$$

with probability at least $1 - (2n)^{-C'} - 2e^{-n}$ for some universal constant $C' > 0$.

Let $\mathbf{B} := \widehat{\mathbf{Y}} - \mathbf{Y}^*$. We have

$$\begin{aligned} S_3 &= \sum_j \sum_a \sum_{i \in C_a} B_{ji} \langle \boldsymbol{\mu}_a, \mathbf{g}_j \rangle \\ &= \ell \sum_j \sum_a \langle \boldsymbol{\mu}_a, \mathbf{g}_j \rangle \left(\frac{1}{\ell} \sum_{i \in C_a^*} B_{ji} \right) \\ &= \ell \sum_j \sum_{a \neq \sigma^*(j)} \left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, \mathbf{g}_j \right\rangle \left(\frac{1}{\ell} \sum_{i \in C_a^*} B_{ji} \right) \end{aligned}$$

where the last step holds since $\sum_{a \neq \sigma^*(j)} \left(\sum_{i \in C_a^*} B_{ji} \right) = - \sum_{i \in C_a^*: a = \sigma^*(j)} B_{ji}$ for each $j \in [n]$ which follows from the row-sum constraint of program (4). By Proposition 3, we have

$$S_4 = -\ell \sum_j \sum_{a \neq \sigma^*(j)} \frac{1}{2} \Delta_{\sigma^*(j),a}^2 \left(\frac{1}{\ell} \sum_{i \in C_a^*} B_{ji} \right).$$

Therefore, we have

$$S_0 = \ell \sum_j \sum_{a \neq \sigma^*(j)} \left(\left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, \mathbf{g}_j \right\rangle - c\Delta_{\sigma^*(j),a}^2 \right) \left(\frac{1}{\ell} \sum_{i \in C_a^*} B_{ji} \right)$$

where $c = \frac{1}{8}$.

To control S_0, we define $\beta_{ja} := \left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, \mathbf{g}_j \right\rangle - c\Delta_{\sigma^*(j),a}^2$ and consider the program

$$\begin{aligned} \max_{\mathbf{X}} \ & \sum_j \sum_{a \neq \sigma^*(j)} \beta_{ja} X_{ja} \\ \text{s.t. } & 0 \le X_{ja} \le 1, \qquad \forall a \neq \sigma^*(j), j \in [n] \\ & \sum_{a \neq \sigma^*(j)} X_{ja} \le 1, \qquad \forall j \in [n] \\ & \sum_j \sum_{a \neq \sigma^*(j)} X_{ja} = R, \end{aligned} \tag{10}$$

where $R \in (0, n]$. Let us denote by $V(R)$ the optimal value of the above program and we let $V(R) = -\infty$ if the program is infeasible. The constraints of program (4) implies that $\frac{\gamma}{2\ell} \in (0, n]$ and

$$\sum_{j \in [n]} \sum_{a \neq \sigma^*(j)} \left(\sum_{i \in C_a^*} B_{ji} \right) = \frac{\gamma}{2}.$$

Hence, by Equation (9), we have

$$0 \le S_0 \le \ell \cdot V\left(\frac{\gamma}{2\ell}\right). \tag{11}$$

B.2. Controlling γ by LP

We show that γ is upper bounded by the objective value of an LP that is related to program (10). If $\gamma = 0$ then the conclusion of Theorem 1 holds trivially. For $\gamma > 0$, we have the following cases:

1. If $\frac{\gamma}{2\ell} \in (0, 1]$, it follows from Equation (11) that the error γ must satisfy

   $$0 \le V\left(\frac{\gamma}{2\ell}\right) = \beta^* \frac{\gamma}{2\ell} \le \beta^* \left\lceil \frac{\gamma}{2\ell} \right\rceil = V\left(\left\lceil \frac{\gamma}{2\ell} \right\rceil\right)$$

   where $\beta^* \coloneqq \max_{j\in[n], a\neq\sigma^*(j)} \beta_{ja}$. This implies

   $$\frac{\gamma}{2\ell} \le \left\lceil \frac{\gamma}{2\ell} \right\rceil \le \max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}.$$

2. If $\frac{\gamma}{2\ell} > 1$, it follows from Equation (11) that the error γ must satisfy

   $$0 \le V\left(\frac{\gamma}{2\ell}\right) \le \max\left\{V\left(\left\lceil \frac{\gamma}{2\ell} \right\rceil\right), V\left(\left\lfloor \frac{\gamma}{2\ell} \right\rfloor\right)\right\} = \max\left\{V\left(\left\lceil \frac{\gamma}{2\ell} \right\rceil\right), V\left(\left\lceil \frac{\gamma}{2\ell} \right\rceil - 1\right)\right\}.$$

   In other words, we have

   $$\begin{aligned}\frac{\gamma}{2\ell} \le \left\lceil \frac{\gamma}{2\ell} \right\rceil &\le \max\left\{R \in \{0, 1, \ldots.\} : V(R) \vee V(R-1) \ge 0\right\} \\ &= 1 + \max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}.\end{aligned}$$

   Note that $\left\lceil \frac{\gamma}{2\ell} \right\rceil \ge 2$, and therefore we must have $1 \le \max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}$. This implies

   $$\frac{\gamma}{2\ell} \le 2\max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}.$$

Consequently, we have

$$\frac{\gamma}{2\ell} \le 2\max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}.$$

B.3. Converting LP to IP

We are now ready to formally establish a connection between the error of the SDP (4) and that of the Oracle IP (6), by relating $\max\left\{R \in \{0, 1, \ldots.\} : V(R) \ge 0\right\}$ to the quantity (7). Note that if $R \ge 0$ is an integer, then there exists an optimal solution $\{w_{ja}\}$ of program (10) such that $w_{ja} \in \{0, 1\}$ for all $j \in [n], a \in [k]$. Therefore, if $R \in \{0, 1, \ldots\}$ is an integer, then

$$V(R) = \texttt{IP}_1(R) \coloneqq \left\{ \begin{array}{ll} \max\limits_{\mathbf{X}} \sum\limits_{j} \sum\limits_{a\neq\sigma^*(j)} \beta_{ja} X_{ja} & \\ \text{s.t. } X_{ja} \in \{0, 1\}, & \forall a \neq \sigma^*(j), j \in [n] \\ \sum\limits_{a\neq\sigma^*(j)} X_{ja} \le 1, & \forall j \in [n] \\ \sum\limits_{j} \sum\limits_{a\neq\sigma^*(j)} X_{ja} = R & \end{array} \right\}. \tag{12}$$

Combining the last two display equations we obtain that

$$\begin{aligned}
\frac{\gamma}{2\ell} &\le 2\max\{R \in \{0,1,\ldots.\} : \texttt{IP}_1(R) \ge 0\} \\
&= 2 \cdot \left\{ \begin{array}{ll} \max\limits_{R,\mathbf{X}} R & \\ \text{s.t. } R \in \{0,1,\ldots\} & \\ \quad \sum\limits_j \sum\limits_{a \neq \sigma^*(j)} \beta_{ja} X_{ja} \ge 0 & \\ \quad X_{ja} \in \{0,1\}, & \forall a \neq \sigma^*(j), j \in [n] \\ \quad \sum\limits_{a \neq \sigma^*(j)} X_{ja} \le 1, & \forall j \in [n] \\ \quad \sum\limits_j \sum\limits_{a \neq \sigma^*(j)} X_{ja} = R, & \end{array} \right\} \\
&= 2 \cdot \texttt{IP}_2 \coloneqq 2 \cdot \left\{ \begin{array}{ll} \max\limits_{\mathbf{X}} \sum\limits_j \sum\limits_{a \neq \sigma^*(j)} X_{ja} & \\ \text{s.t. } \sum\limits_j \sum\limits_{a \neq \sigma^*(j)} \beta_{ja} X_{ja} \ge 0 & \\ \quad X_{ja} \in \{0,1\}, & \forall a \neq \sigma^*(j), j \in [n] \\ \quad \sum\limits_{a \neq \sigma^*(j)} X_{ja} \le 1, & \forall j \in [n] \end{array} \right\}. \qquad (13)
\end{aligned}$$

Let us reparameterize the integer program $\texttt{IP}_2$ by a change of variable. Recall that

$$\mathcal{F} \coloneqq \left\{ \mathbf{F} \in \{0,1\}^{n \times k} : \mathbf{F}\mathbf{1}_k = \mathbf{1}_n \right\}$$

is the set of all possible assignment matrices and $\mathbf{F}^* \in \mathcal{F}$ is the true assignment matrix; that is, $F^*_{ja} = \mathbb{I}\{a = \sigma^*(j)\}$ for all $j \in [n], a \in [k]$. Consider any feasible solution $\mathbf{X}$ of $\texttt{IP}_2$; here for each $j \in [n]$, we may fix $X_{j,\sigma^*(j)} = -\sum_{a \neq \sigma^*(j)} X_{ja}$ — doing so does not affect the feasibility and objective value of $\mathbf{X}$ w.r.t. $\texttt{IP}_2$. Define the new variable $\mathbf{F} \coloneqq \mathbf{F}^* + \mathbf{X} \in \mathcal{F}$. The objective value and constraints of the old variable $\mathbf{X}$ can be mapped to those of $\mathbf{F}$; in particular, we have

$$\sum_j \sum_{a \neq \sigma^*(j)} X_{ja} = \sum_j \sum_{a \neq \sigma^*(j)} (F_{ja} - F^*_{ja}) = \frac{1}{2}\|\mathbf{F} - \mathbf{F}^*\|_1$$

and

$$\left. \begin{array}{c} X_{ja} \in \{0,1\}, \forall a \neq \sigma^*(j), j \in [n] \\ \sum_{a \neq \sigma^*(j)} X_{ja} \le 1, \forall j \in [n] \\ X_{j,\sigma^*(j)} = -\sum_{a \neq \sigma^*(j)} X_{ja}, \forall j \in [n] \end{array} \right\} \Longleftrightarrow \mathbf{F} \in \mathcal{F}$$

and

$$\sum_j \sum_{a \neq \sigma^*(j)} \beta_{ja} X_{ja} \overset{(i)}{=} \sum_j \sum_a \beta_{ja} X_{ja} = \sum_j \sum_a \beta_{ja} F_{ja} - \sum_j \sum_a \beta_{ja} F^*_{ja} \overset{(ii)}{=} \sum_j \sum_a \beta_{ja} F_{ja},$$

where steps (i) and (ii) both follow from the fact that $\beta_{j,\sigma^*(j)} = 0, \forall j$. It follows that $\texttt{IP}_2$ has the same optimal value as a corresponding integer program in terms of $\mathbf{X}$; in particular, we have

$$\texttt{IP}_2 = \texttt{IP}_3 \coloneqq \left\{ \begin{array}{l} \max\limits_{\mathbf{F}} \dfrac{1}{2}\|\mathbf{F} - \mathbf{F}^*\|_1 \\ \text{s.t.} \sum\limits_j \sum\limits_a \beta_{ja} F_{ja} \geq 0 \\ \quad \mathbf{F} \in \mathcal{F} \end{array} \right\}.$$

Combining with equation (13), we see that the error γ satisfies

$$\frac{\gamma}{2\ell} \leq 2 \cdot \texttt{IP}_3. \tag{14}$$

We further simplify the first constraint in $\texttt{IP}_3$. Recall that $\bar{\mathbf{h}}_i \coloneqq \boldsymbol{\mu}_{\sigma^*(i)} + (2c)^{-1}\mathbf{g}_i$ for each $i \in [n]$. Note that $\{\bar{\mathbf{h}}_i\}$ can be viewed as data points generated from the Sub-Gaussian Mixture Model but with $(2c)^{-1}$ times the standard deviation. By definition of β_{ja}, we have

$$\begin{aligned}
\beta_{ja} &= \left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, \mathbf{g}_j \right\rangle - c\Delta^2_{\sigma^*(j),a} \\
&= c\left(2\left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, (2c)^{-1}\mathbf{g}_j \right\rangle - \Delta^2_{\sigma^*(j),a}\right) \\
&= c\left(2\left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, (2c)^{-1}\mathbf{g}_j \right\rangle - \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}\|_2^2\right) \\
&= c\left(2\left\langle \boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}, (2c)^{-1}\mathbf{g}_j \right\rangle - \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_{\sigma^*(j)}\|_2^2 - \|(2c)^{-1}\mathbf{g}_j\|_2^2 + \|(2c)^{-1}\mathbf{g}_j\|_2^2\right) \\
&= c\left(-\|\boldsymbol{\mu}_{\sigma^*(j)} - \boldsymbol{\mu}_a + (2c)^{-1}\mathbf{g}_j\|_2^2 + \|(2c)^{-1}\mathbf{g}_j\|_2^2\right) \\
&= c\left(-\|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 + \|(2c)^{-1}\mathbf{g}_j\|_2^2\right).
\end{aligned}$$

For any $\mathbf{F} \in \mathcal{F}$, we then have

$$\begin{aligned}
\sum_j \sum_a \beta_{ja} F_{ja} &= c \sum_j \sum_a \left(-\|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 + \|(2c)^{-1}\mathbf{g}_j\|_2^2\right) F_{ja} \\
&= c\left(-\sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja} + \sum_j \|(2c)^{-1}\mathbf{g}_j\|_2^2 \sum_a F_{ja}\right) \\
&\overset{(i)}{=} c\left(-\sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja} + \sum_j \|(2c)^{-1}\mathbf{g}_j\|_2^2 \sum_a F^*_{ja}\right) \\
&= c\left(-\sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja} + \sum_j \sum_a \|(2c)^{-1}\mathbf{g}_j\|_2^2 F^*_{ja}\right) \\
&= c\left(-\sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja} + \sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_{\sigma^*(j)}\|_2^2 F^*_{ja}\right) \\
&\overset{(ii)}{=} c\left(-\sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja} + \sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F^*_{ja}\right),
\end{aligned}$$

where step (i) holds because $\sum_a F_{ja} = 1 = \sum_a F^*_{ja}, \forall j$, and step (ii) holds because $F^*_{ja} = 1$ only if $a = \sigma^*(j)$. Again recall the shorthand

$$\eta(\mathbf{F}) \coloneqq \sum_j \sum_a \|\bar{\mathbf{h}}_j - \boldsymbol{\mu}_a\|_2^2 F_{ja}.$$

We have the more compact expression

$$\sum_j \sum_a \beta_{ja} F_{ja} = c\left(\eta(\mathbf{F}^*) - \eta(\mathbf{F})\right) \tag{15}$$

It follows that for any $\mathbf{F} \in \mathcal{F}$, the first constraint in $\texttt{IP}_3$ is satisfied if and only if

$$\eta(\mathbf{F}) \le \eta(\mathbf{F}^*).$$

Combining with the (14), we obtain that

$$\frac{\gamma}{2\ell} \le 2 \cdot \texttt{IP}_3 = 2 \cdot \left\{ \begin{array}{c} \max\limits_{\mathbf{F}} \dfrac{1}{2}\|\mathbf{F} - \mathbf{F}^*\|_1 \\ \text{s.t. } \eta(\mathbf{F}) \le \eta(\mathbf{F}^*) \\ \mathbf{F} \in \mathcal{F} \end{array} \right\}.$$

Rearranging terms, we have the bound

$$\gamma \le 2\ell \cdot \max\left\{\|\mathbf{F} - \mathbf{F}^*\|_1 : \eta(\mathbf{F}) \le \eta(\mathbf{F}^*), \mathbf{F} \in \mathcal{F}\right\}. \tag{16}$$

The result follows from the fact that $\|\mathbf{Y}^*\|_1 = n\ell$ and $\|\mathbf{F}^*\|_1 = n$.

B.4. Proof of Proposition 1

In this section we control S_1. We can further decompose S_1 as

$$\begin{aligned} S_1 &= \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\right\rangle + \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\mathbf{U}\mathbf{U}^\top\right\rangle \\ &\quad - \left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\mathbf{U}\mathbf{U}^\top\right\rangle \\ &\le 2\left|\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\right\rangle\right| + \left|\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\mathbf{U}\mathbf{U}^\top\right\rangle\right| \\ &=: 2T_1 + T_2 \end{aligned}$$

By the generalized Holder's inequality, we have

$$T_1 \le \gamma \cdot \|\mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\|_\infty$$

and

$$\begin{aligned} T_2 &= \left|\left\langle \widehat{\mathbf{Y}} - \mathbf{Y}^*, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\mathbf{U}\mathbf{U}^\top\right\rangle\right| \\ &= \left|\left\langle \left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\mathbf{U}\mathbf{U}^\top, \mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\right\rangle\right| \end{aligned}$$

$$\leq \gamma \cdot \|\mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\|_\infty$$

where the last inequality holds since

$$\|\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\mathbf{U}\mathbf{U}^\top\|_1 \leq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 = \gamma.$$

Combining the above, we have

$$S_1 \leq 3\gamma \cdot \|\mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)\|_\infty.$$

Note that there are $m = nk$ distinct random variables in $\mathbf{U}\mathbf{U}^\top \left(\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right)$ and let us call them $X_1, \ldots, X_m$. For each i, we can see that X_i is the average of ℓ entries in $\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top$ and we let $\mathbf{B}_i$ be an $n \times n$ matrix with ℓ entries equal to 1 and the others equal to 0 such that $\ell X_i = \left\langle \mathbf{B}_i, \mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top \right\rangle$. To proceed, we need the Hanson-Wright inequality (an extension of Exercise 6.2.7 on pp. 140 in Vershynin (2017)).

Lemma 1 (Higher-dimensional Hanson-Wright inequality) *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independent, mean zero, sub-Gaussian random vectors in $\mathbb{R}^M$. Let $\mathbf{B}$ be an $N \times N$ matrix. For every $t \geq 0$ and some universal constant $c > 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i,j} B_{ij}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle - \mathbb{E}\sum_{i,j} B_{ij}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle\right| \geq t\right] \leq 4\exp\left[-c\min\left(\frac{t^2}{K^4 M\|\mathbf{B}\|_F^2}, \frac{t}{K^2\|\mathbf{B}\|}\right)\right]$$

where $K \coloneqq \max_i \|\mathbf{x}_i\|_{\psi_2}$.

The proof is given in Section D.1. Using Lemma 1, we see that for any $t \geq 0$

$$\mathbb{P}\left\{\ell X_i \geq t\right\} = \mathbb{P}\left\{\left\langle \mathbf{B}_i, \mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\right\rangle \geq t\right\} \leq 4\exp\left[-c\min\left(\frac{t^2}{K^4 d\ell}, \frac{t}{K^2\sqrt{\ell}}\right)\right].$$

We can choose $t^* = DK^2\sqrt{\ell}\left(\sqrt{d\log m} + \log m\right)$ with $K = \tau$ and $D > 0$ a universal constant. Apply the union bound, we have

$$S_1 \leq 3\gamma \cdot \frac{1}{\ell} \cdot t^*$$

with probability at least $1 - m \cdot \mathbb{P}\left\{\ell X \geq t\right\} \geq 1 - \exp\left(-C' \log m\right) = 1 - m^{-C'}$ where $C' > 0$ is a universal constant. The result follows from the condition of the proposition.

B.5. Proof of Proposition 2

In this section we control S_2. We have

$$\begin{aligned} S_2 &= \left\langle \mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right), \mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top \right\rangle \\ &\leq \mathrm{Tr}\left[\mathcal{P}_{T^\perp}\left(\widehat{\mathbf{Y}} - \mathbf{Y}^*\right)\right] \cdot \|\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\|_{\mathrm{op}} \\ &\leq \frac{\gamma}{\ell} \cdot \|\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\|_{\mathrm{op}}. \end{aligned}$$

Let $\mathrm{Var}\left(g_{ij}\right) = \nu^2$. We record a fact about the sub-Gaussian property of columns of $\mathbf{G}$.

Fact 1 *Let $\mathbf{x} \in \mathbb{R}^n$ be an arbitrary column of $\mathbf{G}$. We have*

$$\| \langle \mathbf{x}, \mathbf{w} \rangle \|_{\psi_2} \leq C \frac{\tau}{\nu} \sqrt{\mathbb{E} \langle \mathbf{x}, \mathbf{w} \rangle^2} \qquad \text{for any } \mathbf{w} \in \mathbb{R}^n,$$

where $C > 0$ is a universal constant and $C\frac{\tau}{\nu} \geq 1$.

The proof is given in Section D.2. Applying Lemma 8 with $\rho_0 = \frac{\tau}{\nu}$, we have

$$\|\frac{1}{d}\mathbf{G}\mathbf{G}^\top - \frac{1}{d}\mathbb{E}\mathbf{G}\mathbf{G}^\top\|_{\text{op}} \leq C_1 \rho_0^2 \left(\sqrt{\frac{2n}{d}} + \frac{2n}{d} \right) \|\frac{1}{d}\mathbb{E}\mathbf{G}\mathbf{G}^\top\|_{\text{op}}$$

with probability at least $1 - 2e^{-n}$. Here we let $m = d, u = n$ and define $\mathbf{x}_i$ to be the i-th column of $\mathbf{G}$ and $\mathbf{x}$ to be a vector independent of but identically distributed as each column of $\mathbf{G}$ (note that columns of $\mathbf{G}$ are identically distributed). We also use the fact that $\mathbb{E}\mathbf{x}\mathbf{x}^\top = \frac{1}{d}\mathbb{E}\mathbf{G}\mathbf{G}^\top = \nu^2\mathbf{I}$. Multiplying d on both sides of the above equation yields

$$\|\mathbf{G}\mathbf{G}^\top - \mathbb{E}\mathbf{G}\mathbf{G}^\top\|_{\text{op}} \leq C_1 \left(\sqrt{\frac{2n}{d}} + \frac{2n}{d} \right) d\tau^2.$$

Hence, we have

$$S_2 \leq \frac{\gamma}{\ell} \cdot C_1 \left(\sqrt{\frac{2n}{d}} + \frac{2n}{d} \right) d\tau^2 = 2C_1\gamma k \left(\sqrt{\frac{d}{n}} + 1 \right) \frac{\Delta^2}{s^2}$$

The result follows from the condition of the proposition.

B.6. Proof of Proposition 3

We can compute

$$\left(\mathbb{E}\mathbf{H}\mathbf{H}^\top\right)_{ij} = \begin{cases} d\nu^2 + \|\boldsymbol{\mu}_{\sigma^*(i)}\|_2^2 & \text{if } i = j \\ \|\boldsymbol{\mu}_{\sigma^*(i)}\|_2^2 & \text{if } i \neq j \text{ and } \sigma^*(i) = \sigma^*(j) \\ \left\langle \boldsymbol{\mu}_{\sigma^*(i)}, \boldsymbol{\mu}_{\sigma^*(j)} \right\rangle & \text{otherwise.} \end{cases}$$

We partition the matrix $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ into k^2 of $\ell \times \ell$ blocks, and note that T_{ab} denotes the sum of entries within the (a, b)-th block. The constraints of program (4) implies that

1. $T_{aa} \leq 0$ for each $a \in [k]$ and $T_{ab} \geq 0$ for each $a \neq b \in [k]$;
2. $T_{ab} = T_{ba}$ for each $a, b \in [k]$;
3. $-T_{aa} = \sum_{b \in [k]: b \neq a} T_{ab}$ for each $a \in [k]$;
4. $-\sum_{a \in [k]} T_{aa} + \sum_{a,b \in [k]: a \neq b} T_{ab} = \gamma$ and thus $-\sum_{a \in [k]} T_{aa} = \sum_{a,b \in [k]: a \neq b} T_{ab} = \frac{\gamma}{2}$.

Since $\widehat{\mathbf{Y}} - \mathbf{Y}^*$ has zero diagonal, we can write

$$\begin{aligned} S_4 &= \sum_{a\in[k]} T_{aa}\|\boldsymbol{\mu}_a\|_2^2 + 2\sum_{a,b\in[k]:a<b} T_{ab}\left\langle \boldsymbol{\mu}_a, \boldsymbol{\mu}_b\right\rangle \\ &= -\sum_{a,b\in[k]:a<b} T_{ab}\Delta_{ab}^2 \\ &= -\frac{1}{2}\sum_{a,b\in[k]:a\neq b} T_{ab}\Delta_{ab}^2 \\ &\leq -\frac{1}{2}\sum_{a,b\in[k]:a\neq b} T_{ab}\Delta^2 \\ &= -\frac{1}{4}\Delta^2\gamma. \end{aligned}$$

Appendix C. Proof of Theorem 2

We define the shorthand

$$\gamma_{\texttt{IP}} \coloneqq \max\left\{\frac{1}{2}\|\mathbf{F}-\mathbf{F}^*\|_1 : \eta(\mathbf{F}) \leq \eta(\mathbf{F}^*), \mathbf{F}\in\mathcal{F}\right\}.$$

It is not hard to see that $\gamma_{\texttt{IP}}$ takes integer values in $[0, n]$. If $\gamma_{\texttt{IP}} = 0$ then we are done. We therefore focus on the case $\gamma_{\texttt{IP}} \in [n]$.

Suppose $\gamma_{\texttt{IP}} > 3nke^{-s^2/C_0^2}$ for a fixed $C_0 > D/c$. Note that

$$3nke^{-s^2/C_0^2} \overset{(i)}{\leq} nk\cdot\frac{1}{k}\cdot e^{-s^2/\left(2C_0^2\right)} \leq ne^{-s^2/\left(2C_0^2\right)} < n$$

where step (i) holds since we have assumed $s^2 \geq C_s k$ for some universal constant $C_s > 0$. We record an important result for our proof.

Lemma 2 *Let $m \geq 4$ and $g \geq 1$ be integers. Let $\mathbf{X} \in \mathbb{R}^{m\times g}$ be a matrix such that each X_{ja} is a sub-Gaussian random variable with its mean equal to λ_{ja} and its sub-Gaussian norm no larger than ρ_{ja}, and each pair X_{ja} and X_{ib} are independent for $j \neq i$ and $a, b \in [g]$. Then for some universal constant $D > 0$ and for any $\beta \in (0, m]$, we have*

$$\sum_{j,a} X_{ja}M_{ja} \leq D\sqrt{\lceil\beta\rceil\left(\sum_{j,a}\rho_{ja}^2 M_{ja}\right)\log\left(3mg/\beta\right)} + \sum_{j,a}\lambda_{ja}M_{ja},$$
$$\forall \mathbf{M}\in\{0,1\}^{m\times g} : \mathbf{M}\mathbf{1}_g \leq \mathbf{1}_m, \|\mathbf{M}\|_1 = \lceil\beta\rceil,$$

with probability at least $1 - \frac{1.5}{m}$.

The proof is given in Section D.3. Define the set

$$\mathcal{M} \coloneqq \left\{\mathbf{M}\in\{0,1\}^{n\times k} : \mathbf{M}\mathbf{1}_k \leq \mathbf{1}_n, \|\mathbf{M}\|_1 = \gamma_{\texttt{IP}}, M_{j,\sigma^*(j)} = 0\ \forall j\in[n]\right\}.$$

For any $\mathbf{F}$ feasible to $\mathtt{IP}_3$, we have

$$
\begin{aligned}
0 &\leq \frac{1}{c}\left(\eta(\mathbf{F}^*) - \eta(\mathbf{F})\right) \\
&\overset{(i)}{=} \sum_{j\in[n]}\sum_{a\in[k]} \beta_{ja} F_{ja} \\
&= \sum_{(j,a):F_{ja}=1, a\neq\sigma^*(j)} \beta_{ja} \\
&\leq \max_{\mathbf{M}\in\mathcal{M}} \sum_j \sum_{a\neq\sigma^*(j)} \beta_{ja} M_{ja} \\
&\overset{(ii)}{\leq} \max_{\mathbf{M}\in\mathcal{M}} \left[D\sqrt{\gamma_{\mathtt{IP}}\tau^2 \left(\sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja}\right) \log\left(3n\left(k-1\right)/\gamma_{\mathtt{IP}}\right)} - c\sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja} \right] \\
&\leq \max_{\mathbf{M}\in\mathcal{M}} \left[D\sqrt{\gamma_{\mathtt{IP}}\tau^2 \left(\sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja}\right) \frac{s^2}{C_0^2}} - c\sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja} \right] \\
&\leq \left(\frac{D}{C_0} - c\right) \cdot \max_{\mathbf{M}\in\mathcal{M}} \sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja}
\end{aligned}
$$

where step (i) holds by Equation (15), step (ii) holds by Lemma 2 with $g = k-1$ since only $k-1$ entries of $\{\beta_{ja}\}$ are considered for each j in the sum above (ii), and the last step holds since $\gamma_{\mathtt{IP}}\Delta^2 \leq \sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja}$. Since $C_0 > D/c$ and $\sum_j \sum_{a\neq\sigma^*(j)} \Delta^2_{\sigma^*(j),a} M_{ja} > 0$, the RHS above is negative, which is a contradiction. Hence, we must have $\gamma_{\mathtt{IP}} \leq 3nke^{-s^2/C_0^2} \leq ne^{-s^2/\left(2C_0^2\right)}$ and the result follows from the fact that $\|\mathbf{F}^*\|_1 = n$.

Appendix D. Proof of technical results

In this section we provide the proofs of the technical results used in the proofs of our main theorems.

D.1. Proof of Lemma 1

We record the following lemma (Exercise 6.2.7 on pp. 140 in Vershynin (2017)).

Lemma 3 (Higher-dimensional Hanson-Wright inequality) *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be independent, mean zero, sub-Gaussian random vectors in $\mathbb{R}^M$. Let $\mathbf{B} = \{B_{ij}\}$ be an $N \times N$ matrix. There exists some universal constant $c > 0$ such that for every $t \geq 0$*

$$
\mathbb{P}\left[\left|\sum_{i,j:i\neq j}^{N} B_{ij}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle\right| \geq t\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4 M\|\mathbf{B}\|_F^2}, \frac{t}{K^2\|\mathbf{B}\|_{\text{op}}}\right)\right]
$$

where $K := \max_i \|\mathbf{x}_i\|_{\psi_2}$.

With this result, we only need to prove the same tail bound for $\mathbb{P}\left[\left|\sum_{i=1}^N B_{ii}\left(\|\mathbf{x}_i\|_2^2 - \mathbb{E}\|\mathbf{x}_i\|_2^2\right)\right| \geq t\right]$. To prove that, we cite another useful lemma (Theorem 2.8.2 on pp. 36 in Vershynin (2017)).

Lemma 4 (Bernstein's inequality for sub-exponential random variables) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $\mathbf{a} \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{K_1^2\|\mathbf{a}\|_2^2}, \frac{t}{K_1\|\mathbf{a}\|_\infty}\right)\right]$$

where $K_1 \coloneqq \max_i \|X_i\|_{\psi_1}$.

Here, $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm; see Vershynin (2017) for more details. We work under the premise of Lemma 3. Since $\mathbf{x}_i$ are independent sub-Gaussian random vectors, each $\|\mathbf{x}_i\|_2^2 - \mathbb{E}\|\mathbf{x}_i\|_2^2$ is the sum of M independent, mean zero, sub-exponential random variables with sub-exponential norm equal to K^2. Then Lemma 4 implies

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} B_{ii}\left(\|\mathbf{x}_i\|_2^2 - \mathbb{E}\|\mathbf{x}_i\|_2^2\right)\right| \geq t\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4 M\|\mathbf{B}\|_F^2}, \frac{t}{K^2\|\mathbf{B}\|_{\text{op}}}\right)\right]$$

as required.

D.2. Proof of Fact 1

We prove the following equivalent statement

$$\|\langle \mathbf{x}, \mathbf{w}\rangle\|_{\psi_2}^2 \leq C\frac{\tau^2}{\nu^2}\mathbb{E}\langle \mathbf{x}, \mathbf{w}\rangle^2 \qquad \text{for any } \mathbf{w} \in \mathbb{R}^n,$$

where $C > 0$ is a universal constant and $C\frac{\tau^2}{\nu} \geq 1$. We first establish a relationship between τ^2 and $\text{Var}(x_1)$: Proposition 2.5.2 on pp. 24 of Vershynin (2017) implies that $\frac{C'\tau^2}{\nu^2} \geq \frac{1}{2}$ for some universal constant $C' > 0$. Hence, we have

$$\begin{aligned}
\|\langle \mathbf{x}, \mathbf{w}\rangle\|_{\psi_2}^2 &\overset{(i)}{\leq} 2C'\sum_{i\in[n]} w_i^2\|x_i\|_{\psi_2}^2 \\
&= 2C'\frac{\tau^2}{\nu^2}\sum_{i\in[n]} w_i^2\nu^2 \\
&\overset{(ii)}{=} 2C'\frac{\tau^2}{\nu^2}\mathbb{E}\langle \mathbf{x}, \mathbf{w}\rangle^2,
\end{aligned}$$

where (i) holds according to Proposition 2.6.1 on pp. 28 of Vershynin (2017), and (ii) holds since x_i are i.i.d. and $\mathbb{E}x_i = 0$. Letting $C = 2C'$ completes the proof.

D.3. Proof of Lemma 2

We define

$$L_{\mathbf{M}} \coloneqq \sum_{j,a}(X_{ja} - \lambda_{ja})M_{ja},$$

$$R_{\beta,\mathbf{M}} \coloneqq D\sqrt{\lceil\beta\rceil\left(\sum_{j,a}\rho_{ja}^2 M_{ja}\right)\log(3mg/\beta)},$$

$$\mathcal{M}_\beta \coloneqq \left\{\mathbf{M} \in \{0,1\}^{m\times g} : \mathbf{M}\mathbf{1}_g \le \mathbf{1}_m, \|\mathbf{M}\|_1 = \lceil\beta\rceil\right\}.$$

To establish a uniform bound in β, we apply a discretization argument to the possible values of β. Define the shorthand $E \coloneqq (0, m]$. We can cover E by the sub-intervals $E_t \coloneqq (t-1, t]$ for $t \in [m]$. For each $t \in [m]$ we define the probability

$$\alpha_t \coloneqq \mathbb{P}\left\{\exists \beta \in E_t, \exists \mathbf{M} \in \mathcal{M}_\beta : L_{\mathbf{M}} > R_{\beta,\mathbf{M}}\right\}.$$

We bound each of these probabilities:

$$\begin{aligned}
\alpha_t &\overset{(i)}{\le} \mathbb{P}\left\{\exists \mathbf{M} \in \mathcal{M}_t : L_{\mathbf{M}} > R_{t,\mathbf{M}}\right\} \\
&\le \mathbb{P}\left\{\bigcup_{\mathbf{M}\in\mathcal{M}_t} \left\{L_{\mathbf{M}} > R_{t,\mathbf{M}}\right\}\right\} \\
&\le \sum_{\mathbf{M}\in\mathcal{M}_t} \mathbb{P}\left\{L_{\mathbf{M}} > R_{t,\mathbf{M}}\right\}, \qquad (17)
\end{aligned}$$

where step (i) holds since $\beta \in E_t$ implies $\beta \le \lceil\beta\rceil = t$.

Note that each $X_{ja} - \lambda_{ja}$ is an independent zero-mean sub-Gaussian random variable and the squared sub-Gaussian norm of $L_{\mathbf{M}}$ is at most $C_{\psi_2}\sum_{j,a}\rho_{ja}^2 M_{ja}$ where $C_{\psi_2} > 0$ is a universal constant. We apply Hoeffding inequality (Lemma 7) to bound the probability on the RHS of (17):

$$\begin{aligned}
\mathbb{P}\left\{L_{\mathbf{M}} > R_{t,\mathbf{M}}\right\} &\le \exp\left\{-\frac{cD^2 t\left(\sum_{j,a}\rho_{ja}^2 M_{ja}\right)\log(3mg/t)}{C_{\psi_2}\sum_{j,a}\rho_{ja}^2 M_{ja}}\right\} \\
&\le \exp\left\{-4t\log(3mg/t)\right\}
\end{aligned}$$

where $c > 0$ is a universal constant. Plugging this back to (17), we have for each $t \in [m]$,

$$\begin{aligned}
\alpha_t &\le \sum_{\mathbf{M}\in\mathcal{M}_t} \exp\left\{-4t\log(3mg/t)\right\} \\
&= \binom{m}{t} g^t \exp\left\{-4t\log(3mg/t)\right\} \\
&\le \left(\frac{me}{t}\right)^t g^t \exp\left\{-4t\log(3mg/t)\right\} \\
&\le \exp\left\{t\log(3mg/t) + t - 4t\log(3mg/t)\right\} \\
&\le \exp\left\{-t\log(3mg/t)\right\} = \left(\frac{t}{3mg}\right)^t, \qquad (18)
\end{aligned}$$

where the last inequality follows from $t \le t\log(3mg/t)$ for $t \in [m]$. It follows that

$$\begin{aligned}
&\mathbb{P}\left\{\exists \beta \in E, \exists \mathbf{M} \in \mathcal{M}_\beta : L_{\mathbf{M}} > R_{\beta,\mathbf{M}}\right\} \\
&\le \mathbb{P}\left\{\bigcup_{t=1}^{m}\left\{\exists \beta \in E_t, \exists \mathbf{M} \in \mathcal{M}_\beta : L_{\mathbf{M}} > R_{\beta,\mathbf{M}}\right\}\right\}
\end{aligned}$$

$$\leq \sum_{t=1}^{m} \alpha_t$$
$$\leq \sum_{t=1}^{m} \left(\frac{t}{3mg}\right)^t =: P_1(m).$$

It remains to show that $P_1(m) \leq \frac{1.5}{m}$. Since

$$\begin{aligned} P_1(m) &\leq \sum_{t=1}^{m} \left(\frac{t}{3m}\right)^t \\ &\leq \frac{1}{3m} + \sum_{t=2}^{m} \left(\frac{t}{3m}\right)^t \\ &\leq \frac{1}{3m} + m \cdot \max_{t=2,3,\ldots,m} \left(\frac{t}{3m}\right)^t, \end{aligned}$$

the proof is completed if for each integer $t = 2, 3, \ldots, m$, we can show the bound $\left(\frac{t}{3m}\right)^t \leq \frac{1}{m^2}$, or equivalently $f(t) := t(\log 3m - \log t) \geq 2 \log m$. Since $t \leq m$, $f(t)$ has derivative

$$f'(t) = \log 3m - \log t - 1 \geq \log 3m - \log\left(\frac{3m}{3}\right) - 1 = \log 3 - 1 \geq 0.$$

Therefore, $f(t)$ is non-decreasing for $2 \leq t \leq m$ and therefore $f(t) \geq f(2) = 2 \log 3m - 2 \log 2 \geq 2 \log m$. Hence, $P_1(m) \leq \frac{1.5}{m}$.

Appendix E. Proof of Theorem 3

We only need to prove the first part of the theorem. The second part follows immediately from the first part and Theorem 1.

The proof follows similar lines as those of Theorem 17 and Lemma 18 in Makarychev et al. (2016). In the rest of the section, we work under the context of Algorithms 1 and 2. Recall that $k' = \left|\{B_t\}_{t\geq 1}\right|$ and we let $\epsilon := \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 / \|\mathbf{Y}^*\|_1$. We have the following lemma.

Lemma 5 *There exists a partial matching π' between $[k]$ and $[k']$ and a universal constant $C > 0$ such that*

$$\left|\bigcup_{t=\pi'(a)} C_a^* \cap B_t\right| \geq (1 - C\epsilon)\, n.$$

The proof is given in Section E.1. The next lemma concerns the quality of clustering by Algorithm 3.

Lemma 6 *There exists a permutation π on $[k]$ and a universal constant $C > 0$ such that*

$$\left|\bigcup_{t=\pi(a)} C_a^* \cap U_t\right| \geq (1 - C\epsilon)\, n.$$

The proof is given in Section E.2. The result follows from combining the above lemmas and the fact that

$$\texttt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*) = 1 - \frac{1}{n} \max_{\pi \in S_k} \left| \bigcup_{t=\pi(a)} C_a^* \cap U_t \right|.$$

E.1. Proof of Lemma 5

We define $\mathbf{y}_a$ to be an arbitrary row of $\mathbf{Y}^*$ whose index is in C_a^*.

$$\begin{aligned} G_a &:= \left\{ i \in C_a^* : \|\widehat{\mathbf{Y}}_{i\bullet} - \mathbf{y}_a\|_1 \le \frac{\ell}{8} \right\}, \qquad \forall a \in [k] \\ G &:= \bigcup_{a \in [k]} G_a, \\ H &:= V \backslash G. \end{aligned}$$

We construct a partial matching π' between sets C_a^* and B_t by matching every cluster C_a^* with the first B_t that intersects G_a, and we let $\pi'(a) = t$. Since each $i \in [n]$ belongs to some B_t, we are able to match every C_a^* with some B_t. The fact that we cannot match two distinct clusters C_a^* and C_b^* with the same B_t as well as the rest of the proof are given by the following fact.

Fact 2 *We have*

1. *For each $a \in [k]$ and $t \in [k']$ such that $t = \pi'(a)$, we have $B_t \cap G_b = \emptyset$ for any $b \in [k] \setminus \{a\}$ and $B_t \subset G_a \cup H$;*
2. *For each $a \in [k]$ and $t \in [k']$ such that $t = \pi'(a)$, we have*
$$|B_t \cap C_a^*| \ge |G_a| - |B_t \cap H|.$$
3. *We have*
$$\sum_{t=\pi'(a)} |B_t \cap C_a^*| \ge |V| - 2|H|.$$
4. *There exists a universal constant $C > 0$ such that $|H| \le C\epsilon n$.*

The proof is given below.

E.1.1. PROOF OF FACT 2

1. Suppose that there exist B_t and $b \in [k]$ such that $b \neq a$ and $B_t \cap G_b \neq \emptyset$. Let $u \in B_t \cap G_a$ and $v \in B_t \cap G_b$. Since G_a and G_b are disjoint, we know that $u \neq v$. Let $w \in B_t$. Then we have
$$\begin{aligned} \|\widehat{\mathbf{Y}}_{u\bullet} - \widehat{\mathbf{Y}}_{w\bullet}\|_1 &\le \frac{\ell}{4} \\ \|\widehat{\mathbf{Y}}_{v\bullet} - \widehat{\mathbf{Y}}_{w\bullet}\|_1 &\le \frac{\ell}{4}. \end{aligned}$$

Therefore

$$\|\widehat{\mathbf{Y}}_{u\bullet} - \widehat{\mathbf{Y}}_{v\bullet}\|_1 \leq \|\widehat{\mathbf{Y}}_{u\bullet} - \widehat{\mathbf{Y}}_{w\bullet}\|_1 + \|\widehat{\mathbf{Y}}_{v\bullet} - \widehat{\mathbf{Y}}_{w\bullet}\|_1 \leq \frac{\ell}{2}.$$

This implies

$$\begin{aligned}\|\mathbf{y}_a - \mathbf{y}_b\|_1 &\leq \|\mathbf{y}_a - \widehat{\mathbf{Y}}_{u\bullet}\|_1 + \|\widehat{\mathbf{Y}}_{u\bullet} - \widehat{\mathbf{Y}}_{v\bullet}\|_1 + \|\mathbf{y}_b - \widehat{\mathbf{Y}}_{v\bullet}\|_1 \\ &\leq \frac{\ell}{8} + \frac{\ell}{2} + \frac{\ell}{8} < \ell,\end{aligned}$$

which is a contradiction to the fact that $\|\mathbf{y}_a - \mathbf{y}_b\|_1 = 2\ell$. To complete the proof, we note that for any $i \in B_t$ we have either $i \in G_a$ or $i \in H$.

2. Fix $i \in G_a$ for some $a \in [k]$. For any $j \in G_a$ we have $j \in B(i)$ since

$$\|\widehat{\mathbf{Y}}_{i\bullet} - \widehat{\mathbf{Y}}_{j\bullet}\|_1 \leq \|\mathbf{y}_a - \widehat{\mathbf{Y}}_{i\bullet}\|_1 + \|\mathbf{y}_a - \widehat{\mathbf{Y}}_{j\bullet}\|_1 \leq \frac{\ell}{4}.$$

Therefore, by definition

$$|B_t| \geq |B(i)| \geq |G_a|.$$

We have

$$\begin{aligned}|B_t \cap C_a^*| &\overset{(i)}{\geq} |B_t \cap G_a| \\ &= |B_t| - |B_t \backslash G_a| \\ &\overset{(ii)}{=} |B_t| - |B_t \cap H| \\ &\geq |G_a| - |B_t \cap H|,\end{aligned}$$

where step (i) holds since $G_a \subset C_a^*$ and step (ii) holds since $B_t \subset G_a \cup H$.

3. Summing the LHS of the above equation over $t = \pi'(a)$ gives

$$\begin{aligned}\sum_{t=\pi'(a)} |B_t \cap C_a^*| &= \sum_{a\in[k]} |G_a| - \sum_{t=\pi'(a)} |B_t \cap H| \\ &\geq \sum_{a\in[k]} |G_a| - \sum_{t\geq 1} |B_t \cap H| \\ &\overset{(i)}{=} |G| - |V \cap H| \\ &= |V| - 2|H|,\end{aligned}$$

where step (i) holds since $B_t \cap H$ are disjoint and $\bigcup_{t\geq 1} B_t = V$.

4. We have

$$|H| \cdot \frac{\ell}{8} \leq \sum_{i\in H} \|\widehat{\mathbf{Y}}_{i\bullet} - \mathbf{y}_{\sigma^*(i)}\|_1 \leq \|\widehat{\mathbf{Y}} - \mathbf{Y}^*\|_1 \leq \epsilon\|\mathbf{Y}^*\|_1 = \epsilon \cdot n\ell$$

where the last step follows from the fact that $\|\mathbf{Y}^*\|_1 = n\ell$. The result follows.

E.2. Proof of Lemma 6

Let π' be the partial matching between C_a^* and B_t from Lemma 5. Define $\pi(a) = \pi'(a)$ for $\pi'(a) \le k$. If the resulting π is a partial permutation, we extend π to a permutation defined on $[k]$ in an arbitrary way. We may assume that $\{U_t\}_{t\in[k]}$ are $\{B_t\}_{t\in[k]}$ WLOG, and that U_t consists of B_t and some elements from sets B_u with $u > k$. We have

$$\begin{aligned}\left|\bigcup_{t=\pi(a)} C_a^* \cap U_t\right| &\ge \left|\bigcup_{t=\pi'(a)\le k} C_a^* \cap B_t\right| \\ &= \left|\bigcup_{t=\pi'(a)} C_a^* \cap B_t\right| - \left|\bigcup_{t=\pi'(a)>k} C_a^* \cap B_t\right| \\ &\ge \left(1 - C'\epsilon\right) n - \left|\bigcup_{t=\pi'(a)>k} C_a^* \cap B_t\right|\end{aligned}$$

where $C' > 0$ is a universal constant. Define

$$\begin{aligned}T_1 &\coloneqq \left\{t > k : t = \pi'(a) \text{ for some } a \in [k]\right\}, \\ T_2 &\coloneqq \left\{t \in [k] : t \ne \pi'(a) \text{ for any } a \in [k]\right\}.\end{aligned}$$

Note that $|T_1| = |T_2|$ and for any $t_1 \in T_1$ and $t_2 \in T_2$ we have $|B_{t_1}| \le |B_{t_2}|$. Therefore,

$$\begin{aligned}\left|\bigcup_{t=\pi'(a)>k} C_a^* \cap B_t\right| &\le \left|\bigcup_{t\in T_1} B_t\right| \\ &\le \left|\bigcup_{t\in T_2} B_t\right| \\ &\le |V| - \left|\bigcup_{t=\pi'(a)} C_a^* \cap B_t\right| \\ &= C'\epsilon n.\end{aligned}$$

The result follows by setting $C \coloneqq 2C'$.

Appendix F. Proof of Theorem 4

Let $\mathrm{Var}\left(g_{ij}\right) = \nu^2$. For $a \in [k]$, define $\widehat{C}_a \coloneqq \{i \in [n] : \widehat{\sigma}_i = a\}$ the estimated clusters encoded in $\widehat{\boldsymbol{\sigma}}$, and recall that our cluster center estimators are defined by $\hat{\boldsymbol{\mu}}_a \coloneqq \ell^{-1}\sum_{i\in\widehat{C}_a} \mathbf{h}_i$. We assume $\left\{\widehat{C}_a\right\}$ achieves the lowest clustering error as given in Theorem 3 WLOG. For each $a \in [k]$, we have

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_a\|_2 &\le \|\frac{1}{\ell}\sum_{i\in\widehat{C}_a} \mathbf{h}_i - \frac{1}{\ell}\sum_{j\in C_a^*} \mathbf{h}_j\|_2 + \|\frac{1}{\ell}\sum_{j\in C_a^*} \mathbf{h}_j - \boldsymbol{\mu}_a\|_2 \\ &=: Q_1 + Q_2.\end{aligned}$$

F.1. Controlling Q_1

Define $\epsilon \coloneqq \mathtt{err}(\widehat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^*)$. We work on the event that the result Theorem 3 is true. We have

$$Q_1 = \frac{1}{\ell} \| \sum_{i \in \widehat{C}_a \backslash C_a^*} \mathbf{h}_i - \sum_{j \in C_a^* \backslash \widehat{C}_a} \mathbf{h}_j \|_2$$

Note that $\left|\widehat{C}_a \backslash C_a^*\right| = \left|C_a^* \backslash \widehat{C}_a\right|$ so we can pair each point in $\widehat{C}_a \backslash C_a^*$ with a point in $C_a^* \backslash \widehat{C}_a$. Let us pair ith point in $\widehat{C}_a \backslash C_a^*$ with $j(i)$th point in $C_a^* \backslash \widehat{C}_a$, and define $\mathcal{M} \coloneqq \{(i, j(i))\}$. We have $|\mathcal{M}| \leq n\epsilon$ and we can write

$$\begin{aligned}
Q_1 &= \frac{1}{\ell} \| \sum_{(i,j(i)) \in \mathcal{M}} \left(\mathbf{h}_i - \mathbf{h}_{j(i)}\right) \|_2 \\
&\leq \frac{1}{\ell} \sum_{(i,j(i)) \in \mathcal{M}} \|\mathbf{h}_i - \mathbf{h}_{j(i)}\|_2 \\
&\leq \frac{1}{\ell} \sum_{(i,j(i)) \in \mathcal{M}} \left(\Delta_{\sigma^*(i), \sigma^*(j(i))} + \|\mathbf{g}_i - \mathbf{g}_{j(i)}\|_2\right) \\
&\leq \frac{1}{\ell} \sum_{(i,j(i)) \in \mathcal{M}} \left(C_q \Delta + \|\mathbf{g}_i - \mathbf{g}_{j(i)}\|_2\right),
\end{aligned}$$

where the last step holds for some universal constant $C_q > 0$ given that $\max_{a,b \in [k]} \Delta_{ab} \leq C_q \Delta$. By Theorem 3.1.1 on pp. 41 of Vershynin (2017), $\frac{1}{\sqrt{2}\nu} \|\mathbf{g}_i - \mathbf{g}_{j(i)}\|_2 - \sqrt{d}$ is a sub-Gaussian random variable with sub-Gaussian norm at most $C_{\psi_2} \frac{\tau^2}{\nu^2}$ where $C_{\psi_2} > 0$ is a universal constant. Then Lemma 7 implies that

$$\mathbb{P}\left[\frac{1}{\sqrt{2}\nu} \|\mathbf{g}_i - \mathbf{g}_{j(i)}\|_2 - \sqrt{d} \geq C \frac{\tau^2}{\nu^2} \sqrt{\log n}\right] \leq n^{-C'}$$

for some universal constants $C, C' > 2$. By the union bound and the facts that $|\mathcal{M}| \leq n$ and $\nu \lesssim \tau$, we have

$$\max_{(i,j) \in \mathcal{M}} \|\mathbf{g}_i - \mathbf{g}_{j(i)}\|_2 \leq C_g \left(\tau\sqrt{2d} + C\tau\sqrt{2 \log n}\right)$$

with probability at least $1 - n^{-C_1}$ where $C_g, C_1 > 0$ are universal constants.

Therefore, we have

$$\begin{aligned}
Q_1 &\leq C_0 \left(\Delta + \tau\sqrt{d} + \tau\sqrt{\log n}\right) \cdot k \exp\left[-\frac{s^2}{C_e}\right] \\
&\leq C_0 \left(\Delta + \tau\sqrt{d} + \tau\sqrt{\log n}\right) \cdot \exp\left[-\frac{s^2}{2C_e}\right]
\end{aligned}$$

for some universal constant $C_0, C_e > 0$ with probability at least $1 - n^{-C_1}$, where the last step holds since $s^2 \geq k$. The fact that $e^x \geq 1 + x > x$ for any x implies

$$\exp\left[-\frac{s^2}{4C_e}\right] \leq \frac{4C_e}{s^2} = \frac{\tau}{\Delta} \cdot \frac{4C_e}{s} \leq 4C_e \frac{\tau}{\Delta}$$

where the last step holds since we have $s \geq 1$ by the conditions of Theorem 3. Hence, we have

$$\begin{aligned} Q_1 &\leq C_0\tau\left(4C_e + \sqrt{d} + \sqrt{\log n}\right) \cdot \exp\left[-\frac{s^2}{4C_e}\right] \\ &\leq C_1\tau\left(1 + \sqrt{d} + \sqrt{\log n}\right) \cdot \exp\left[-\frac{s^2}{4C_e}\right] \\ &\leq 2C_1\tau\left(\sqrt{d} + \sqrt{\log n}\right) \cdot \exp\left[-\frac{s^2}{4C_e}\right] \end{aligned}$$

where $C_1 > 0$ is a universal constant.

F.2. Controlling Q_2

We have

$$Q_2 = \|\frac{1}{\ell}\sum_{j\in C_a^*} \mathbf{g}_j\|_2.$$

We see that $\frac{1}{\ell}\sum_{j\in C_a^*} g_{ji}$ has variance $\frac{1}{\ell}\nu^2$. By Proposition 2.6.1 on pp. 28 and Theorem 3.1.1 on pp. 41 of Vershynin (2017), $\frac{\sqrt{\ell}}{\nu}\|\frac{1}{\ell}\sum_{j\in C_a^*}\mathbf{g}_j\|_2 - \sqrt{d}$ is a sub-Gaussian random variable with sub-Gaussian norm at most $C_{\psi_2}\frac{\tau^2}{\nu^2}$ where $C_{\psi_2} > 0$ is a universal constant. Then Lemma 7 implies that

$$\mathbb{P}\left[\frac{\sqrt{\ell}}{\nu}\|\frac{1}{\ell}\sum_{j\in C_a^*}\mathbf{g}_j\|_2 - \sqrt{d} \geq C\frac{\tau^2}{\nu^2}\sqrt{\log n}\right] \leq n^{-C'}$$

for some universal constants $C, C' > 0$. Since $\nu \lesssim \tau$, there exists a universal constant $C_0 > 0$ such that

$$Q_2 \leq C_0\tau\left(\sqrt{\frac{kd}{n}} + \sqrt{\frac{k\log n}{n}}\right)$$

with probability at least $1 - n^{-C'}$.

Appendix G. Technical lemmas

The following lemma is Theorem 2.6.2 on pp. 28 in Vershynin (2017).

Lemma 7 (General Hoeffding's inequality) *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-Gaussian random variables. Then, for every $t \geq 0$ we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^N X_i\right| \geq t\right] \leq 2\exp\left[-\frac{ct^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2}\right],$$

where $c > 0$ is a universal constant.

The following lemma is Exercise 4.7.3 in Vershynin (2017).

Lemma 8 (Tail bound of covariance matrix of sub-Gaussians) *Let* $\mathbf{x}$ *be a sub-Gaussian vector and let* $\mathbf{x}_1, \ldots, \mathbf{x}_m$ *be independent samples of* $\mathbf{x}$*. Let* m *be a positive integer and define*

$$\mathbf{\Sigma} := \mathbb{E}\mathbf{x}\mathbf{x}^\top,$$
$$\mathbf{\Sigma}_m := \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i{}^\top.$$

Let $\rho_0 \geq 1$ *be such that*

$$\| \langle \mathbf{x}, \mathbf{w} \rangle \|_{\psi_2} \leq \rho_0 \sqrt{\mathbb{E} \langle \mathbf{x}, \mathbf{w} \rangle^2} \qquad \text{for any } \mathbf{w} \in \mathbb{R}^N.$$

For any $u \geq 0$*, we have for a universal constant* $C > 0$*,*

$$\|\mathbf{\Sigma}_m - \mathbf{\Sigma}\|_{\text{op}} \leq C\rho_0^2 \left(\sqrt{\frac{N+u}{m}} + \frac{N+u}{m} \right) \|\mathbf{\Sigma}\|_{\text{op}}$$

with probability at least $1 - 2e^{-u}$*.*