Sketch-and-Lift: Scalable Subsampled Semidefinite Program for K-means Clustering

Yubo Zhuang **Xiaohui Chen** **Yun Yang**
University of Illinois at Urbana-Champaign

Abstract

Semidefinite programming (SDP) is a powerful tool for tackling a wide range of computationally hard problems such as clustering. Despite the high accuracy, semidefinite programs are often too slow in practice with poor scalability on large (or even moderate) datasets. In this paper, we introduce a linear time complexity algorithm for approximating an SDP relaxed K-means clustering. The proposed *sketch-and-lift* (SL) approach solves an SDP on a subsampled dataset and then propagates the solution to all data points by a nearest-centroid rounding procedure. It is shown that the SL approach enjoys a similar exact recovery threshold as the K-means SDP on the full dataset, which is known to be information-theoretically tight under the Gaussian mixture model. The SL method can be made adaptive with enhanced theoretic properties when the cluster sizes are unbalanced. Our simulation experiments demonstrate that the statistical accuracy of the proposed method outperforms state-of-the-art fast clustering algorithms without sacrificing too much computational efficiency, and is comparable to the original K-means SDP with substantially reduced runtime.

1 INTRODUCTION

Clustering is a widely explored unsupervised machine learning task to partition data into a fewer number of unknown groups. The K-means clustering is a classical clustering method with good empirical performance on recovering the cluster labels for Euclidean data (MacQueen, 1967). Under the Gaussian mixture model (GMM) with isotropic noise, K-means clustering is equivalent to the maximum likelihood estimator (MLE) for cluster labels, which is known to be worst-case NP-hard (Aloise et al., 2009). Fast approximation algorithms to solve the K-means such as Lloyd's algorithm (Lloyd, 1982; Lu and Zhou, 2016) and spectral clustering (Meila and Shi, 2001; Ng et al., 2001; Vempala and Wang, 2004; Achlioptas and McSherry, 2005; von Luxburg, 2007; von Luxburg et al., 2008) provably yield consistent recovery when different groups are well separated. Recently, semi-definite programming (SDP) relaxations (Peng and Wei, 2007; Mixon et al., 2016; Li et al., 2017; Fei and Chen, 2018; Chen and Yang, 2021a; Royer, 2017; Giraud and Verzelen, 2018; Bunea et al., 2016) have emerged as an important approach for clustering due to its superior empirical performance (Peng and Wei, 2007), robustness against outliers and adversarial attack (Fei and Chen, 2018), and attainment of the information-theoretic limit (Chen and Yang, 2021b). Despite having polynomial time complexity, the SDP relaxed K-means has notoriously poor scalability to large (or even moderate) datasets for instance by interior point methods (Alizadeh, 1995; Jiang et al., 2020), as the typical runtime complexity of an interior point algorithm for solving the SDP is at least $O(n^{3.5})$, where n is the sample size. Hence the goal of this paper is to derive a computationally cheap approximation to the SDP relaxed K-means formulation for reducing the time complexity while maintaining statistical optimality.

Sketching, a popular numerical technique in randomized linear algebra to speed up matrix computations via compressing a matrix to a much smaller one by multiplying a random matrix (Drineas and Mahoney, 2017), has been deployed in recent years as a valuable tool in many data science applications at scale. We refer papers from Bluhm and Stilck França (2019); Yurtsever et al. (2017) for sketching semidefinite programs. In this paper, we consider *subsampling sketches*, constructed by subsampling m out of the total n data points (seen as a random projection with indepen-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

dent Bernoulli weights). Subsampling sketches perform sample size reduction to substantially reduce the runtime complexity when $m \ll n$ by instead solving a much smaller m-dimensional SDP.

In this paper, we propose a *sketch-and-lift* (SL) approach for fast and scalable clustering. The main goal for the SL approach is to look for the smallest possible projected sample size m in order to maximally reduce the computational cost without sacrificing too much statistical accuracy. Under the GMM, we show that the subsampled size m can be made almost independent of n to guarantee exact recovery on randomly subsampled data points when the signal is above a threshold depending on the down-sampling ratio m/n. To reconstruct a solution to the full dataset, we need to project back (or lift) from cluster labels estimated by an m-dimensional SDP to cluster labels of the entire n data points. This back projection step takes $O(n)$ complexity. Thus the proposed SL approach has an overall *linear* time complexity as long as $m = O(n^c)$ for some constant $c \in (0, 1)$, which substantially mitigates the high polynomial runtime complexity of solving the original SDP relaxed K-means. For instance, we can set $c = 2/7$ if the interior point method is used to solve the SDP (Jiang et al., 2020).

The baseline SL procedure begins with a uniform subsampling on the entire dataset. When the cluster sizes are unequal, the single down-sampling parameter γ creates a non-trivial bias on restricting the data dimension growth rate. Motivated from this observation, we propose two SL variants: one based on bias-correction by equalizing the size of the estimated clusters from the subsampled SDP, and the other based on adaptively choosing the sampling weights on the input data points. By doing so, we show that the constraint on the data dimension is unnecessary after bias correction, and the bias-corrected SL and weighted SL boost numeric performance for unequal cluster size case.

The rest of paper is structured as follows. In Section 2, we describe some background on the K-means clustering and its SDP relaxation. In Section 3, we present our SL approach and its variants. In Section 4, we derive the guarantees for exact recovery of the SL methods under the standard Gaussian mixture model. In Section 5, we show some statistical and computational comparisons for the SL methods and the state-of-the-art K-means algorithm in various settings.

2 BACKGROUND

We first provide some background on the K-means clustering. After that, we describe a matrix-lifting semidefinite relaxation scheme which turns the mixed integer program associated with the K-means into a convex one by throwing away the integer constraints, and review its theoretical properties.

2.1 K-means clustering

Let $X_1, \dots, X_n$ be a sequence of p-dimensional vectors and $X = (X_1, \dots, X_n) \in \mathbb{R}^{p\times n}$ denote the data matrix with n data points in $\mathbb{R}^p$. Suppose that there is a clustering structure $G_1^*, \dots, G_K^*$ (i.e., a partition on $[n] := \{1, \dots, n\}$ such that $\bigsqcup_{k=1}^K G_k^* = [n]$, where $\bigsqcup$ denotes the disjoint union) on the n data point indices. To recover the true clustering structure $G_1^*, \dots, G_K^*$ from data, we consider the K-means defined as the following constrained combinatorial optimization problem:

$$\begin{aligned} &\max_{G_1,\dots,G_K} \sum_{k=1}^K \frac{1}{|G_k|} \sum_{i,j\in G_k} \langle X_i, X_j\rangle \\ &\text{subject to } \bigsqcup_{k=1}^K G_k = [n], \end{aligned} \tag{1}$$

where $\langle X_i, X_j\rangle = X_i^\top X_j$ is the Euclidean inner product in $\mathbb{R}^p$ that represents the similarity between two vectors X_i and X_j. Note that the objective function in (1) is proportional to the log-likelihood function of cluster labels after profiling (maximizing) out the cluster centers as nuisance parameters in the Gaussian mixture model with constant and isotropic noise. Therefore, solving for (1) is equivalent to computing the maximum likelihood estimator.

The standard approach to finding an approximate solution of the K-means problem is the Lloyd's algorithm, also known as Voronoi iteration or K-means algorithm, which repeatedly finds the centroid of the points within each cluster G_k and then re-assigns points to the K clusters according to which of these centroids is closest. To overcome some shortcomings of the K-means algorithm that is unstable in both its running time and approximation accuracy, the K-means++ algorithm (Arthur and Vassilvitskii, 2007) is proposed and becomes a state-of-the-art clustering algorithm hereafter. The K-means++ algorithm addresses the unstability issue by specifying a careful initialization procedure to seed the K-means algorithm. See the paper from Saxena et al. (2017) for a recent review on different clustering techniques and their developments.

We end this subsection with a brief review on theoretical developments of the K-means clustering. Consistency of the K-means estimation of the clustering centers is studied by Pollard (1981), without concerning the computational complexity. Kumar and Kannan (2010) and Awasthi and Sheffet (2012) show that if the true cluster centers are sufficiently well-separated relative to their spreads, then the Lloyd's algorithm

initialized by spectral clustering achieves exact recovery of the cluster labels. Partial recovery bounds of Lloyd's algorithm for local search solution to the K-means are derived by Lu and Zhou (2016).

2.2 K-means as mixed integer program

Next, we describe an equivalent formulation of the K-means optimization (1) that will be useful in motivating its convex relaxation in the next subsection. Note that we can express the cluster labels for each partition $G_1, \ldots, G_K$ of $[n]$ by their *one-hot encoding*: we can associate each $(G_k)_{k=1}^K$ with a binary *assignment matrix* $H = (h_{ik}) \in \{0,1\}^{n\times K}$, where $h_{ij} = 1$ indicates X_i belongs to cluster k, and $h_{ik} = 0$ otherwise. Because each row of H contains exactly one non-zero entry, there is one-to-one mapping (up to assignment labeling) between the partition and the assignment matrix. Thus recovery of the true clustering structure is equivalent to recovery of the associated assignment matrix, and the K-mean clustering problem can be re-expressed as a (non-convex) mixed integer program:

$$\begin{aligned} &\max_{H} \langle A, HBH^\top \rangle \\ &\text{subject to } H \in \{0,1\}^{n\times K},\ H\mathbf{1}_K = \mathbf{1}_n, \end{aligned} \tag{2}$$

where $A = X^\top X$ is the $n \times n$ similarity matrix and $\mathbf{1}_n$ denotes the n-dimensional vector of all ones.

2.3 SDP relaxed K-means

Relaxing the above mixed integer program (2) by changing variable $Z = HBH^\top$, we arrive at the SDP relaxed approximation of the K-means clustering problem:

$$\begin{aligned} &\hat{Z} = \arg\max_{Z\in\mathbb{R}^{n\times n}} \langle A, Z \rangle \\ &\text{subject to } Z \succeq 0,\ \mathrm{tr}(Z) = K,\ Z\mathbf{1}_n = \mathbf{1}_n,\ Z \geq 0, \end{aligned} \tag{3}$$

where $Z \geq 0$ means each entry $Z_{ij} \geq 0$ and $Z \succeq 0$ means the matrix Z is symmetric and positive semi-definite. This SDP approximation relaxes the integer constraint on H into two linear constraints $\mathrm{tr}(Z) = K$ and $Z \geq 0$ that are satisfied by any $Z = HBH^T$ as H ranges over feasible solutions of problem (2). The SDP in (3) was first introduced by Peng and Wei (2007) and it was shown that this SDP relaxing the integer constraint is information-theoretically tight under the standard Gaussian mixture model (Chen and Yang, 2021b) (see our brief review below).

Membership matrix Z^* corresponding to the true partition $G_1^*, \ldots, G_K^*$ is a block diagonal matrix with K blocks, each of which has size $n_k \times n_k$ with all entries equal to n_k^{-1}. Here $n_k = |G_k^*|$ is the size of cluster k.

Note that the SDP solution $\hat{Z}$ of (3) is generally not integral in the sense that $\hat{Z}$ cannot be directly used to recover a partition estimate of the data points. In such case, we can apply some rounding technique to project $\hat{Z}$ back to yield a partition as the solution to the original discrete optimization problem (1). For instance, we may cluster the top K eigenvectors of $\hat{Z}$ to extract the estimated partition structure $\hat{G}_1, \ldots, \hat{G}_K$. On the other hand, it is known that rounding is not necessary as the relaxed SDP solution can be directly used to recover the K-means in (1), when the separation of cluster centers is large enough, a property often referred in literature as the *hidden integrality* (Fei and Chen, 2018; Chen and Yang, 2021b; Ndaoud, 2018; Awasthi et al., 2015).

Formally, consider the standard GMM where n_k observations from the k-th cluster follow i.i.d. $N(\mu_k, \sigma^2 I_p)$ for $k \in [K]$. It is proved by Chen and Yang (2021b) that for any $\alpha > 0$, if the squared minimal separation $\Delta^2 = \min_{1\leq k\neq l\leq K} \|\mu_k - \mu_l\|^2$ satisfies $\Delta^2 \geq (1+\alpha)\bar{\Delta}_*^2$, where

$$\bar{\Delta}_*^2 = 4\sigma^2 \left(1 + \sqrt{1 + \frac{p}{n_* \log n}}\right) \log n, \tag{4}$$

with $n_* = \min_{1\leq k\neq l\leq K} 2n_k n_l/(n_k + n_l)$ denoting the smallest pairwise harmonic average and $n = \sum_{k=1}^K n_k$ the total sample size, then with probability at least $1 - c_1 K^2 n^{-c_2\alpha}$ for some constants $c_1, c_2 > 0$, the SDP in (3) will produce the integral solution Z^* that corresponds to exact recovery (cf. Lemma A.1 for a precise statement). Regarding the lower bound, Chen and Yang (2021b) shows that in the equal cluster size case where $n_k = n/K$ for each $k \in [K]$, if $\Delta^2 \leq (1-\alpha)\bar{\Delta}_*^2$ holds for any $\alpha > 0$, then with probability at least $1 - cKn^{-1}$, no clustering algorithm can achieve simultaneous exact recovery of all cluster labels. In other words, $\bar{\Delta}_*^2$ is the cutoff value for exact recovery of GMM.

3 PROPOSED LINEAR TIME APPROXIMATION ALGORITHMS

Now we introduce our proposed linear time complexity algorithm for approximating the SDP relaxed K-means problem. We also discuss some variants that significantly boost the numerical performance and are better suited to handle unequal cluster scenarios.

3.1 Sketch-and-lift for the K-means SDP

We first provide some intuition before formally describing our sketch-and-lift (SL) approach for the K-means SDP (3). As in the Lloyd's algorithm, finding

a best clustering scheme consists of two intermediate steps: 1. estimate the center of each cluster; 2. determine cluster labels based on which of these centers is closest. It turns out that the loss of statistical accuracy in estimating the centers in the first step due to using fewer but correctly labeled samples is much less severe than that due to using mislabeled samples. This motivates us to apply stable and reliable but computationally more expensive clustering algorithms such as the SDP (3) to a smaller subsample of size m to extract correct cluster labels of the subsample. Based on the cluster labels, we obtain an estimator of the cluster centers (as within cluster averages) using the subsample, and then apply the estimated centers for clustering the entire data. As we will illustrate in the theoretical analysis, the sample size m for estimating the centers via the K-means SDP (3) in the first step can be as small as $O\big((\log n)^2\big)$ (see the remark after Theorem 4.3) in order to guarantee the exact recovery of entire data cluster labels in the second step under suitable separation conditions. Due to this extremely low sample size requirement on m, the overall computational complexity will be dominated by the linear $O(n)$ complexity in the second step. Note that the subsampling in the first step corresponds to the "sketch" operation; and the label recovery based on centers estimated from a subsample corresponds to the "lift" operation. In principle, this SL idea can be incorporated with any accurate clustering method. We choose the SDP relaxed K-means in this paper mainly due to its theoretical optimality in cluster label recovery (cf. Section 2.3).

Now we formally describe our SL approach. Let $\gamma \in (0,1)$ be a pre-specified subsampling factor which may depend on the sample size n. We first randomly sample an index subset $T \subset [n]$ with i.i.d. $\mathrm{Ber}(\gamma)$. Denote the subsampled data matrix as $V = (X_i)_{i\in T}$, which is of size p-by-m where $m = |T|$ follows a Binomial $\mathrm{Bin}(n,\gamma)$ distribution. Here assuming the i.i.d. sampling is mainly for technical convenience, and in practice one can also uniformly sample a subset of $[n]$ with fixed size $\lfloor n\gamma \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x. After the subsampling, we apply the SDP relaxed K-means (3) to V:

$$\hat{W} = \arg\max_{Z\in\mathbb{R}^{m\times m}} \langle V^\top V, W\rangle$$
$$\text{subject to } W \succeq 0,\ \mathrm{tr}(W) = K,\ W\mathbf{1}_m = \mathbf{1}_m,\ W \geq 0. \tag{5}$$

Once we obtain a partition estimate $\hat{R}_1,\dots,\hat{R}_K \subset T$ from $\hat{W}$ on the subset V (perhaps after a rounding procedure), we compute the centroids $\bar{X}_k = |\hat{R}_k|^{-1}\sum_{j\in\hat{R}_k} X_j$ based on the estimated partition $T = \bigsqcup_{k=1}^K \hat{R}_k$. Finally, we project back the cluster labels to all data points in $X \setminus V$ by mapping them to the nearest centroid among $\bar{X}_1,\dots,\bar{X}_K$, i.e., for each $X_i \in X \setminus V$, we assign $i \in \hat{G}_k$ when $\|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|$ for all $l \neq k$ and $l \in [K]$, where $\|\cdot\|$ denotes the ℓ_2-norm. The SL algorithm is summarized in Algorithm 1 (steps 2-3 in subroutine Algorithm 2 correspond to rounding).

Algorithm 1: Sketch-and-lift algorithm for K-means SDP with sampling weights $(w_1,\dots,w_n)$.

Input: sampling weights $(w_1,\dots,w_n)$ with $w_1 = \cdots = w_n = \gamma \in (0,1)$ being the subsampling factor.

1 (Sketch) Independent sample an index subset $T \subset [n]$ via $\mathrm{Ber}(w_i)$ and store the subsampled data matrix $V = (X_i)_{i\in T}$.

2 Run subroutine Algorithm 2 with input V to get a partition estimate $\hat{R}_1,\dots,\hat{R}_K$ for T.

3 Compute the centroids $\bar{X}_k = |\hat{R}_k|^{-1}\sum_{j\in\hat{R}_k} X_j$ for $k \in [K]$.

4 (Lift) For each $i \in [n]\setminus T$, assign $i \in \hat{G}_k$ if $\|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \quad \forall l \neq k,\ l \in [K]$.

Output: A partition estimate $\hat{G}_1,\dots,\hat{G}_K$ for $[n]$.

Algorithm 2: Subroutine for solving K-means SDP.

Input: Data matrix $V \in \mathbb{R}^{p\times m}$ containing m points.

1 Solve the SDP in (5) using V to get solution $\hat{W}$.

2 Perform the spectral decomposition of $\hat{W}$ and take the top K eigenvectors $(\hat{u}_1,\dots,\hat{u}_k)$.

3 Run K-means clustering on $(\hat{u}_1,\dots,\hat{u}_k)$ and extract the cluster labels $\hat{R}_1,\dots,\hat{R}_K$ as a partition estimate for $[m]$.

Output: A partition estimate $\hat{R}_1,\dots,\hat{R}_K$ for $[m]$.

We highlight that the SL approach has a *linear* time complexity in the sample size n if we choose $m = O(n^c)$ for some small constant $c \in (0,1)$. Theoretically, it is shown in Section 4 that the subsampling factor γ is allowed to vanish to zero while retaining statistical validity of SL. Obviously, any clustering algorithm has at least a linear time complexity since it should visit at least one time for each data point. On the other hand, it is shown that the SL enjoys a similar exact recovery threshold as the original SDP on all data points, which is known to achieve the information-theoretic limit (Chen and Yang, 2021b). Empirically, we demonstrate in Section 5 that around the sharp threshold of exact recovery, the SL approach statistically outperforms the widely used K-means++ algorithm (Arthur and Vassilvitskii, 2007) by a large margin in terms of the error rates.

Remark 3.1 (Multi-epoch with averaging). We can repeat the above SL procedure for multiple epochs to enhance the empirical performance. A simple way to achieve this is to randomly partition the data points into $\lfloor n/m \rfloor$ blocks, each of which is a sequence of independent Bernoulli trials of size n with success probability γ. Then we run $\lfloor n/m \rfloor$ SL procedure in Algorithm 1 on the independent data blocks and average the centroids estimated from the multiple epochs before lifting. Such procedures can be easily paralleled in a distributed system and therefore the computational burden for running multiple epochs is essentially the same as one sketch-and-lift pass. In Section 5, we present some numerical result for the multi-epoch SL with averaging. This multi-epoch SL approach is summarized in Algorithm 3 below. □

Algorithm 3: Multi-epoch sketch-and-lift algorithm for K-means SDP.

Input: Subsample size m.

Randomly partition data indices $[n]$ into $S = \lfloor n/m \rfloor$ blocks $T_1, \dots, T_S$ with size m.

for $s = 1, \dots, S$ **do**

  (Sketch) Run subroutine Algorithm 2 with input $V_s = (X_i)_{i \in T_s}$ to get a partition estimate $\hat{R}_{s,1}, \dots, \hat{R}_{s,K}$ for T.

  Compute the centroids $\bar{X}_{s,k} = |\hat{R}_{s,k}|^{-1} \sum_{j \in \hat{R}_{s,k}} X_j$ for $k \in [K]$.

Compute the aggregated centroids $\bar{X}_k = S^{-1} \sum_{s=1}^{S} \bar{X}_{s,k}$ for $k \in [K]$.

(Lift) For each $i \in [n] \setminus T$, assign $i \in \hat{G}_k$ if $\|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \quad \forall l \neq k, \, l \in [K]$.

Output: A partition estimate $\hat{G}_1, \dots, \hat{G}_K$ for $[n]$.

Remark 3.2 (Related work on stochastic block models). Mixon and Xie (2020) proposed a subsampled SDP for the two-component stochastic block model (SBM) with equal community size. The approach presented by Mixon and Xie (2020) first randomly subsamples a small vertex set according a Bernoulli process with rate $\gamma \in (0, 1)$ and then solves the community detection problem on the induced subgraph. The solution on the subgraph is finally projected by a majority voting procedure to all nodes in the whole graph. It is shown by Mixon and Xie (2020); Abdalla and Bandeira (2021) that the subsampling factor $\gamma > c$, where $c > 0$ is a constant depending on the edge connecting probabilities within-community and between-communities in the graph, is needed for exact community recovery with high probability. In our clustering problem, we allow the subsampling ratio $\gamma = o(1)$ (cf. Theorem 4.1 below), so the computational cost can be much further reduced than the subsampled SDP for the SBM. In particular, we can choose very small γ such that the reduced SDP problem size $m = \lfloor n\gamma \rfloor$ is nearly independent of n (up to some polylogarithmic factor $\log^c n$). Moreover, the subsampled SDP for SBM proposed by Mixon and Xie (2020) works only for two-component equal cluster size case, while our SL approach works for unbalanced K-component clusters and it can be further enhanced with bias-correction (Section 3.2) and non-uniform sampling weights (Section 3.3) to better handle the general unequal cluster size scenario. □

The SL approach performs the uniform subsampling on n data points, which is natural for equal cluster size case. If the cluster sizes are not equal, the estimated centroids based on the partition given by the sketched SDP in (5) have different variances. Thus by comparing the distances between data point in $X \setminus V$ with $\bar{X}_k$ and $\bar{X}_l$ will create a larger bias than that in the equal cluster case. Theoretically, such an extra bias term will imposes the unnecessary constraint of $p = O\big((\gamma n/K)^2\big)$ (cf. Theorem 4.1). To mitigate this issue, we propose two procedures in the following subsections.

3.2 Bias-corrected sketch-and-lift (BCSL)

Suppose we have obtained a partition $\hat{R}_1, \dots, \hat{R}_K$ for V by the sketched SDP and an estimate of the cluster centers $\bar{X}_k$ and $\bar{X}_l$. Let $\underline{m} := \min_{k \in [K]} |\hat{R}_k|$ be the smallest cluster size. To fairly compare the distances $\|X_i - \bar{X}_k\|$ and $\|X_i - \bar{X}_l\|$ by matching the variance of all cluster center estimates $\{\bar{X}_k : k \in [K]\}$, we further down-sample $\hat{R}_1, \dots, \hat{R}_K$ to have the same size $\underline{m}$. In particular, we can randomly sample a subset $\tilde{R}_k$ with equal size $\underline{m}$ from each $\hat{R}_k$. Then we apply the lift step to propagate $\tilde{R}_1, \dots, \tilde{R}_K$ to the original data to obtain a partition $\hat{G}_1, \dots, \hat{G}_K$. As we will show in Theorem 4.3, this bias correction scheme removes the undesirable constraint on p as needed in the SL approach. The bias-corrected SL (BCSL) algorithm is summarized in Algorithm 4.

3.3 Weighted sketch-and-lift (WSL)

Another bias correcting method is to subsample $X_1, \dots, X_n$ with non-uniform weights that convey the cluster size information, so that in the sketched data matrix V, all clusters have roughly the same number of points. For example, we increase (decrease) the sampling weights for those points from small (large) clusters. Compared to the BCSL approach, this weighted SL (WSL) approach has no waste of information when estimating the cluster centers based on partition centroids, given we know the *ideal* sampling weights. However, the WSL appears to incur a chicken and egg problem as the ideal sampling weights requires knowl-

Algorithm 4: Bias-corrected sketch-and-lift algorithm for K-means SDP.

Input: subsampling factor $\gamma \in (0,1)$.

1 Independently sample an index subset $T \subset [n]$ via $\mathrm{Ber}(\gamma)$ and make the subsampled data matrix $V = (X_i)_{i\in T}$.

2 Run subroutine Algorithm 2 with input V to get a partition estimate $\hat{R}_1, \dots, \hat{R}_K$ for T.

3 For each $\hat{R}_k$, randomly sample a subset $\tilde{R}_k$ with equal size $\underline{m}$.

4 Compute the centroids $\bar{X}_k = |\tilde{R}_k|^{-1} \sum_{j\in\tilde{R}_k} X_j$ for $k \in [K]$.

5 For each $i \in [n] \setminus T$, assign $i \in \hat{G}_k$ if $\|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \quad \forall l \neq k,\, l \in [K]$.

Output: A partition estimate $\hat{G}_1, \dots, \hat{G}_K$ for $[n]$.

edge on the cluster membership of each data point. Fortunately, as we will discuss in Remark 3.3, a multi-round extension of the WSL which iteratively applies the WSL to refine the sampling weights based on the clustering labels in the previous round has surprisingly good numerical performance in that the recovery error decays as the round increases (cf. Figure 18 in the supplement). Now let us formally describe the WSL approach. For each data point X_i for $i \in G_k^*$, we denote $w_i^* = \gamma n/(Kn_k)$ as the ideal sketch weight for X_i. Suppose in practice we have a set of approximating weights $w_i \in [0,1]$ such that there exists a subset $D \subset [n]$ which satisfies for some $(\epsilon, \delta) \in [0,\infty) \times [0,1]$

$$|D| \geq (1-\delta)\, n \quad \text{and} \quad \left| \frac{w_i}{w_i^*} - 1 \right| \leq \epsilon. \tag{6}$$

Condition (6) requires the at least $(1-\delta)$ proportion of constructed sampling weights should be close to the true sampling weights with at most ϵ distortion. We call such weights *a set of* (ϵ, δ)*-weights*. Ideal weights are $(0,0)$-weights. In practice, a priori estimate for the weights w_i can be set through Lloyd's algorithm for the K-means. And the parameter γ can be chosen as small as $O(\log(n)/n)$, which implicitly shows that the weights w_i's would be as small as $o(1)$. The rest of the WSL is the same as Algorithm 1 with a general non-uniform sampling weighs $(w_1, \dots, w_n)$. In addition, we can also combine the BCSL with the WSL to enforce the equal cluster sizes when computing the centroids $\bar{X}_k$'s. The WSL algorithm is summarized in Algorithm 5.

Remark 3.3 (Multi-round WSL). The priori estimate for the weights (e.g., by Lloyd's or K-means++ algorithm) can be viewed as a warm start of WSL. To further boost the performance of the WSL, we can iteratively apply WSL to refine the recovered cluster labels, which is summarized in Algorithm 6 below. In Section 5, we present some superior numerical result for the multi-round WSL. □

Algorithm 5: Weighted sketch-and-lift algorithm for K-means SDP.

Input: subsampling factor $\gamma \in (0,1)$.

1 Run Lloyd's algorithm to obtain an initial partition estimate $\tilde{G}_1, \dots, \tilde{G}_K$ for $[n]$, with sizes $\tilde{n}_1, \dots, \tilde{n}_K$.

2 For each $i \in [n]$, set $w_i = \gamma n/(K\tilde{n}_k)$ if $i \in \tilde{G}_k$.

3 Run Algorithm 1 with weights $(w_1, \dots, w_n)$.

Output: A partition estimate $\hat{G}_1, \dots, \hat{G}_K$ for $[n]$.

Algorithm 6: Multi-round weighted sketch-and-lift algorithm for K-means SDP.

Input: subsampling factor $\gamma \in (0,1)$ and number of rounds R.

1 Run Algorithm 5 to get partition estimate $\tilde{G}_1, \dots, \tilde{G}_K$ for $[n]$.

2 **for** $r = 2, \dots, R$ **do**

3 For each $i \in [n]$, update sampling weight $w_i = \gamma n/(K\tilde{n}_k)$ if $i \in \tilde{G}_k$, where $\tilde{n}_k = |\tilde{G}_k|$.

4 Run Algorithm 1 with weights $(w_1, \dots, w_n)$ to update $\tilde{G}_1, \dots, \tilde{G}_K$.

Output: A partition estimate $\tilde{G}_1, \dots, \tilde{G}_K$ for $[n]$.

4 EXACT RECOVERY GUARANTEES

To study the theoretic properties of SL, we follow the literature by using the standard GMM as our working model. Specifically, we assume $X_1, \dots, X_n$ to be from the following data generating model: if $i \in G_k^*$, then

$$X_i = \mu_k + \epsilon_i, \tag{7}$$

where $\mu_1, \dots, \mu_K \in \mathbb{R}^p$ are the (unobserved) cluster centers and ϵ_i are i.i.d. $N(0, \sigma^2 I_p)$ noise. Recall that $\Delta^2 = \min_{1\leq k\neq l\leq K} \|\mu_k - \mu_l\|^2$ denotes the squared minimal separation between cluster centers.

Theorem 4.1 (Separation bound for exact recovery by SL). Let $\alpha > 0$ and $\gamma \in (0,1)$ be the subsampling ratio. Suppose that $n_1 = \cdots = n_K = n/K$. If $\Delta^2 \geq (1+\alpha)\bar{\Delta}_\gamma^2$, where

$$\bar{\Delta}_\gamma^2 = 4\sigma^2 \left(1 + \sqrt{1 + \frac{Kp}{\gamma n \log n}} \right) \log n, \tag{8}$$

then the output $\hat{G}_1, \dots, \hat{G}_K$ from the SL Algorithm 2 for the K-means SDP achieves exact recovery, i.e., $\hat{G}_k = G_k^*$ for all $k \in [K]$ with probability at least

$1 - C_1(\log(\gamma n))^{-C_2}$, provided that $K \le C_3 \frac{\log(\gamma n)}{\log\log(\gamma n)}$ and $p \le C_4(\gamma n/K)^2$, where $C_i, i = 1, 2, 3, 4$ are constants depending only on α.

Theorem 4.1 considers the equal cluster case, and includes the exact recovery property for the original SDP K-means (3) as a special case when $\gamma = 1$ (and no lift step is needed). Compared to the separation cutoff value (4) for exact recovery of the entire data, the separation requirement in (8) has an extra factor of γ^{-1} inside the square root—the larger $Kp/(\gamma n \log n)$ term corresponds to the statistical fluctuation of using only $\gamma n/K$ samples to estimate the p-dimensional cluster centers instead of n/K samples in the entire data, which appears to be inevitable for any single-epoch SL method. Interestingly, as we empirically observed in Section 5, the multi-epoch SL method summarized in Algorithm 3 has a noticeable improvement over the SL and appears to attain the optimal cutoff value (4) due to the usage of almost all data in estimating the cluster centers (by averaging across subsamples). Based on the numeric evidence, we pose the following conjecture as a future study.

Conjecture 4.2. Multi-epoch SL method with averaging (Algorithm 3) attains the information-theoretic threshold $\bar{\Delta}_*^2$ in (4) as the SDP (3) on the entire data.

For general subsampling ratio $\gamma \in (0, 1)$, we note that there is an additional constraint $p \lesssim (\gamma n/K)^2$ to ensure the exact recovery. This constraint can be shown even stricter $p \lesssim (\gamma n/K)$ for unequal cluster size case. This condition comes from the fact that when lift is needed to obtain the full cluster labels on all data points, we need to ensure that the estimated cluster sizes from the subsampled SDP (5) are approximately equal to match the variances. In contrast, we shall show that in Theorems 4.3 and 4.4 below that the BCSL does not require this condition, and the WSL still requires this condition $p \lesssim (\gamma n/K)^2$, but both will work for the unequal cluster size case as well.

Theorem 4.3 (Separation bound for exact recovery by BCSL). Let $\alpha > 0$, $\gamma \in (0, 1)$ be the subsampling ratio and $\underline{n} = \min_{k\in[K]} n_k$. If $\Delta^2 \ge (1 + \alpha)\bar{\Delta}_\gamma'^2$, where

$$\bar{\Delta}_\gamma'^2 = 4\sigma^2 \left(1 + \sqrt{1 + \frac{p}{\gamma \underline{n} \log n}}\right) \log n, \qquad (9)$$

then the output $\hat{G}_1, \ldots, \hat{G}_K$ from the BCSL Algorithm 4 achieves exact recovery with probability at least $1 - C_1(\log \gamma n)^{-C_2}$, provided that $K \le C_3 \frac{\log(\gamma n)}{\log\log(\gamma n)}$, $\underline{n} \ge C_4 n/\log(\gamma n)$, $\log n/\underline{n} \le C_5\gamma$, where C_i, $i = 1, 2, 3, 4, 5$ are only depend on α.

According to Theorem 4.3, the subsampling factor γ can be as small as $O\big((\log n)^2/n\big)$, corresponding to a minimal subsample size $m = O((\log n)^2)$. Consequently, the overall time complexity is dominated by the $O(n)$ complexity of the lift step.

Theorem 4.4 (Separation bound for exact recovery by WSL). Suppose we have a set of (ϵ, δ)-weights $w_i \in [0, 1]$ satisfying (6). Let $\alpha > 0$. If $\Delta^2 \ge (1 + \alpha)\bar{\Delta}_\gamma^2$, where $\bar{\Delta}_\gamma^2$ is defined in (8), then the WSL Algorithm 5 achieves exact recovery with probability at least $1 - C_1(\log \gamma n)^{-C_2}$, provided that

$$K \le C_3(\log \gamma n)/(\log\log \gamma n), \quad p \le C_4(\gamma n/K)^2,$$
$$\delta \le C_5(\underline{n}/n) \min\big\{1, \sqrt{\gamma n/p}\big\},$$
$$\epsilon \le C_6 \min\big\{1, \gamma n \log n/(Kp)\big\},$$

where constants C_i, $i = 1, \ldots, 6$, only depend on α.

We remark that the WSL by adjusting the bias with adaptive (non-uniform) weights essentially reduces the unequal sizes case to the equal size case. The cost of choosing the adaptive weights is that we need impose size conditions on (ϵ, δ) (e.g., $\epsilon, \delta = o(p^{-c})$ for some $c > 0$) such that they can absorb the effect coming from the growth of p. In Section B in the supplement, we evaluate the effect of initial weights by K-means++ algorithm. In addition, from the numerical results, we conjecture that the multi-round WSL summarized in Algorithm 6 can further relax the conditions to achieve exact recovery, for example, by throwing away the $p \le C_4(\gamma n/K)^2$ constraint. We leave its formal theoretical study to a future direction.

5 NUMERICAL EXPERIMENTS

In this section, we test the numerical performance for the SL method and its variants, and compare them with the K-means++ algorithm on synthetic data. MATLAB code implementing the SL approach and its variants are available at: https://github.com/Yubo02/Sketch-and-Lift-Scalable-Subsampled-Semidefinite-Program-for-K-means-Clustering

5.1 Setup

We generate data from a 4-component Gaussian mixture model (7) parametrized by (p, n, λ^*), where parameter $\lambda^* > 0$ characterizes the cluster center separation through $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$ and recall that $\bar{\Delta}_*^2$ is the theoretical cutoff in (4). We compare the following clustering methods.

- M_0 is the Matlab build-in K-means clustering implementation (default algorithm is K-means++).
- M_1 is the sketch-and-lift (SL) method described in Algorithm 1.
- M_2 is the bias-corrected sketch-and-lift (BCSL) method described in Algorithm 4.

- M_3 is the weighted sketch-and-lift (WSL) method described in Algorithm 5.
- M_4 is the multi-epoch sketch-and-lift (ME-SL) with averaging described in Algorithm 3.
- M_5 is the multi-round weighted sketch-and-lift (MR-WSL) method described in Algorithm 6 with output at the 4-th round.

For the SL methods (M1-M5), we vary the subsampling factor γ. We choose round number as 4 in M5 since according to the additional numerical results reported in the supplement, the MR-WSL typically reaches its best performance after 3-4 rounds. We also compare with the original SDP relaxed K-means method (3) (method O) on the entire data points whenever it is feasible to run (in our case when $n \leq 3000$). We report the error rate in recovering cluster labels and the running time for these algorithms averaged over 100 simulations.

5.2 Results

Due to the space limit, we report simulation results for equal cluster size case in this subsection. For complete numerical experiment results including the unequal cluster size case, we refer to Section B in the supplementary material.

The baseline setup is $p = 1000$, $n = 2000$, $\gamma = 0.1$ and $\lambda^* = 1.2$, expect when γ is changing, we use $n = 10000$. In each simulation setting, we vary one parameter and report the error rate, which is summarized in Figure 1. Figure 2 compares the runtimes as n changes. We observe that all SL methods have *significantly and uniformly smaller* error rate than the state-of-the-art K-means++ method (blue solid curve) in nearly all setups. We also note that runtime curves of the SL-methods on the log-scale are *parallel* to the K-means++ algorithm, indicating that SL methods have the same linear $O(n)$ complexity as the fast K-means++ (with difference only occurring in the leading constant). In comparison, the original SDP (O) has super linear complexity as expected.

One interesting observation is that the multi-epoch SL method (M4) is almost always the best method across all four settings in Figure 1, and has comparable performance as the original SDP method (O) in the range when it is feasible to run. Note that with sample size $n = 2000$ as the baseline, only misspecifying one cluster label in 1 out of 100 replicates will incur an error rate of 5×10^{-6}, meaning that an average error rate (over 100 replicates) of order below (or around) 10^{-5} can be viewed as perfect clustering (due to the log-scale, we display 10^{-6} when the actual error is zero). This empirical observation provides the numerical evidence for supporting Conjec-

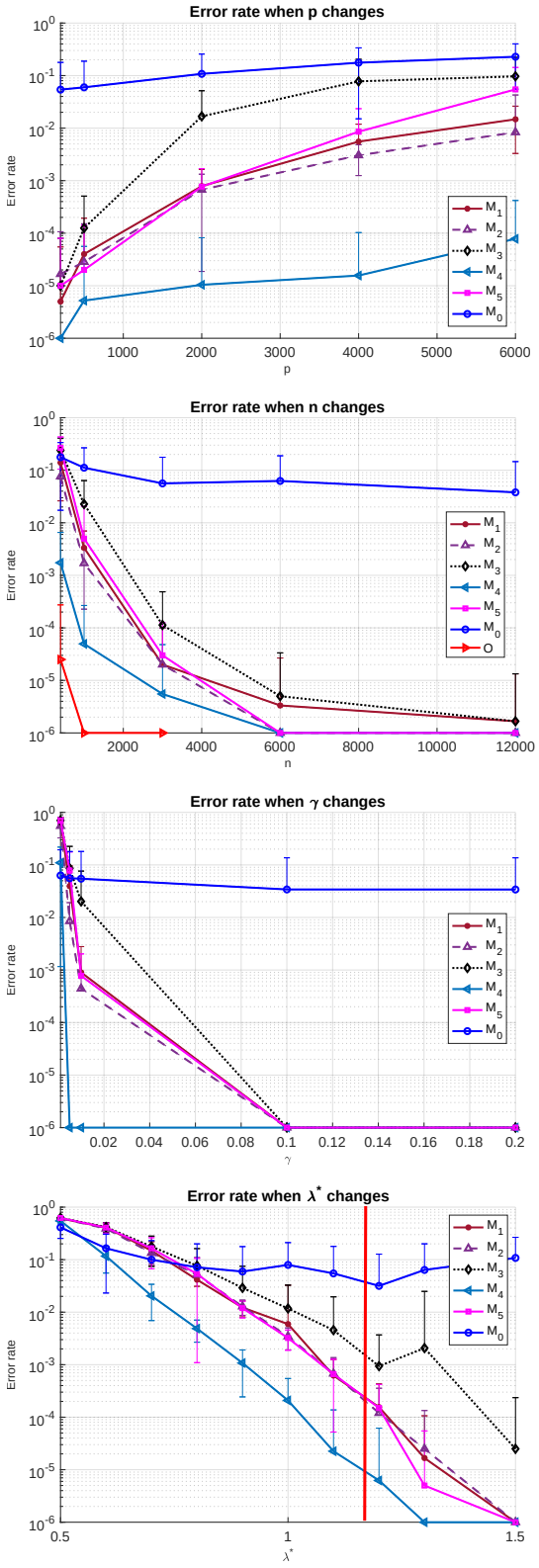

Figure 1: Log-scale error rates (with error bars) when one parameter varies. Zero error is displayed as 10^{-6} in the log-scale plot. Red vertical line in the lowest plot indicates theoretical threshold $\bar{\Delta}_\gamma^2$ for SL methods.

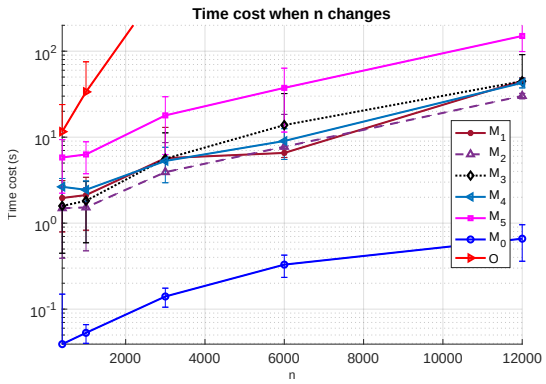

Figure 2: Log-scale runtime (with error bars) v.s. n.

ture 4.2 about the information-theoretic optimality of the multi-epoch SL.

We also report the error decay for the MR-WSL with K-means++ algorithm as the warm start in Section B in the supplement, where we observed that the MR-WSL has a surprisingly good recovery performance after 3-4 rounds. In particular, from Figure 1 we can see that the MR-WSL (M5) is the second best method in most settings (with the best being M4), and has significant improvements over its single round counterpart WSL (M3) due to progressive refinements on the estimated sampling weights (cf. Section B).

The plot at the very bottom in Figure 1 shows that the error rates as we change the separation parameter $\Delta^2 = \min_{1\le k\neq l\le K} \|\mu_k - \mu_l\|^2$, where the red vertical line indicates the theoretical threshold $\bar{\Delta}_\gamma^2$ for SL methods given in Theorem 4.1. Since error of order 10^{-5} is very close to perfect clustering, the numerical results in the plot are consistent with our theory that $\bar{\Delta}_\gamma^2$ characterizes the cutoff value for SL methods.

We also assess the impact of initialization (i.e., warm start effect) of WSL by K-means by looking at the estimated (ϵ, δ) parameters (Table 1 for the 1-st round in Section B). In particular, we find that for fixed $\epsilon = 0.2$, $\delta = 0.25373, 0.25598, 0.37208, 0.33282, 0.38929$ for $p = 200, 500, 2000, 4000, 6000$, respectively. Thus for increasing p corresponding to more difficult clustering problems, the quality of initial weights deteriorates by the K-means. Still, our WLS (in particular, the MR-WLS refinement) maintains good quality of cluster label recovery. And $\delta = 0.000005, 0.00006, 0.028845, 0.088965, 0.15886$, respectively when we perform the 2-nd round of WSL (Table 2 in Section B). Finally we can see that $\delta = 0$ uniformly for the 4-th round WSL (Table 4 in Section B). This shows that the multi-round WSL refines the clustering errors and can eventually achieve the exact recovery as if we initialize with the ideal weights.

Finally, we applied our method to two benchmark datasets. The first one considers 32-dimensional mass cytometry (CyTOF) dataset, consisting of protein expression levels of healthy human bone marrow mononuclear cells (BMMCs) from two healthy individuals. Following Levine et al. (2015), we run clustering analysis on individual H1, where $n = 72463$ cells were assigned to populations and $p = 32$. We report clustering results with $K = 14$ and $\gamma = 0.02$. The misclassification error for kmeans++ (our Algorithm 6 with 1-st round using kmeans++ as initialization) is 0.5709 (0.4719) with time cost 0.8757 (226.5155).

The second one is for unbalanced synthetic 2-D Gaussian clusters data presented by Rezaei and Fränti (2016), where $n = 6500, p = 2, K = 8, \gamma = 0.01$. The misclassification error for kmeans++ (our Algorithm 6 with 1-st round using kmeans++ as initialization) is 0.4301 (0.2213) and time cost is 0.0122 (0.5257). Thus, the SL method is robust to the GMM assumption and can improve the accuracy on top of kmeans++ with similar scalability (both time costs increased by $O(10^2)$ times as n becomes $O(10^2)$ times larger).

Acknowledgements

Xiaohui Chen was partially supported by NSF CAREER Award DMS-1752614. Yun Yang was partially supported by NSF DMS-1907316.

References

P. Abdalla and A. S. Bandeira. Community detection with a subsampled semidefinite program, 2021.

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In P. Auer and R. Meir, editors, *Learning Theory*, pages 458–469, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Optim.*, 5(1):13–51, 1995.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

P. Awasthi and O. Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, pages 191–200, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3333-7.

A. Bluhm and D. Stilck França. Dimensionality reduction of sdps through sketching. *Linear Algebra and its Applications*, 563:461–475, 2019. ISSN 0024-3795.

F. Bunea, C. Giraud, M. Royer, and N. Verzelen. PECOK: a convex optimization approach to variable clustering. *arXiv:1606.05100*, 2016.

X. Chen and Y. Yang. Diffusion k-means clustering on manifolds: Provable exact recovery via semidefinite relaxations. *Applied and Computational Harmonic Analysis*, 52:303–347, 2021a. ISSN 1063-5203.

X. Chen and Y. Yang. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021b.

P. Drineas and M. W. Mahoney. Lectures on randomized numerical linear algebra, 2017.

Y. Fei and Y. Chen. Hidden integrality of sdp relaxation for sub-gaussian mixture models. *arXiv:1803.06510*, 2018.

C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed $k$means. *arXiv:1807.07547*, 2018.

H. Jiang, T. Kathuria, Y. T. Lee, S. Padmanabhan, and Z. Song. A faster interior point method for semidefinite programming. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 910–918, 2020.

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

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E. ad D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, and G. P. Nolan. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015. ISSN 0092-8674.

X. Li, Y. Li, S. Ling, T. Stohmer, and K. Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *arXiv:1710.06008*, 2017.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

Y. Lu and H. Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv:1612.02099*, 2016.

J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*, pages 281–297, 1967.

M. Meila and J. Shi. Learning segmentation by random walks. In *In Advances in Neural Information Processing Systems*, pages 873–879. MIT Press, 2001.

D. G. Mixon and K. Xie. Sketching semidefinite programs for faster clustering, 2020.

D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv:1602.06612v2*, 2016.

M. Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *arXiv:1812.08078*, 2018.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. OPTIM*, 18(1):186–205, 2007.

D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 1981.

M. Rezaei and P. Fränti. Set-matching methods for external cluster validity. *IEEE Trans. on Knowledge and Data Engineering*, 28(8):2173–2186, 2016.

M. Royer. Adaptive clustering through semidefinite programming. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1795–1803. Curran Associates, Inc., 2017.

A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci*, 68: 2004, 2004.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36 (2):555–586, 2008.

A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage, 2017.

Supplementary Material: Sketch-and-Lift: Scalable Subsampled Semidefinite Program for K-means Clustering

A PROOF OF MAIN RESULTS

In this section, we prove the main result of this paper.

A.1 Auxiliary lemmas

Recall that $n_k = |G_k^*|$ denotes the number of data points in k-th cluster and $\Delta^2 = \min_{1\le k\neq l\le K} \|\mu_k - \mu_l\|^2$ is the minimal separation between cluster centers. Set $N = \min_{1\le k\neq l\le K}\{\frac{2n_k n_l}{n_k+n_l}\}$ and $\underline{n} = \min_{k\in[K]} n_k$.

Lemma A.1 (Separation bound for exact recovery by the full SDP: general case). If there exist constants $\tilde{\delta} > 0$ and $\beta \in (0,1)$ such that

$$\log n \ge \frac{(1-\beta)^2}{\beta^2}\frac{C_1 n}{N},\ \tilde{\delta} \le \frac{\beta^2}{(1-\beta)^2}\frac{C_2}{K},\ N \ge \frac{4(1+\tilde{\delta})^2}{\tilde{\delta}^2},$$

and

$$\Delta^2 \ge \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{p}{N\log n}+C_3 r_n}\right)\log n$$

with

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

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

Lemma A.1 is proved by Chen and Yang (2021b) (Theorem II.1). Specializing Lemma A.1 to equal cluster case $n_1 = \cdots = n_K = n/K$, we have the following corollary.

Corollary A.2 (Separation bound for exact recovery by the full SDP: equal cluster case). Let $\alpha > 0$ and $\overline{\Delta}_1^2$ be defined in (8). Suppose that the cluster sizes are equal and $K \le C_1 \log(n)/\log\log(n)$ for some small constant $C_1 > 0$ depending only on α. If $\Delta^2 \ge (1+\alpha)\overline{\Delta}_1^2$, then the SDP in (3) achieves exact recovery with probability at least $1 - C_2(\log n)^{-c_3}$, where C_2, c_3 are constants depending only on α.

A.2 Proof outline for Theorem 4.1

Before presenting the rigorous proof, we first discuss the overall strategy for proving Theorem 4.1 for equal cluster size case, which can be divided into three steps. Proofs for Theorems 4.3 and 4.4 have the same architecture. We define the events

$$\begin{aligned} A :=& \left\{\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*\right\},\\ B :=& \left\{\hat{R}_1 = R_1, \ldots, \hat{R}_K = R_K\right\},\\ B_\tau :=& \left\{(1-\tau)\frac{n\gamma}{K} \le |R_k| \le (1+\tau)\frac{n\gamma}{K},\ \forall k\in[K]\right\}, \end{aligned}$$

where $\tau \in (0,1)$ and $R_k = G_k^* \cap T$. Since probability of wrong recovery satisfies

$$\mathbb{P}(A^c) \le 2\mathbb{P}(B_\tau^c) + \mathbb{P}(B^c|B_\tau) + \mathbb{P}(A^c \cap B|B_\tau), \tag{10}$$

it suffices to bound the three terms on the right-hand side of (10), where the first term is due to the tolerance of approximate equality of the cluster sizes on subsampled data (step 1), the second term is due to wrong recovery using the subsampled SDP (step 2), and the third term is due to the lifting procedure (step 3) on the data points that are not sampled in step 1.

In step 1, we reduce the exact recovery problem on the entire n data points to the subsampled $m \approx n\gamma$ data points with K clusters with approximately equal size that resembles the problem structure of the original SDP clustering problem. The violation probability of approximate cluster size in step 1 can be controlled by the classical Chernoff bound

$$\mathbb{P}(B_\tau^c) \le 2n^{-1} \quad \text{for } \tau \asymp \sqrt{K\log(n)/n\gamma}.$$

In step 2, since the subsampling procedure is independent of the original data points, we can treat the subsampled data points $V = (X_i)_{i\in T}$ as a new clustering problem. Based on this observation, we establish the exact recovery guarantee on subsampled data using the separation upper bound proved for the general unbalanced GMM by Chen and Yang (2021b), from which we show that if the minimal separation $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$ and $K \lesssim \log(\gamma n)/\log\log(\gamma n)$, then

$$\mathbb{P}(B^c|B_\tau) \lesssim 1/\log^c(\gamma n).$$

In step 3, we use the nearest centroids $\bar{X}_1, \dots, \bar{X}_K$ estimated from step 2 for propagating the solution from the subsampled SDP to all data points that are not sampled in step 1. The key structure for the lift step to be successful is the independence between V and $X\setminus V$. In particular, the independence among $\bar{X}_k, \bar{X}_l, X_i, i \in [n]/T$ entails that under the minimal separation $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$, the error probability for the nearest-centroid procedure assigning $i \in \hat{G}_k$ via $\|X_i - \bar{X}_l\|^2 > \|X_i - \bar{X}_k\|^2$ for all $i \in G_k^* \setminus T$ vanishes

$$\mathbb{P}(A^c \cap B|B_\tau) \le K^2/n^c.$$

Combining the above three steps with the master bound (10), we conclude that the probability of exact recovery $\mathbb{P}(A) \ge 1 - C\log^{-c}(\gamma n)$ for large enough n, ensuring correctness of the SL approach.

A.3 Proof of Theorem 4.1

Step 1: reduction to subsampled data points. Let $T \subset [n]$ be the subsampled data point indices so that $V = (X_i)_{i\in T} \subset X$ and $m = |T|$. Let $R_k = G_k^* \cap T$. Define the events

$$\begin{aligned}
A :=& \Big\{\hat{G}_1 = G_1^*, \dots, \hat{G}_K = G_K^*\Big\},\\
B :=& \Big\{\hat{R}_1 = R_1, \dots, \hat{R}_K = R_K\Big\},\\
B_\tau :=& \Big\{(1-\tau)\frac{n\gamma}{K} \le |R_k| \le (1+\tau)\frac{n\gamma}{K},\ \forall k \in [K]\Big\},
\end{aligned}$$

where $\tau \in (0,1)$. Observe that

$$\begin{aligned}
\mathbb{P}(A^c) =& \mathbb{P}(A^c \cap (B \cap B_\tau)) + \mathbb{P}(A^c \cap (B^c \cup B_\tau^c))\\
\le& \mathbb{P}(A^c \cap B|B_\tau)\mathbb{P}(B_\tau) + \mathbb{P}(A^c \cap B^c) + \mathbb{P}(A^c \cap B_\tau^c)\\
\le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(A^c \cap B^c \cap B_\tau) + 2\mathbb{P}(B_\tau^c)\\
\le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(A^c \cap B^c|B_\tau) + 2\mathbb{P}(B_\tau^c)\\
\le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(B^c|B_\tau) + 2\mathbb{P}(B_\tau^c).
\end{aligned}$$

Thus to bound the error probability for exact recovery, it suffices to bound the three terms on the right-hand side of the last inequality. Since the subsampled data points V from $X = (X_1, \dots, X_n)$ are drawn with i.i.d. $\mathrm{Ber}(\gamma)$, we apply the Chernoff bound and the union bound to get

$$\mathbb{P}(B_\tau^c) \le 2K\exp\Big(-\frac{\tau^2 n\gamma}{3K}\Big).$$

Choosing $\tau = \sqrt{6K\log(n)/n\gamma}$, we have

$$\mathbb{P}(B_\tau^c) \le 2n^{-1}.$$

Step 2: exact recovery for subsampled data. Since the subsampling procedure is independent of the original X_i points, we can treat the $X_i \in V$ as the new cluster problem to apply Lemma A.1 with $T = \bigcup_{k=1}^{K} R_k$, $n = m$ and $n_k = m_k$, where $m_k = |R_k|$. In particular, if there exist constants $\tilde{\delta} > 0$ and $\beta \in (0,1)$ such that

$$\log m \geq \frac{(1-\beta)^2}{\beta^2}\frac{C_1 m}{M},$$
$$\tilde{\delta} \leq \frac{\beta^2}{(1-\beta)^2}\frac{C_2}{K}, \quad N \geq \frac{4(1+\tilde{\delta})^2}{\tilde{\delta}^2},$$

and

$$\Delta^2 \geq \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{p}{M\log m}+C_3 r_m}\right)\log m$$

with

$$r_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log m}\Big(\frac{\sqrt{p\log m}}{\underline{m}}+\frac{\log m}{\underline{m}}\Big),$$

where $\underline{m} = \min_{k\in[K]} m_k$ and $M = \min_{1\leq k\neq l\leq K} \frac{2m_l m_k}{m_l+m_k}$, then the SDP achieves exact recovery, i.e., $\hat{R}_k = R_k$, $\forall k \in [K]$, with probability at least $1 - C_4 K^2 m^{-\tilde{\delta}}$, where $C_i,\ i = 1,2,3,4$ are universal constants. Note that on event B_τ, we have

$$(1-\tau)n\gamma \leq m \leq (1+\tau)n\gamma,$$
$$\frac{2m_l m_k}{m_l+m_k} = \frac{2}{m_l^{-1}+m_k^{-1}} \geq (1-\tau)\frac{n\gamma}{K}.$$

Thus on the event B_τ, we can choose an upper bound Δ'^2 :

$$\Delta'^2 := \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{pK/((1-\tau)\gamma n)}{\log((1+\tau)\gamma n)}+C_3 r'_m}\right)\log((1+\tau)\gamma n)$$

with

$$r'_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log((1+\tau)\gamma n)}\left(\frac{K\sqrt{p\log((1+\tau)\gamma n)}}{(1-\tau)\gamma n}+\frac{K\log((1+\tau)\gamma n)}{(1-\tau)\gamma n}\right).$$

Note that $\tau = \sqrt{6K\log(n)/n\gamma} = o(1)$ under the assumption $\frac{K\log n}{n} = o(\gamma)$. Fix an $\alpha > 0$. By choosing small enough β and $\tilde{\delta}$ that may also also depend on α, we have for large enough n, if $K \leq C_1\frac{\log(\gamma n)}{\log\log(\gamma n)}$ for some constant C_1 depending on α and $\Delta^2 \geq (1+\alpha)\bar{\Delta}_\gamma^2$, where

$$\bar{\Delta}_\gamma^2 = 4\sigma^2\left(1+\sqrt{1+\frac{Kp}{\gamma n\log n}}\right)\log n,$$

then SDP achieves exact recovery with probability at least $1 - C_2(\log(\gamma n))^{-C_3}$, where C_2, C_3 depend only on α. Thus we conclude that

$$\mathbb{P}(B^c|B_\tau) \leq C_2(\log(\gamma n))^{-C_3}.$$

Remark A.3 (Lower bound for γ). It can be seen that the lower bound condition $\frac{K\log n}{n} = o(\gamma)$ for γ can be relaxed to $K = o(n\gamma/\log\log(\gamma n))$ given $K \leq C_1\frac{\log(\gamma n)}{\log\log(\gamma n)}$. i.e., we can think of $K \leq C_1\frac{\log(\gamma n)}{\log\log(\gamma n)}$ as another way to interpret the lower bound for γ.

Step 3: lift solution from sketched SDP to all the data points. Recall that the lift solution to all n data points is defined as

$$\hat{G}_k = \Big\{i \in [n]\setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \neq k\Big\} \cup \hat{R}_k,$$

where $\bar{X}_k = \sum_{j\in\hat{R}_k} X_j/m_k$ is the centroid of the k-th cluster output from the subsampled SDP. Since our goal in this step is to bound $\mathbb{P}(A^c \cap B|B_\tau)$, the subsequent analysis will be on the event B, that is $\hat{R}_k = R_k$ for all $k \in [K]$. Then we have $\bar{X}_k = \sum_{j\in R_k} X_j/m_k$ and

$$\hat{G}_k = \Big\{ i \in [n] \setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \neq k \Big\} \cup R_k.$$

Let $\mathcal{A}_{kl}^{(i)} = \Big\{ \|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > \xi \Big\}$, where $i \in G_k^* \backslash T$, where ξ is some number to be determined. Recall that $X_i = \mu_k + \epsilon_i, \forall i \in G_k^*$, where ϵ_i are i.i.d. $N(0, \sigma^2 I_p)$. Denote similarly $\bar{X}_k = \mu_k + \bar{\epsilon}_k$, where $\bar{\epsilon}_k = \sum_{j\in R_k} \epsilon_i/m_k$. For $i \in G_k^* \backslash T$, we note that $\epsilon_i, \bar{\epsilon}_k, \bar{\epsilon}_l$ are independent. We can write

$$\begin{aligned}
&\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 \\
=&\|\theta + \epsilon_i - \bar{\epsilon}_l\|^2 - \|\epsilon_i - \bar{\epsilon}_k\|^2 \\
=&\|\theta\|^2 + \|\bar{\epsilon}_l\|^2 - \|\bar{\epsilon}_k\|^2 - 2\langle\theta, \bar{\epsilon}_l\rangle + 2\langle\theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i\rangle,
\end{aligned}$$

where $\theta = \mu_k - \mu_l$. Set $\zeta_n = 2\log(Kn)$ and define

$$\begin{aligned}
\mathcal{B}_{kl,1}^{(i)} := \Big\{ &\|\bar{\epsilon}_l\|^2 \geq m_l^{-1}(p - 2\sqrt{p\zeta_n}), \\
&\|\bar{\epsilon}_k\|^2 \leq m_k^{-1}(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle\theta, \bar{\epsilon}_l\rangle \leq \sqrt{2m_l^{-1}\zeta_n}\|\theta\| \Big\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B}_{kl,2}^{(i)} := \Big\{ &\|\bar{\epsilon}_l - \bar{\epsilon}_k\|^2 \leq (m_l^{-1} + m_k^{-1})(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle\theta, \bar{\epsilon}_k - \bar{\epsilon}_l\rangle \leq \sqrt{2(m_l^{-1} + m_k^{-1})\zeta_n}\|\theta\| \Big\}.
\end{aligned}$$

Let $\mathcal{B}_{kl}^{(i)} = \mathcal{B}_{kl,1}^{(i)} \cup \mathcal{B}_{kl,2}^{(i)}$. Using the standard tail probability bound for χ^2 distribution (Laurent and Massart, 2000), we have $\mathbb{P}(\mathcal{B}_{kl}^{(i)^c}) \leq 5/(n^2K^2)$. Since

$$\langle\theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i\rangle | \{\bar{\epsilon}_l, \bar{\epsilon}_k\} \sim N(0, \|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2),$$

we have on the event $\mathcal{B}_{kl}^{(i)}$ that

$$\begin{aligned}
\mathcal{C}^* :&= \mathbb{P}\Big(2\langle\theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i\rangle \leq -(1-\beta)\|\theta\|^2 \Big| \bar{\epsilon}_k, \bar{\epsilon}_l\Big) \\
&= 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{\|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2}}\right) \\
&\leq 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}}\right),
\end{aligned}$$

where $\beta \in (0,1)$,

$$r_n'' = \|\theta\|^2 + 2\sqrt{2(m_l^{-1} + m_k^{-1})\zeta_n}\|\theta\| + (m_l^{-1} + m_k^{-1})(p + 2\sqrt{p\zeta_n} + 2\zeta_n).$$

Note that $\|\theta\|^2 \geq \Delta^2 \geq 8\log n$, which implies $\sqrt{2(m_l^{-1} + m_k^{-1})\zeta_n} \leq \|\theta\|\sqrt{2/M}$. Now we choose $\eta > 0$ such that $\frac{1+2\eta}{1+\eta} \geq 1 + 2\sqrt{2/M}$ (i.e., $M \geq 8(1+\eta^{-1})^2$). In order to have $\mathcal{C}^*$ be bounded by $n^{-(1+\eta)}$, it is sufficient to require that

$$1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}}\right) \leq 1 - \Phi(\sqrt{2(1+\eta)\log n}).$$

or further

$$\frac{(1-\beta)^2}{8(1+\eta)\log n}\|\theta\|^4 - (1+2\sqrt{2/M})\|\theta\|^2 - (p+2\sqrt{p\zeta_n}+2\zeta_n)(m_l^{-1}+m_k^{-1}) \geq 0.$$

A sufficient condition for the last display is

$$\Delta^2 \geq \frac{4\sigma^2(1+2\eta)}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+2\eta)}\frac{p}{M\log n}+2r_n'''}\right)\log n,$$

where

$$r_n''' = \frac{(1-\beta)^2}{(1+2\eta)\log n}\Big(\frac{\sqrt{p\log(nK)}}{\underline{m}}+\frac{\log(nK)}{\underline{m}}\Big).$$

Now if we put

$$\xi = \frac{m_k-m_l}{m_km_l}p+\beta\|\theta\|^2 - 4\sqrt{\frac{\log(nK)}{m_l}}\|\theta\| - 2\frac{m_k+m_l}{m_km_l}\sqrt{2p\log(nK)} - 4\frac{\log(nK)}{m_k},$$

then we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\|X_i-\bar{X}_l\|^2-\|X_i-\bar{X}_k\|^2>\xi,\\
&\qquad\qquad \forall i\in G_k^*\backslash T,\ \forall 1\leq k\neq l\leq K\Big\}^c\Big)\\
&=\mathbb{P}\Big(\bigcup_{i=1}^n\bigcup_{1\leq k\neq l\leq K}A_{kl}^{(i)^c}\Big)\\
&\leq\sum_{i=1}^n\sum_{1\leq k\neq l\leq K}\mathbb{P}\Big(A_{kl}^{(i)^c}\cap\mathcal{B}_{kl}^{(i)}\Big)+\mathbb{P}\Big(\mathcal{B}_{kl}^{(i)^c}\Big)\\
&\leq\sum_{i=1}^n\sum_{1\leq k\neq l\leq K}\mathbb{E}[\mathcal{C}^*1_{\mathcal{B}_{kl,2}^{(i)}}]+\frac{5}{n}\\
&\leq\frac{K^2}{n^\eta}+\frac{7}{n}.
\end{aligned}$$

Next we claim that $\xi>0$. Recall that on the event B_τ, we have $m_k\in[(1-\tau)m_*,(1+\tau)m_*]$, $1/M\in[\frac{1}{(1+\tau)m_*},\frac{1}{(1-\tau)m_*}]$, where $m_*=n\gamma/K$. Then,

$$\Big|\frac{m_k-m_l}{m_km_l}p\Big|\leq\frac{2\tau}{(1-\tau)^2}\frac{p}{m_*}\leq\frac{6p\sqrt{\log n}}{(1-\tau^2)m_*^{3/2}}.$$

Note that

$$\|\theta\|^2\geq\bar{\Delta}_\gamma^2\geq 4\sigma^2\left(1+\sqrt{1+\frac{p}{m_*\log n}}\right)\log n.$$

So if $p=O(\gamma n/K)^2)$, then

$$\Big|\frac{m_k-m_l}{m_km_l}p\Big|\leq\frac{\beta}{5}\|\theta\|^2$$

for large enough n. Similarly, we have

$$\begin{aligned}
&4\sqrt{\log(nK)/m_l}\|\theta\|\leq\frac{\beta}{5}\|\theta\|^2,\\
&2\frac{m_k+m_l}{m_km_l}\sqrt{2p\log(nK)}\leq\frac{\beta}{5}\|\theta\|^2,\\
&4m_k^{-1}\log(nK)\leq\frac{\beta}{5}\|\theta\|^2.
\end{aligned}$$

For $\alpha > 0$, we can choose small enough $\beta := \beta(\alpha, \sigma) > 0$ and $\eta := \eta(\alpha)$. Then for n large, we have if $\Delta^2 \geq (1+\alpha)\bar{\Delta}_\gamma^2$, then

$$\begin{aligned}
&\mathbb{P}(A^c \cap B | B_\tau) \\
=&\mathbb{P}\Big(\Big\{\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > 0, \forall i \in G_k^* \backslash T,\ \forall 1 \leq k \neq l \leq K\Big\}^c\Big) \\
\leq&\frac{K^2}{n^\eta}.
\end{aligned}$$

Now, combining all pieces together, we conclude that, for all n large enough,

$$\mathbb{P}(\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*) \geq 1 - C(\log(\gamma n))^{-c}.$$

A.4 Proof outline for Theorem 4.3

The overall strategy for proving Theorem 4.3 for unequal cluster size case is identical to the proof of Theorem 4.1, which can be divided into three steps. We will briefly talk about the difference here. And all the details are contained in the proof.

In step 1, we reduce the exact recovery problem on the entire n data points to the subsampled $m \approx n\gamma$ data points with K clusters with approximately γn_k many points for each cluster $\hat{G}_k$ that resembles the problem structure of the original SDP clustering problem. The parameter τ in the classical Chernoff bound now should be

$$\tau \asymp \sqrt{\log(n)/\underline{n}\gamma}.$$

In step 2, since the subsampling procedure is independent of the original data points, we can treat the subsampled data points $V = (X_i)_{i \in T}$ as a new clustering problem. Same as proof of Theorem 4.1, we establish the exact recovery guarantee on subsampled data using the separation upper bound proved for the general unbalanced GMM by Chen and Yang (2021b), from which we get the similar conditions. i.e., $\Delta^2 \geq (1+\alpha)\bar{\Delta}_\gamma'^2$ and $K \lesssim \log(\gamma n)/\log\log(\gamma n)$.

In step 3, we first get the subsampled clusters from step 2 and for each cluster, we randomly down-sample same size (the minimum one) of points to make the new clusters have the same sample size. Then we use the nearest centroids $\bar{X}_1, \ldots, \bar{X}_K$ of the new clusters for propagating the solution from the subsampled SDP to all data points that are not sampled in step 1. The independence among $\bar{X}_k, \bar{X}_l, X_i, i \in [n]/T$ entails that under the minimal separation $\Delta^2 \geq (1+\alpha)\bar{\Delta}_\gamma^2$, the error probability for the nearest-centroid procedure assigning $i \in \hat{G}_k$ via $\|X_i - \bar{X}_l\|^2 > \|X_i - \bar{X}_k\|^2$ for all $i \in G_k^* \setminus T$ vanishes.

A.5 Proof of Theorem 4.3

Step 1: reduction to subsampled data points. Let $T \subset [n]$ be the subsampled data point indices so that $V = (X_i)_{i \in T} \subset X$ and $m = |T|$. Let $n_k = |G_k^*|$, $R_k = G_k^* \cap T$. Define the events

$$\begin{aligned}
A :=&\Big\{\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*\Big\}, \\
B :=&\Big\{\hat{R}_1 = R_1, \ldots, \hat{R}_K = R_K\Big\}, \\
B_\tau :=&\Big\{(1-\tau)n_k\gamma \leq |R_k| \leq (1+\tau)n_k\gamma,\ \forall k \in [K]\Big\},
\end{aligned}$$

where $\tau \in (0,1)$. Observe that

$$\begin{aligned}
\mathbb{P}(A^c) =&\mathbb{P}(A^c \cap (B \cap B_\tau)) + \mathbb{P}(A^c \cap (B^c \cup B_\tau^c)) \\
\leq&\mathbb{P}(A^c \cap B | B_\tau)\mathbb{P}(B_\tau) + \mathbb{P}(A^c \cap B^c) + \mathbb{P}(A^c \cap B_\tau^c) \\
\leq&\mathbb{P}(A^c \cap B | B_\tau) + \mathbb{P}(A^c \cap B^c \cap B_\tau) + 2\mathbb{P}(B_\tau^c) \\
\leq&\mathbb{P}(A^c \cap B | B_\tau) + \mathbb{P}(A^c \cap B^c | B_\tau) + 2\mathbb{P}(B_\tau^c) \\
\leq&\mathbb{P}(A^c \cap B | B_\tau) + \mathbb{P}(B^c | B_\tau) + 2\mathbb{P}(B_\tau^c).
\end{aligned}$$

Thus to bound the error probability for exact recovery, it suffices to bound the three terms on the right-hand side of the last inequality. Since the subsampled data points V from $X = (X_1, \dots, X_n)$ are drawn with i.i.d. $\mathrm{Ber}(\gamma)$, we apply the Chernoff bound and the union bound to get

$$\mathbb{P}(B_\tau^c) \le 2K \exp\Big(- \frac{\tau^2 \underline{n}\gamma}{3}\Big).$$

Choosing $\tau = \sqrt{6\log(n)/\underline{n}\gamma}$, we have

$$\mathbb{P}(B_\tau^c) \le 2n^{-1}.$$

Step 2: exact recovery for subsampled data. Since the subsampling procedure is independent of the original X_i points, we can treat the $X_i \in V$ as the new cluster problem to apply Lemma A.1 with $T = \bigcup_{k=1}^K R_k$, $n = m$ and $n_k = m_k$, where $m_k = |R_k|$. In particular, if there exist constants $\tilde{\delta} > 0$ and $\beta \in (0,1)$ such that

$$\log m \ge \frac{(1-\beta)^2}{\beta^2}\frac{C_1 m}{M}, \quad \tilde{\delta} \le \frac{\beta^2}{(1-\beta)^2}\frac{C_2}{K}, \quad N \ge \frac{4(1+\tilde{\delta})^2}{\tilde{\delta}^2}$$

and

$$\Delta^2 \ge \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1 + \sqrt{1 + \frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{p}{M\log m} + C_3 r_m}\right)\log m$$

with

$$r_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log m}\Big(\frac{\sqrt{p\log m}}{\underline{m}} + \frac{\log m}{\underline{m}}\Big),$$

where $\underline{m} = \min_{k\in[K]} m_k$ and $M = \min_{1\le k\ne l\le K} \frac{2m_l m_k}{m_l + m_k}$, then the SDP achieves exact recovery, i.e., $\hat{R}_k = R_k$, $\forall k \in [K]$, with probability at least $1 - C_4 K^2 m^{-\tilde{\delta}}$, where C_i, $i = 1,2,3,4$ are universal constants. Note that on event B_τ, we have

$$\begin{aligned}
&(1-\tau)n\gamma \le m \le (1+\tau)n\gamma,\\
&\frac{2m_l m_k}{m_l + m_k} = \frac{2}{m_l^{-1} + m_k^{-1}} \ge (1-\tau)\underline{n}\gamma.
\end{aligned}$$

Thus on the event B_τ, we can choose an upper bound Δ'^2 :

$$\Delta'^2 := \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1 + \sqrt{1 + \frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{p/((1-\tau)\gamma\underline{n})}{\log((1+\tau)\gamma n)} + C_3 r'_m}\right)\log((1+\tau)\gamma n)$$

with

$$r'_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log((1+\tau)\gamma n)}\left(\frac{\sqrt{p\log((1+\tau)\gamma n)}}{(1-\tau)\gamma\underline{n}} + \frac{\log((1+\tau)\gamma n)}{(1-\tau)\gamma\underline{n}}\right).$$

Note that $\tau = \sqrt{6\log(n)/\underline{n}\gamma} = o(1)$ under the assumption $\frac{\log n}{\underline{n}} = o(\gamma)$. Fix an $\alpha > 0$. By choosing small enough β and $\tilde{\delta}$ that may also also depend on α, we have for large enough n, if $K \le C_1 \frac{\log(\gamma n)}{\log\log(\gamma n)}$, $\underline{n} \ge C_2 n/\log(\gamma n)$ for some constant C_1, C_2 depending on α and $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$, where

$$\bar{\Delta}_\gamma^2 = 4\sigma^2\left(1 + \sqrt{1 + \frac{p}{\gamma\underline{n}\log n}}\right)\log n,$$

then SDP achieves exact recovery with probability at least $1 - C_3(\log(\gamma n))^{-C_4}$, where C_3, C_4 depend only on α. Thus we conclude that

$$\mathbb{P}(B^c | B_\tau) \le C_2(\log(\gamma n))^{-C_3}.$$

Step 3: lift solution from sketched SDP to all the data points.

Recall that the lift solution to all n data points is defined as

$$\hat{G}_k = \Big\{ i \in [n] \setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \neq k \Big\} \cup \hat{R}_k,$$

where $\bar{X}_k = |\tilde{R}_k|^{-1} \sum_{j \in \tilde{R}_k}$ is the revised centroid of the k-th cluster output from the subsampled SDP. And $\tilde{R}_k$ is a randomly sampled subset of $\hat{R}_k$ with equal size $\underline{m}$. Since our goal in this step is to bound $\mathbb{P}(A^c \cap B | B_\tau)$, the subsequent analysis will be on the event B, that is $\hat{R}_k = R_k$ for all $k \in [K]$.

Then we have $\bar{X}_k = \sum_{j \in \tilde{R}_k \subseteq R_k} X_j / \underline{m}$ and

$$\hat{G}_k = \Big\{ i \in [n] \setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \neq k \Big\} \cup R_k.$$

Let $\mathcal{A}_{kl}^{(i)} = \Big\{ \|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > \xi \Big\}$, where $i \in G_k^* \setminus T$, where ξ is some number to be determined. If we further make the analysis on $T' \subseteq T$, let $\mathcal{A}_{kl,t'}^{(i)} = \mathcal{A}_{kl}^{(i)} \cap \{T' = t'\}$, where $t' = \bigsqcup_{k=1}^K R_k'$ is any realization of T'. Recall that $X_i = \mu_k + \epsilon_i, \forall i \in G_k^*$, where ϵ_i are i.i.d. $N(0, \sigma^2 I_p)$. Denote similarly $\bar{X}_k = \mu_k + \bar{\epsilon}_k$, where $\bar{\epsilon}_k = \sum_{j \in R_k} \epsilon_i / \underline{m}$. For $i \in G_k^* \setminus T$, we note that $\epsilon_i, \bar{\epsilon}_k, \bar{\epsilon}_l$ are independent. We can write

$$\begin{aligned}
&\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 \\
=&\|\theta + \epsilon_i - \bar{\epsilon}_l\|^2 - \|\epsilon_i - \bar{\epsilon}_k\|^2 \\
=&\|\theta\|^2 + \|\bar{\epsilon}_l\|^2 - \|\bar{\epsilon}_k\|^2 - 2\langle \theta, \bar{\epsilon}_l \rangle + 2\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle,
\end{aligned}$$

where $\theta = \mu_k - \mu_l$. Set $\zeta_n = 2\log(Kn)$ and define

$$\begin{aligned}
\mathcal{B}_{kl,1}^{(i)} := \Big\{ &\|\bar{\epsilon}_l\|^2 \geq \underline{m}^{-1}(p - 2\sqrt{p\zeta_n}), \\
&\|\bar{\epsilon}_k\|^2 \leq \underline{m}^{-1}(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle \theta, \bar{\epsilon}_l \rangle \leq \sqrt{2\underline{m}^{-1}\zeta_n}\|\theta\| \Big\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B}_{kl,2}^{(i)} := \Big\{ &\|\bar{\epsilon}_l - \bar{\epsilon}_k\|^2 \leq 2\underline{m}^{-1})(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle \theta, \bar{\epsilon}_k - \bar{\epsilon}_l \rangle \leq 2\sqrt{(\underline{m}^{-1})\zeta_n}\|\theta\| \Big\}.
\end{aligned}$$

Let $\mathcal{B}_{kl}^{(i)} = \mathcal{B}_{kl,1}^{(i)} \cup \mathcal{B}_{kl,2}^{(i)}$. Using the standard tail probability bound for χ^2 distribution (Laurent and Massart, 2000), we have $\mathbb{P}(\mathcal{B}_{kl}^{(i)^c}) \leq 5/(n^2K^2)$. Since

$$\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle | \{\bar{\epsilon}_l, \bar{\epsilon}_k\} \sim N(0, \|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2),$$

we have on the event $\mathcal{B}_{kl}^{(i)}$ that

$$\begin{aligned}
\mathcal{C}^* :&= \mathbb{P}\Big(2\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle \leq -(1-\beta)\|\theta\|^2 \Big| \bar{\epsilon}_k, \bar{\epsilon}_l \Big) \\
&= 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{\|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2}} \right) \\
&\leq 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}} \right),
\end{aligned}$$

where $\beta \in (0,1)$,

$$r_n'' = \|\theta\|^2 + 4\sqrt{\underline{m}^{-1}\zeta_n}\|\theta\| + 2\underline{m}^{-1}(p + 2\sqrt{p\zeta_n} + 2\zeta_n).$$

Note that $\|\theta\|^2 \geq \Delta^2 \geq 8\log n$, which implies $2\sqrt{\underline{m}^{-1}\zeta_n} \leq \|\theta\|\sqrt{2/\underline{m}}$. Now we choose $\eta > 0$ such that $\frac{1+2\eta}{1+\eta} \geq 1 + 2\sqrt{2/\underline{m}}$ (i.e., $\underline{m} \geq 8(1+\eta^{-1})^2$). In order to have $\mathcal{C}^*$ be bounded by $n^{-(1+\eta)}$, it is sufficient to require that

$$1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}}\right) \leq 1 - \Phi(\sqrt{2(1+\eta)\log n}).$$

or further

$$\frac{(1-\beta)^2}{8(1+\eta)\log n}\|\theta\|^4 - (1+2\sqrt{2/M})\|\theta\|^2 - 2(p + 2\sqrt{p\zeta_n} + 2\zeta_n)\underline{m}^{-1} \geq 0.$$

A sufficient condition for the last display is

$$\Delta^2 \geq \frac{4\sigma^2(1+2\eta)}{(1-\beta)^2}\left(1 + \sqrt{1 + \frac{(1-\beta)^2}{(1+2\eta)}\frac{p}{\underline{m}\log n} + 2r_n'''}\right)\log n,$$

where

$$r_n''' = \frac{(1-\beta)^2}{(1+2\eta)\log n}\Big(\frac{\sqrt{p\log(nK)}}{\underline{m}} + \frac{\log(nK)}{\underline{m}}\Big).$$

Now if we put

$$\xi = \beta\|\theta\|^2 - 4\sqrt{\frac{\log(nK)}{\underline{m}}}\|\theta\| - 4\underline{m}^{-1}\sqrt{2p\log(nK)} - 4\frac{\log(nK)}{\underline{m}},$$

notice that

$$\mathbb{P}\Big(A_{kl}^{(i)^c}\Big) = \sum_{t'\in\mathcal{T}'}\mathbb{P}\Big(\mathcal{A}_{kl,t'}^{(i)}\Big)\cdot\mathbb{P}\Big(T' = t'\Big) \leq \max_{t'\in\mathcal{T}'}\mathbb{P}\Big(\mathcal{A}_{kl,t'}^{(i)}\Big),$$

where $\mathcal{T}'$ is all the possible subset of T s.t. $|\tilde{R}_i| = \underline{m}$. Then we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > \xi, \\
&\qquad\qquad \forall i \in G_k^*\backslash T,\ \forall 1 \leq k \neq l \leq K\Big\}^c\Big) \\
&= \mathbb{P}\Big(\bigcup_{i=1}^n \bigcup_{1\leq k\neq l\leq K} A_{kl}^{(i)^c}\Big) \\
&\leq \sum_{i=1}^n \sum_{1\leq k\neq l\leq K} \mathbb{P}\Big(A_{kl}^{(i)^c}\Big) \\
&\leq \sum_{i=1}^n \sum_{1\leq k\neq l\leq K} \max_{t'\in\mathcal{T}'}\mathbb{P}\Big(\mathcal{A}_{kl,t'}^{(i)} \cap \mathcal{B}_{kl}^{(i)}\Big) + \mathbb{P}\Big(\mathcal{B}_{kl}^{(i)^c}\Big) \\
&\leq \sum_{i=1}^n \sum_{1\leq k\neq l\leq K} \mathbb{E}[\mathcal{C}^* 1_{\mathcal{B}_{kl,2}^{(i)}}] + \frac{5}{n} \\
&\leq \frac{K^2}{n^\eta} + \frac{7}{n}.
\end{aligned}$$

Next we claim that $\xi > 0$. Recall that on the event B_τ, we have $m_k \in [(1-\tau)\gamma n_k, (1+\tau)\gamma n_k]$, $1/\underline{m} \in [\frac{1}{(1+\tau)\gamma\underline{n}}, \frac{1}{(1-\tau)\gamma\underline{n}}]$. Note that

$$\|\theta\|^2 \geq \bar{\Delta}_\gamma^2 \geq 4\sigma^2\left(1 + \sqrt{1 + \frac{p}{\gamma\underline{n}\log n}}\right)\log n.$$

So if $\gamma\underline{n} \to \infty$ as $n \to \infty, n$ large, then we have

$$4\sqrt{\log(nK)/\underline{m}}\|\theta\| \le \frac{\beta}{5}\|\theta\|^2,$$
$$4\underline{m}^{-1}\sqrt{2p\log(nK)} \le \frac{\beta}{5}\|\theta\|^2,$$
$$4\underline{m}^{-1}\log(nK) \le \frac{\beta}{5}\|\theta\|^2.$$

For $\alpha > 0$, we can choose small enough $\beta := \beta(\alpha, \sigma) > 0$ and $\eta := \eta(\alpha)$. Then for n large, we have if $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$, then

$$\begin{aligned}
&\mathbb{P}(A \cap B | B_\tau) \\
=&\mathbb{P}\Big(\Big\{\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > 0, \forall i \in G_k^* \backslash T,\ \forall 1 \le k \ne l \le K\Big\}^c\Big) \\
\le&\frac{K^2}{n^\eta}.
\end{aligned}$$

Now, combining all pieces together, we conclude that, for all n large enough,

$$\mathbb{P}(\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*) \ge 1 - C(\log(\gamma n))^{-c}.$$

A.6 Proof outline for Theorem 4.4

The overall strategy for proving Theorem 4.4 for unequal cluster size case is identical to the proof of Theorem 4.1, which can be divided into three steps. Again we will briefly talk about the difference here.

In step 1, we reduce the exact recovery problem on the entire n data points to the subsampled $m \approx n\gamma$ data points with K clusters with approximately equal size that resembles the problem structure of the original SDP clustering problem. Here we use the weighted sampling ratio w_i for each point $i \in [n]$, where majority of them should be around the true weights $w_i^*,\ i \in [n]$. We will apply the classical Chernoff bound by setting appropriate τ through the definition of (ϵ, δ) pairs. i.e.,

$$\tau \asymp \sqrt{\log(n)/\gamma\underline{n}} + \epsilon + \delta n/\underline{n}.$$

In step 2, we assume $\tau = o(1)$ by setting each summand of τ to be $o(1)$. And we set the same conditions for minimal separation $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$ and $K \lesssim \log(\gamma n)/\log\log(\gamma n)$.

In step 3, we use the nearest centroids $\bar{X}_1, \ldots, \bar{X}_K$ estimated from step 2 for propagating the solution from the subsampled SDP to all data points that are not sampled in step 1. The discussion here is similar to the proof of Theorem 4.1. The only difference here is that now we bound the size difference by τ through (ϵ, δ) pairs. i.e.,

$$|m_k - m_l| < 2\tau \asymp \sqrt{\log(n)/\gamma\underline{n}} + \epsilon + \delta n/\underline{n},$$

where $k \ne l \in [n]$.

A.7 Proof of Theorem 4.4

Step 1: reduction to subsampled data points. Let $T \subset [n]$ be the subsampled data point indices so that $V = (X_i)_{i \in T} \subset X$ and $m = |T|$. Let $R_k = G_k^* \cap T$. Define the events

$$\begin{aligned}
A :=&\Big\{\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*\Big\}, \\
B :=&\Big\{\hat{R}_1 = R_1, \ldots, \hat{R}_K = R_K\Big\}, \\
B_\tau :=&\Big\{(1-\tau)\frac{n\gamma}{K} \le |R_k| \le (1+\tau)\frac{n\gamma}{K},\ \forall k \in [K]\Big\}, \\
B_{\tau_1} :=&\Big\{(1-\tau_1)m_k^* \le |R_k| \le (1+\tau_1)m_k^*,\ \forall k \in [K]\Big\},
\end{aligned}$$

where $m_k^* = \sum_{i\in G_k^*} w_i, \tau, \tau_1 \in (0,1)$. Observe that

$$\begin{aligned}\mathbb{P}(A^c) =& \mathbb{P}(A^c \cap (B \cap B_\tau)) + \mathbb{P}(A^c \cap (B^c \cup B_\tau^c)) \\ \le& \mathbb{P}(A^c \cap B|B_\tau)\mathbb{P}(B_\tau) + \mathbb{P}(A^c \cap B^c) + \mathbb{P}(A^c \cap B_\tau^c) \\ \le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(A^c \cap B^c \cap B_\tau) + 2\mathbb{P}(B_\tau^c) \\ \le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(A^c \cap B^c|B_\tau) + 2\mathbb{P}(B_\tau^c) \\ \le& \mathbb{P}(A^c \cap B|B_\tau) + \mathbb{P}(B^c|B_\tau) + 2\mathbb{P}(B_\tau^c).\end{aligned}$$

Thus to bound the error probability for exact recovery, it suffices to bound the three terms on the right-hand side of the last inequality. Since the subsampled data points V from $X = (X_1, \ldots, X_n)$ are drawn with i.i.d. Ber(γ), we apply the Chernoff bound and the union bound to get

$$\mathbb{P}(B_{\tau_1}^c) \le 2 \sum_{k\in[K]} \exp\Big(-\frac{\tau^2 m_k^*}{3}\Big).$$

Choosing $\tau_1 = \sqrt{6\log(n)/\underline{m}^*}$, we have

$$\mathbb{P}(B_{\tau_1}^c) \le 2n^{-1}.$$

If we choose $\tau = \tau_1 + \epsilon + \delta(1+\epsilon)\bar{w}^* K/\gamma$, then

$$\mathbb{P}(B_\tau^c) \le \mathbb{P}(B_{\tau_1}^c) \le 2n^{-1}.$$

Step 2: exact recovery for subsampled data. Since the subsampling procedure is independent of the original X_i points, we can treat the $X_i \in V$ as the new cluster problem to apply Lemma A.1 with $T = \bigcup_{k=1}^K R_k$, $n = m$ and $n_k = m_k$, where $m_k = |R_k|$. In particular, if there exist constants $\tilde{\delta} > 0$ and $\beta \in (0,1)$ such that

$$\log m \ge \frac{(1-\beta)^2}{\beta^2}\frac{C_1 m}{M}, \quad \tilde{\delta} \le \frac{\beta^2}{(1-\beta)^2}\frac{C_2}{K}, \quad N \ge \frac{4(1+\tilde{\delta})^2}{\tilde{\delta}^2}$$

and

$$\Delta^2 \ge \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1 + \sqrt{1 + \frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{p}{M\log m} + C_3 r_m}\right)\log m$$

with

$$r_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log m}\Big(\frac{\sqrt{p\log m}}{\underline{m}} + \frac{\log m}{\underline{m}}\Big),$$

where $\underline{m} = \min_{k\in[K]} m_k$ and $M = \min_{1\le k\ne l\le K} \frac{2m_l m_k}{m_l + m_k}$, then the SDP achieves exact recovery, i.e., $\hat{R}_k = R_k$, $\forall k \in [K]$, with probability at least $1 - C_4 K^2 m^{-\tilde{\delta}}$, where C_i, $i = 1,2,3,4$ are universal constants. Note that on event B_τ, we have

$$\begin{aligned}&(1-\tau)n\gamma \le m \le (1+\tau)n\gamma, \\ &\frac{2m_l m_k}{m_l + m_k} = \frac{2}{m_l^{-1} + m_k^{-1}} \ge (1-\tau)\frac{n\gamma}{K}.\end{aligned}$$

Thus on the event B_τ, we can choose an upper bound Δ'^2 :

$$\Delta'^2 := \frac{4\sigma^2(1+2\tilde{\delta})}{(1-\beta)^2}\left(1 + \sqrt{1 + \frac{(1-\beta)^2}{(1+\tilde{\delta})}\frac{pK/((1-\tau)\gamma n)}{\log((1+\tau)\gamma n)} + C_3 r'_m}\right)\log((1+\tau)\gamma n)$$

with

$$r'_m = \frac{(1-\beta)^2}{(1+\tilde{\delta})\log((1+\tau)\gamma n)}\left(\frac{K\sqrt{p\log((1+\tau)\gamma n)}}{(1-\tau)\gamma n} + \frac{K\log((1+\tau)\gamma n)}{(1-\tau)\gamma n}\right).$$

Note that $\tau = \tau_1 + \epsilon + \delta(1+\epsilon)\bar{w}^* K/\gamma = o(1)$ under the assumption $\tau_1 = o(1), \epsilon = o(1), \delta(1+\epsilon)\bar{w}^* K/\gamma = o(1)$. i.e. $\frac{K \log n}{n} = o(\gamma), \delta = o(\underline{n}/n), \epsilon = o(1)$.Fix an $\alpha > 0$. By choosing small enough β and δ that may also also depend on α, we have for large enough n, if $K \le C_1 \frac{\log(\gamma n)}{\log\log(\gamma n)}$ for some constant C_1 depending on α and $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$, where

$$\bar{\Delta}_\gamma^2 = 4\sigma^2 \left(1 + \sqrt{1 + \frac{Kp}{\gamma n \log n}}\right) \log n,$$

then SDP achieves exact recovery with probability at least $1 - C_2(\log(\gamma n))^{-C_3}$, where C_2, C_3 depend only on α. Thus we conclude that

$$\mathbb{P}(B^c | B_\tau) \le C_2(\log(\gamma n))^{-C_3}.$$

Step 3: lift solution from sketched SDP to all the data points. Recall that the lift solution to all n data points is defined as

$$\hat{G}_k = \left\{ i \in [n] \setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \ne k \right\} \cup \hat{R}_k,$$

where $\bar{X}_k = \sum_{j \in \hat{R}_k} X_j / m_k$ is the centroid of the k-th cluster output from the subsampled SDP. Since our goal in this step is to bound $\mathbb{P}(A^c \cap B | B_\tau)$, the subsequent analysis will be on the event B, that is $\hat{R}_k = R_k$ for all $k \in [K]$. Then we have $\bar{X}_k = \sum_{j \in R_k} X_j / m_k$ and

$$\hat{G}_k = \left\{ i \in [n] \setminus T : \|X_i - \bar{X}_k\| < \|X_i - \bar{X}_l\|, \forall l \ne k \right\} \cup R_k.$$

Let $\mathcal{A}_{kl}^{(i)} = \left\{ \|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > \xi \right\}$, where $i \in G_k^* \setminus T$, where ξ is some number to be determined. Recall that $X_i = \mu_k + \epsilon_i, \forall i \in G_k^*$, where ϵ_i are i.i.d. $N(0, \sigma^2 I_p)$. Denote similarly $\bar{X}_k = \mu_k + \bar{\epsilon}_k$, where $\bar{\epsilon}_k = \sum_{j \in R_k} \epsilon_i / m_k$. For $i \in G_k^* \setminus T$, we note that $\epsilon_i, \bar{\epsilon}_k, \bar{\epsilon}_l$ are independent. We can write

$$\begin{aligned}
&\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 \\
=&\|\theta + \epsilon_i - \bar{\epsilon}_l\|^2 - \|\epsilon_i - \bar{\epsilon}_k\|^2 \\
=&\|\theta\|^2 + \|\bar{\epsilon}_l\|^2 - \|\bar{\epsilon}_k\|^2 - 2\langle \theta, \bar{\epsilon}_l \rangle + 2\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle,
\end{aligned}$$

where $\theta = \mu_k - \mu_l$. Set $\zeta_n = 2\log(Kn)$ and define

$$\begin{aligned}
\mathcal{B}_{kl,1}^{(i)} := \Big\{ &\|\bar{\epsilon}_l\|^2 \ge m_l^{-1}(p - 2\sqrt{p\zeta_n}), \\
&\|\bar{\epsilon}_k\|^2 \le m_k^{-1}(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle \theta, \bar{\epsilon}_l \rangle \le \sqrt{2 m_l^{-1} \zeta_n} \|\theta\| \Big\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{B}_{kl,2}^{(i)} := \Big\{ &\|\bar{\epsilon}_l - \bar{\epsilon}_k\|^2 \le (m_l^{-1} + m_k^{-1})(p + 2\sqrt{p\zeta_n} + 2\zeta_n), \\
&\langle \theta, \bar{\epsilon}_k - \bar{\epsilon}_l \rangle \le \sqrt{2(m_l^{-1} + m_k^{-1})\zeta_n} \|\theta\| \Big\}.
\end{aligned}$$

Let $\mathcal{B}_{kl}^{(i)} = \mathcal{B}_{kl,1}^{(i)} \cup \mathcal{B}_{kl,2}^{(i)}$. Using the standard tail probability bound for χ^2 distribution (Laurent and Massart, 2000), we have $\mathbb{P}(\mathcal{B}_{kl}^{(i)^c}) \le 5/(n^2 K^2)$. Since

$$\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle | \{\bar{\epsilon}_l, \bar{\epsilon}_k\} \sim N(0, \|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2),$$

we have on the event $\mathcal{B}_{kl}^{(i)}$ that

$$\begin{aligned}
\mathcal{C}^* :&= \mathbb{P}\Big(2\langle \theta - \bar{\epsilon}_l + \bar{\epsilon}_k, \epsilon_i \rangle \le -(1-\beta)\|\theta\|^2 \Big| \bar{\epsilon}_k, \bar{\epsilon}_l \Big) \\
&= 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{\|\theta - \bar{\epsilon}_l + \bar{\epsilon}_k\|^2}} \right) \\
&\le 1 - \Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}} \right),
\end{aligned}$$

where $\beta \in (0,1)$,

$$r_n'' = \|\theta\|^2 + 2\sqrt{2(m_l^{-1}+m_k^{-1})\zeta_n}\|\theta\| + (m_l^{-1}+m_k^{-1})(p+2\sqrt{p\zeta_n}+2\zeta_n).$$

Note that $\|\theta\|^2 \geq \Delta^2 \geq 8\log n$, which implies $\sqrt{2(m_l^{-1}+m_k^{-1})\zeta_n} \leq \|\theta\|\sqrt{2/M}$. Now we choose $\eta > 0$ such that $\frac{1+2\eta}{1+\eta} \geq 1+2\sqrt{2/M}$ (i.e., $M \geq 8(1+\eta^{-1})^2$). In order to have $\mathcal{C}^*$ be bounded by $n^{-(1+\eta)}$, it is sufficient to require that

$$1-\Phi\left(\frac{(1-\beta)\|\theta\|^2}{2\sqrt{r_n''}}\right) \leq 1-\Phi(\sqrt{2(1+\eta)\log n}).$$

or further

$$\frac{(1-\beta)^2}{8(1+\eta)\log n}\|\theta\|^4 - (1+2\sqrt{2/M})\|\theta\|^2 - (p+2\sqrt{p\zeta_n}+2\zeta_n)(m_l^{-1}+m_k^{-1}) \geq 0.$$

A sufficient condition for the last display is

$$\Delta^2 \geq \frac{4\sigma^2(1+2\eta)}{(1-\beta)^2}\left(1+\sqrt{1+\frac{(1-\beta)^2}{(1+2\eta)}\frac{p}{M\log n}+2r_n'''}\right)\log n,$$

where

$$r_n''' = \frac{(1-\beta)^2}{(1+2\eta)\log n}\Big(\frac{\sqrt{p\log(nK)}}{\underline{m}}+\frac{\log(nK)}{\underline{m}}\Big).$$

Now if we put

$$\xi = \frac{m_k-m_l}{m_k m_l}p + \beta\|\theta\|^2 - 4\sqrt{\frac{\log(nK)}{m_l}}\|\theta\| - 2\frac{m_k+m_l}{m_k m_l}\sqrt{2p\log(nK)} - 4\frac{\log(nK)}{m_k},$$

then we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\|X_i-\bar{X}_l\|^2 - \|X_i-\bar{X}_k\|^2 > \xi, \\
&\qquad\qquad \forall i \in G_k^*\backslash T,\ \forall 1 \leq k \neq l \leq K\Big\}^c\Big) \\
&= \mathbb{P}\Big(\bigcup_{i=1}^n \bigcup_{1\leq k\neq l\leq K} A_{kl}^{(i)^c}\Big) \\
&\leq \sum_{i=1}^n \sum_{1\leq k\neq l\leq K} \mathbb{P}\Big(A_{kl}^{(i)^c} \cap \mathcal{B}_{kl}^{(i)}\Big) + \mathbb{P}\Big(\mathcal{B}_{kl}^{(i)^c}\Big) \\
&\leq \sum_{i=1}^n \sum_{1\leq k\neq l\leq K} \mathbb{E}[\mathcal{C}^* 1_{\mathcal{B}_{kl,2}^{(i)}}] + \frac{5}{n} \\
&\leq \frac{K^2}{n^\eta} + \frac{7}{n}.
\end{aligned}$$

Next we claim that $\xi > 0$. Recall that on the event B_τ, we have $m_k \in [(1-\tau)m_*, (1+\tau)m_*]$, $1/M \in [\frac{1}{(1+\tau)m_*}, \frac{1}{(1-\tau)m_*}]$, where $m_* = n\gamma/K$. Then,

$$g\Big|\frac{m_k-m_l}{m_k m_l}p\Big| \leq \frac{2\tau}{(1-\tau)^2}\frac{p}{m_*} \leq \frac{6p\sqrt{\log n}}{(1-\tau)^2 m_*\sqrt{\underline{m}^*}} + \frac{2\epsilon}{(1-\tau)^2}\frac{p}{m_*} + \frac{2(\lambda(1+\epsilon)\bar{w}^*K/\gamma)}{(1-\tau)^2}\frac{p}{m_*}.$$

Note that $\underline{m}^* \geq m_*(1-\epsilon) - \delta n(1+\epsilon)\bar{w}^*$,

$$\|\theta\|^2 \geq \bar{\Delta}_\gamma^2 \geq 4\sigma^2\left(1+\sqrt{1+\frac{p}{m_*\log n}}\right)\log n.$$

So if $p = O(\gamma n/K)^2), \epsilon = O(\frac{\gamma n}{Kp}\log n), \delta = O(\sqrt{\gamma n/p}\cdot \underline{n}/n)$ then

$$\Big|\frac{m_k - m_l}{m_k m_l}p\Big| \le \frac{\beta}{5}\|\theta\|^2$$

for large enough n. Similarly, we have

$$\begin{aligned}
&4\sqrt{\log(nK)/m_l}\|\theta\| \le \frac{\beta}{5}\|\theta\|^2,\\
&2\frac{m_k + m_l}{m_k m_l}\sqrt{2p\log(nK)} \le \frac{\beta}{5}\|\theta\|^2,\\
&4m_k^{-1}\log(nK) \le \frac{\beta}{5}\|\theta\|^2.
\end{aligned}$$

For $\alpha > 0$, we can choose small enough $\beta := \beta(\alpha, \sigma) > 0$ and $\eta := \eta(\alpha)$. Then for n large, we have if $\Delta^2 \ge (1+\alpha)\bar{\Delta}_\gamma^2$, then

$$\begin{aligned}
&\mathbb{P}(A \cap B | B_\tau)\\
=&\mathbb{P}\Big(\Big\{\|X_i - \bar{X}_l\|^2 - \|X_i - \bar{X}_k\|^2 > 0, \forall i \in G_k^* \backslash T,\ \forall 1 \le k \ne l \le K\Big\}^c\Big)\\
\le&\frac{K^2}{n^\eta}.
\end{aligned}$$

Now, combining all pieces together, we conclude that, for all n large enough,

$$\mathbb{P}(\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*) \ge 1 - C(\log(\gamma n))^{-c}.$$

B FULL NUMERICAL EXPERIMENT RESULTS

In this section, we show the full simulation results. The error rates and running times on log-scale are shown in Figures 3 to 18 (with captions describing the simulation setup). The baseline setup is $\lambda^* = 1.2, p = 1000, K = 4, \gamma = 0.1, n = 2000$ by default, expect when γ is changing, we use $n = 10000$. For simplicity, we set all the distance between centers to be equal. i.e., $\|\mu_k - \mu_l\|^2 = \Delta^2$, $\forall 1 \le k \ne l \le K$. In particular, we arrange all the centers on vertices of a regular simplex. i.e., $\mu_l = \Delta/\sqrt{2}\cdot e_l, \forall l \in [K]$, where e_l is a vector that the l-th component is 1 and 0 otherwise. The variance of Gaussian distributions is chosen to be 1. The detailed explanation for methods $M_0 - M_5$ and O can be found in Section 5.

Figures 3 to 9 are under the setting of equal cluster size. Figures 10 to 16 are for the unequal cluster size case. In particular, Figures 3 to 6, Figures 10 to 13 are under the separation condition from the theoretical cutoff $\bar{\Delta}_*^2$ in (4); the separation for Figures 7 to 9 is the theoretical cutoff $\bar{\Delta}_\gamma^2$ for SL methods; and the separation for Figures 14 to 16 is the theoretical cutoff $\bar{\Delta}_\gamma^{'2}$ for BCSL methods. Figures 3 to 6 correspond to Figure 1 and Figure 2 in Section 5. By comparing Figures 3 to 6 (equal cluster size) to Figures 10 to 13 (unequal cluster size), we can see that when p changes, M_2 and M_3 which are the improved methods aiming at handling large p unequal cluster size settings have better performance than M_1, which are consistent with our theory. Moreover, M_5, the multi-round WSL (at 4-th round), is optimal among those methods in Figure 10. For the time cost, we can still see that SL methods have the same linear $O(n)$ complexity as the fast K-means++ (with difference only occurring in the leading constant) for unequal cluster size from Figure 11 corresponding to Figure 2 in Section 5. We consider two threshold settings (Figures 7 to 9 and Figures 14 to 16) in order to show that the thresholds $\bar{\Delta}_\gamma^2$ for SL methods and $\bar{\Delta}_\gamma^{'2}$ for BCSL methods are nearly optimal in the sense that they are neither too large that all methods can achieve exact recovery easily nor too small that we can observe the exact recoveries in most cases, which can be found in those figures.

We would like to make a remark on the initial weights estimated by the K-means++ algorithm, which is the default K-means clustering algorithm in Matlab. From Table 1, we see that as dimension p increases, the fraction $(1-\delta)$ of "good" weights within the ϵ-distortion level $[(1-\epsilon)w_i^*, (1+\epsilon)w_i^*]$ decreases. This is also reflected in Figure 17, where the error rate of cluster label recovery is almost constant (N_2 purple curve) when the separation signal is slightly above the cutoff value of exact recovery. With such a warm-start given by K-means++, our WSL can improve the recovery error rate in one iteration (N_3 black curve). Moreover, if we use the ideal weights,

then the WSL can achieve the exact recovery (N_1 red curve in Figure 17). The one-iteration WSL can be viewed as the multi-round WSL with one round, we can see from Figure 18 that after 3-4 rounds, the multi-round WSL refines the clustering errors and can eventually achieve the exact recovery as if we initialize with the ideal weights. This can be explained by comparing Table 1 with Table 2 to 4, where we can see the decreasing trends for δ, which will become 0 eventually for all ϵ we set on the grid. This shows that the weights for R_4 we get from iteration are fairly close to or identical to 0. On the other hand it can be well explained by Figure 19, which is the corresponding averaged relative weight difference of R_1 to R_5 in Figure 18. Averaged relative weight difference is defined by $\frac{1}{n}\sum_{j\in[n]}|w_j - w_j^*|/w_j^*$. where w_j, w_j^* is defined in Section 3.3. R_i stands for the i-th round of MR-WSL corresponding to Figure 18. By comparing Figure 18 and Figure 19 we can see similar trends between the relative weight difference for R_{i+1} and the error rate for R_i when various parameters change. The relative weight difference has a decreasing trend as we perform more rounds, which accounts for the improvement of MR-WSL shown in Figure 18.

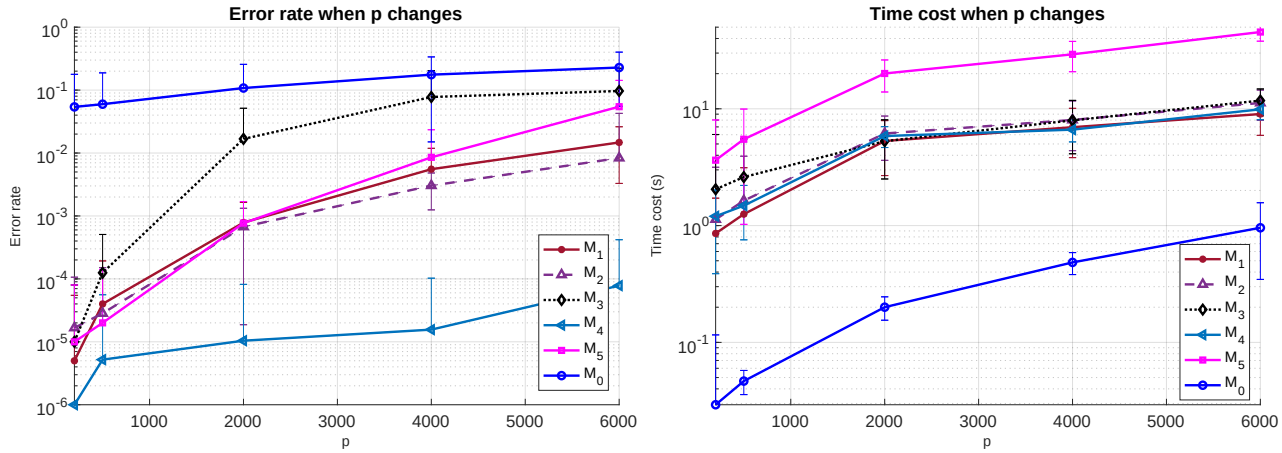

Figure 3: Log-scale error rates and runtime (with error bars) v.s. p under the setting of equal cluster size and $\Delta^2 = (\lambda^*\bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

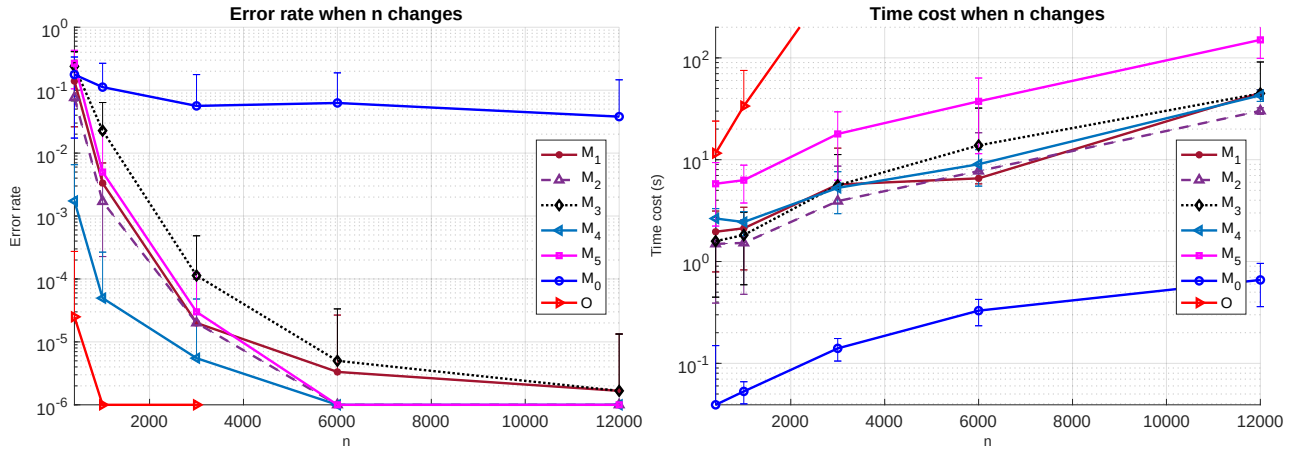

Figure 4: Log-scale error rates and runtime (with error bars) v.s. n under the setting of equal cluster size and $\Delta^2 = (\lambda^*\bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

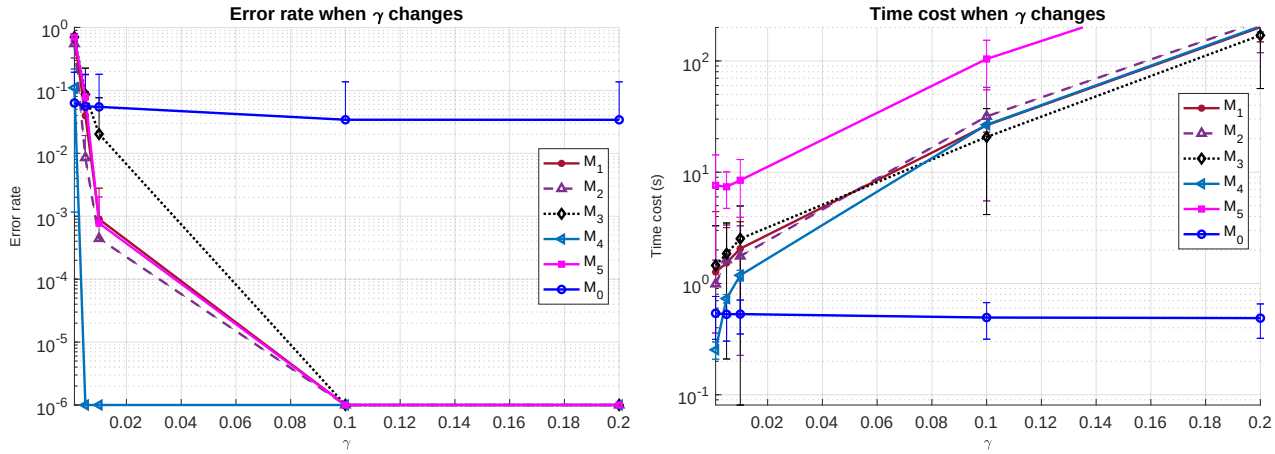

Figure 5: Log-scale error rates and runtime (with error bars) v.s. γ under the setting of equal cluster size and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

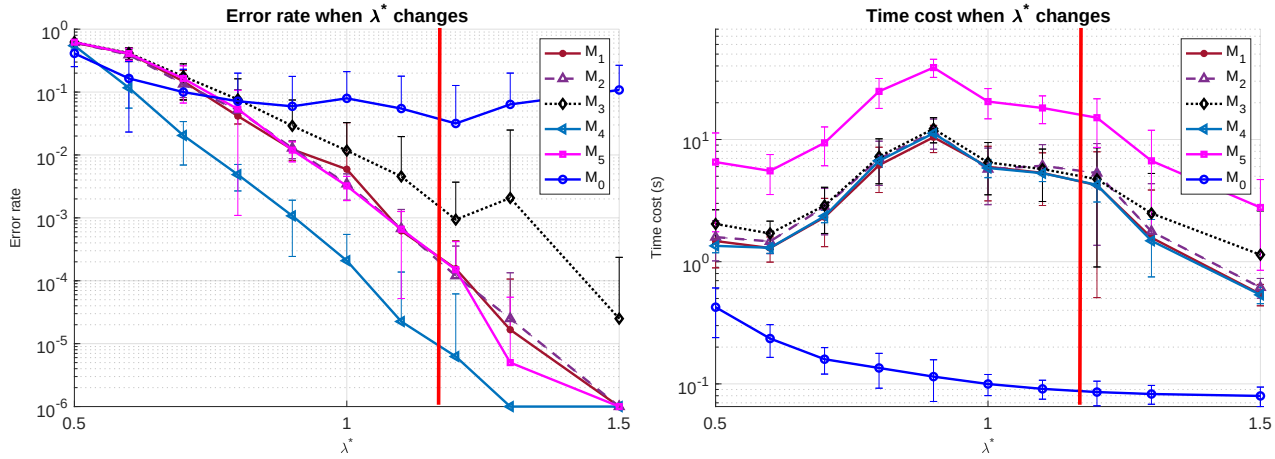

Figure 6: Log-scale error rates and runtime (with error bars) v.s. λ^* under the setting of equal cluster size and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Red vertical line indicates theoretical threshold $\bar{\Delta}_\gamma^2$ for SL methods. Zero error is displayed as 10^{-6} in the log-scale plot.

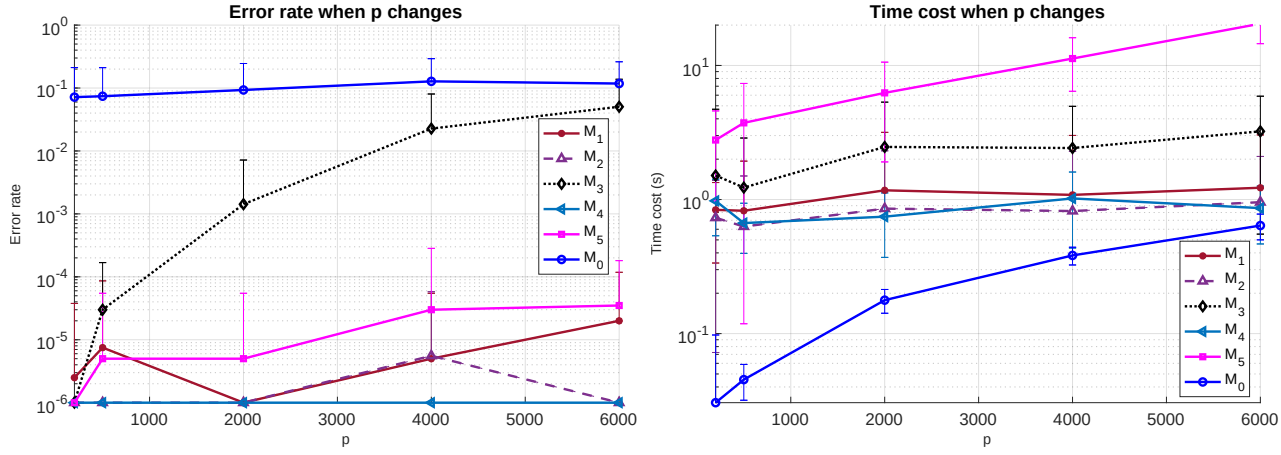

Figure 7: Log-scale error rates and runtime (with error bars) v.s. p under the setting of equal cluster size and $\Delta^2 = (\lambda^* \bar{\Delta}_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

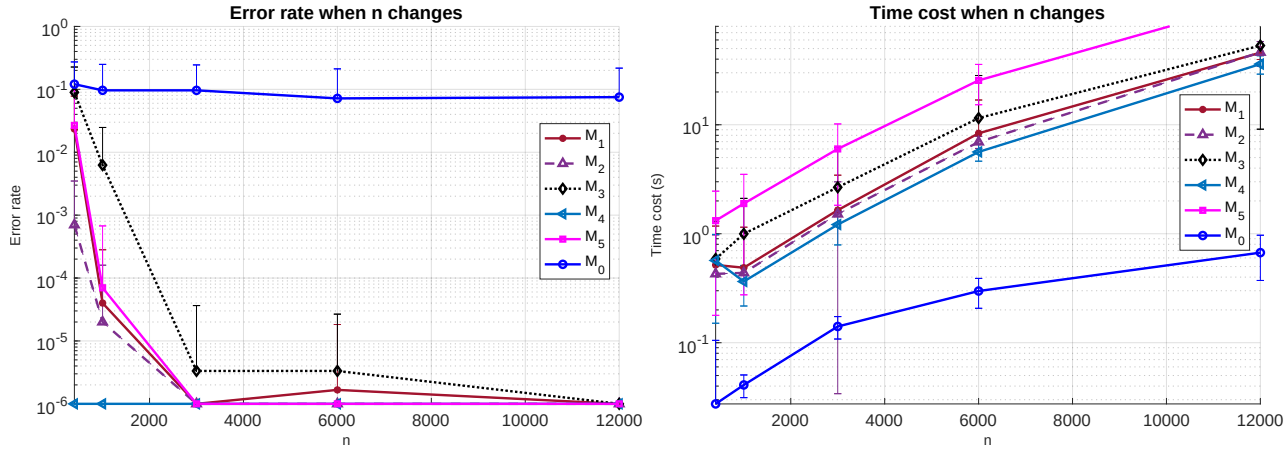

Figure 8: Log-scale error rates and runtime (with error bars) v.s. n under the setting of equal cluster size and $\Delta^2 = (\lambda^* \bar{\Delta}_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

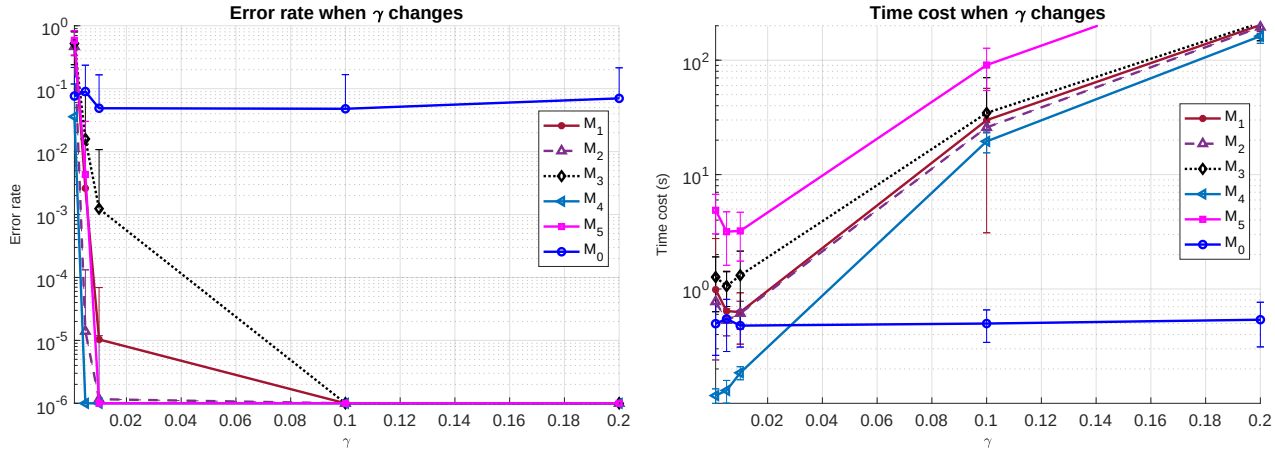

Figure 9: Log-scale error rates and runtime (with error bars) v.s. γ under the setting of equal cluster size and $\Delta^2 = (\lambda^* \bar{\Delta}_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

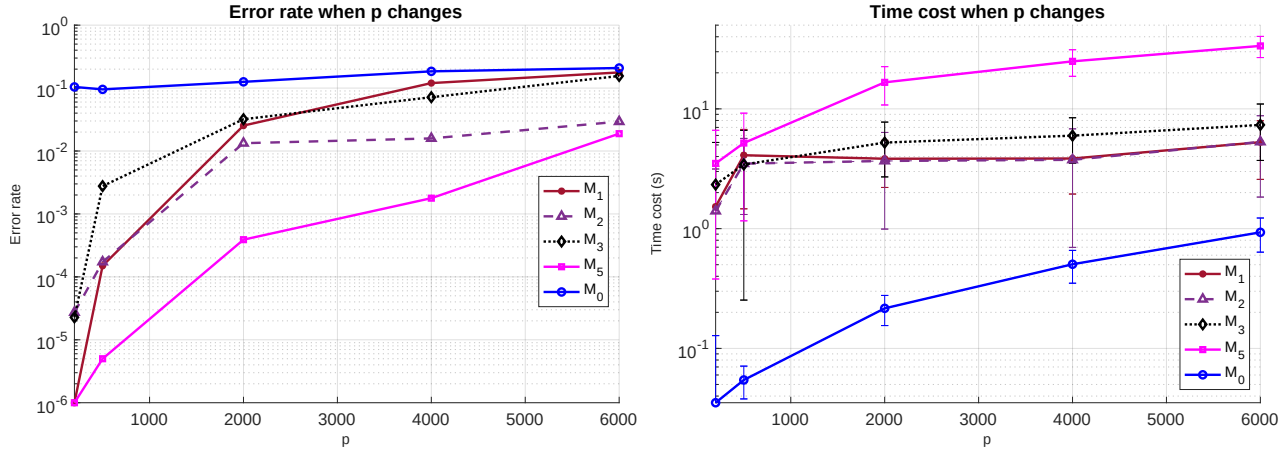

Figure 10: Log-scale error rates and runtime (with error bars) v.s. p under the setting of unequal cluster size $(n_1 = n_2 = n/8, n_3 = n_4 = 3n/8)$ and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

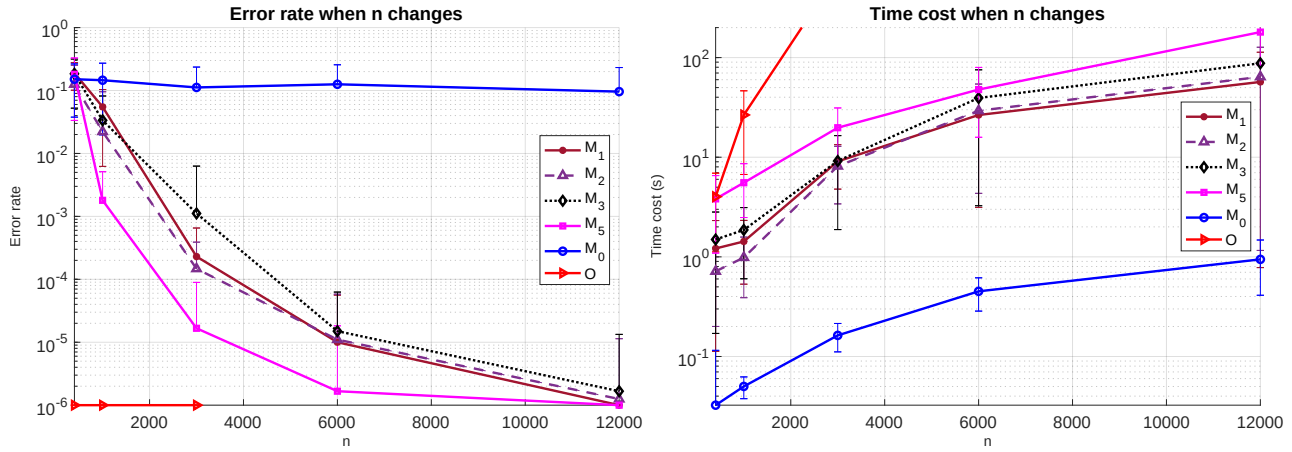

Figure 11: Log-scale error rates and runtime (with error bars) v.s. n under the setting of unequal cluster size ($n_1 = n_2 = n/8, n_3 = n_4 = 3n/8$) and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

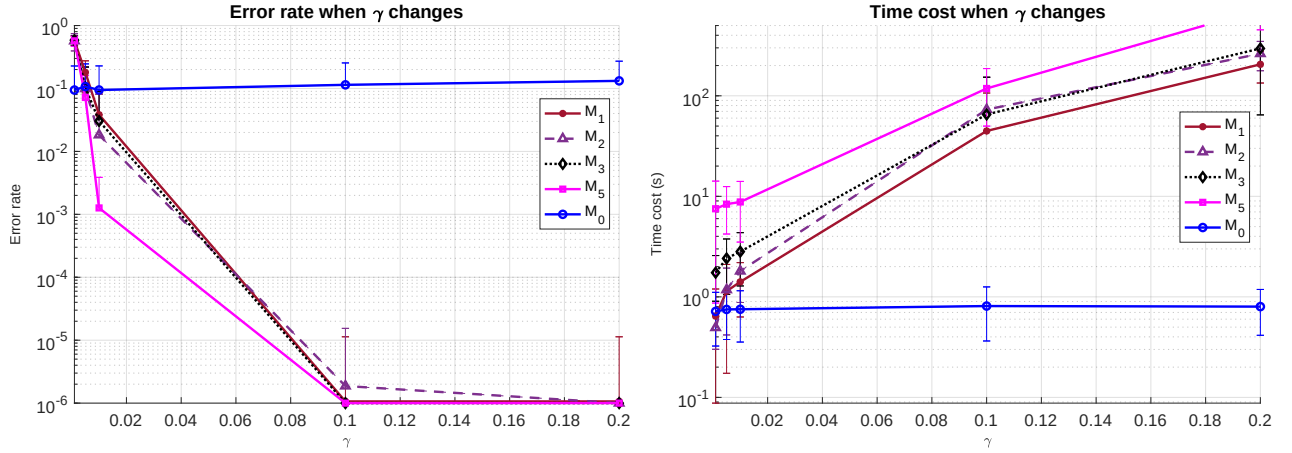

Figure 12: Log-scale error rates and runtime (with error bars) v.s. γ under the setting of unequal cluster size ($n_1 = n_2 = n/8, n_3 = n_4 = 3n/8$) and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

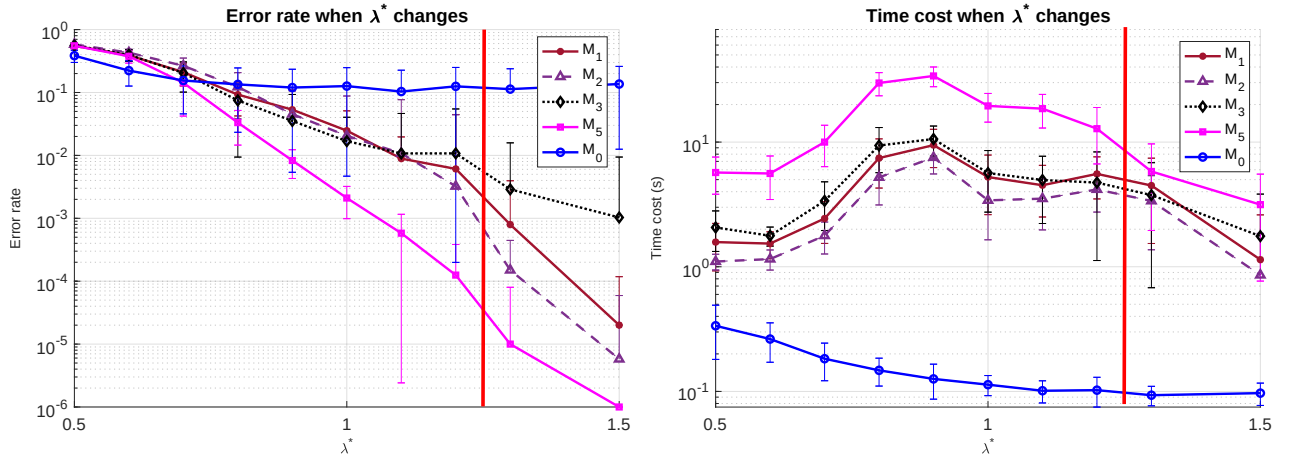

Figure 13: Log-scale error rates and runtime (with error bars) v.s. λ^* under the setting of unequal cluster size ($n_1 = n_2 = n/8, n_3 = n_4 = 3n/8$) and $\Delta^2 = (\lambda^* \bar{\Delta}_*)^2$. Red vertical line indicates theoretical threshold $\bar{\Delta}_\gamma'^2$ for SL methods. Zero error is displayed as 10^{-6} in the log-scale plot.

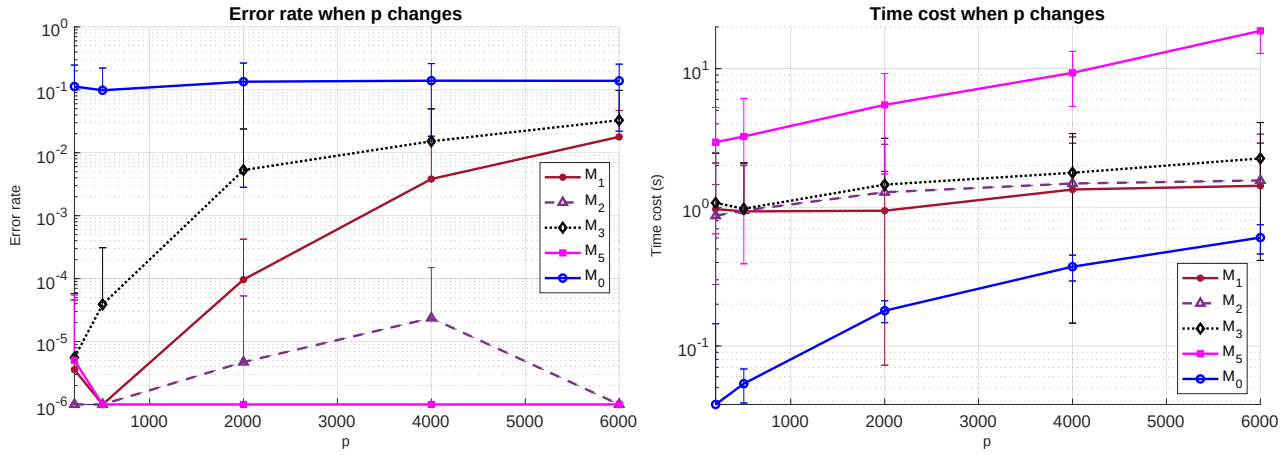

Figure 14: Log-scale error rates and runtime (with error bars) v.s. p under the setting of unequal cluster size $(n_1 = n_2 = n/8, n_3 = n_4 = 3n/8)$ and $\Delta^2 = (\lambda^* \bar{\Delta}'_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

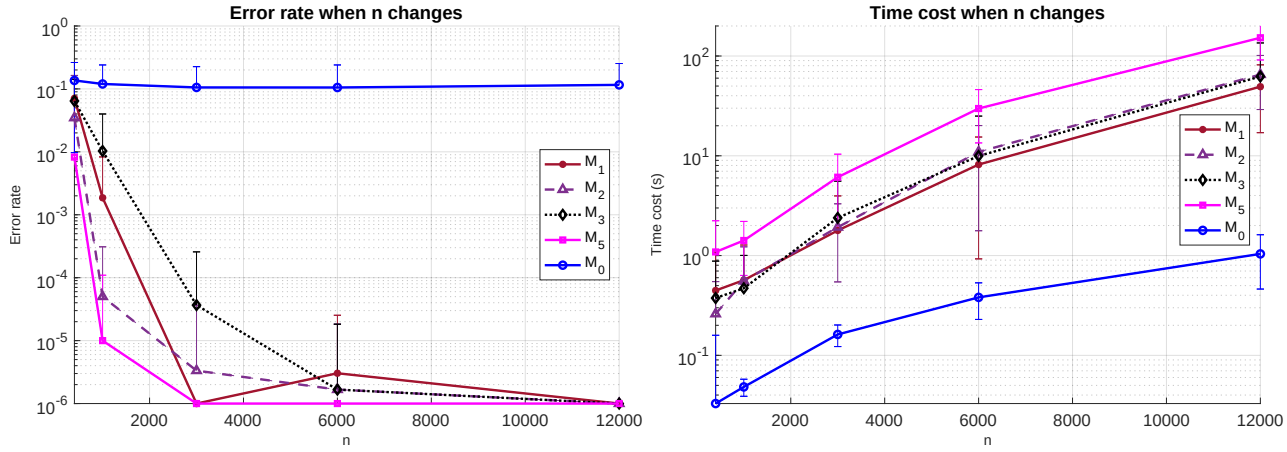

Figure 15: Log-scale error rates and runtime (with error bars) v.s. n under the setting of unequal cluster size $(n_1 = n_2 = n/8, n_3 = n_4 = 3n/8)$ and $\Delta^2 = (\lambda^* \bar{\Delta}'_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

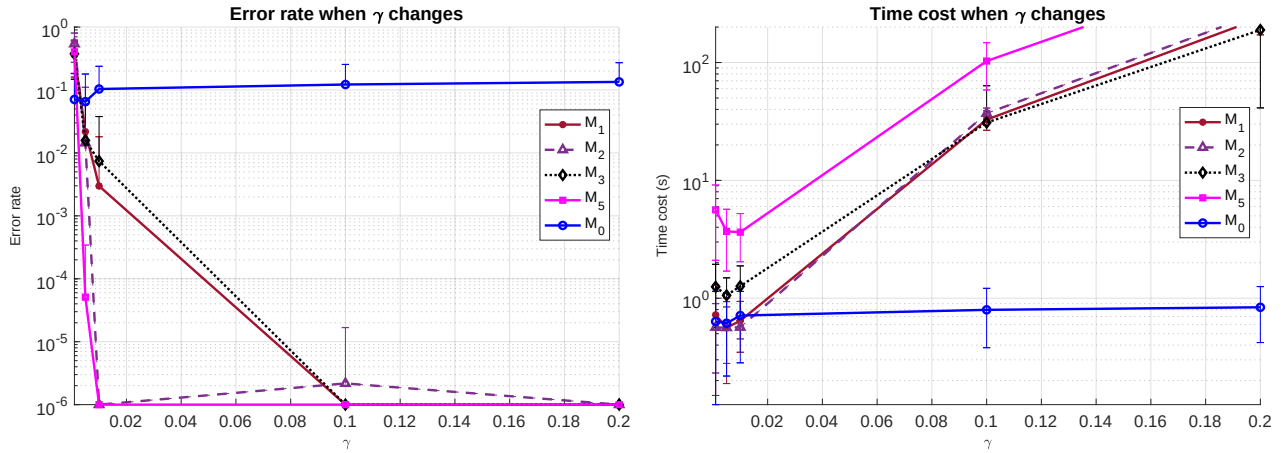

Figure 16: Log-scale error rates and runtime (with error bars) v.s. γ under the setting of unequal cluster size $(n_1 = n_2 = n/8, n_3 = n_4 = 3n/8)$ and $\Delta^2 = (\lambda^* \bar{\Delta}'_\gamma)^2$. Zero error is displayed as 10^{-6} in the log-scale plot.

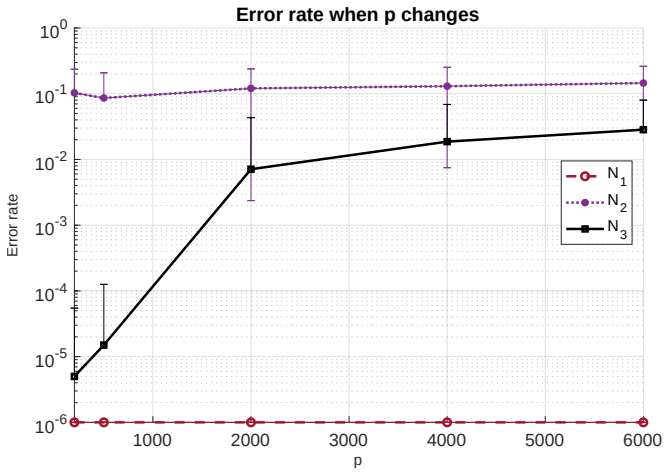

Figure 17: Log-scale error rates (with error bars) v.s. p under the setting of unequal cluster size ($n_1 = n_2 = n/8, n_3 = n_4 = 3n/8$) and $\Delta^2 = (\lambda^* \bar{\Delta}'_\gamma)^2$. N_1 is WSL when we plug in the true weights. N_2 is K-means++ method. N_3 is WSL when we plug in the weights based on K-means++ method. Zero error is displayed as 10^{-6} in the log-scale plot.

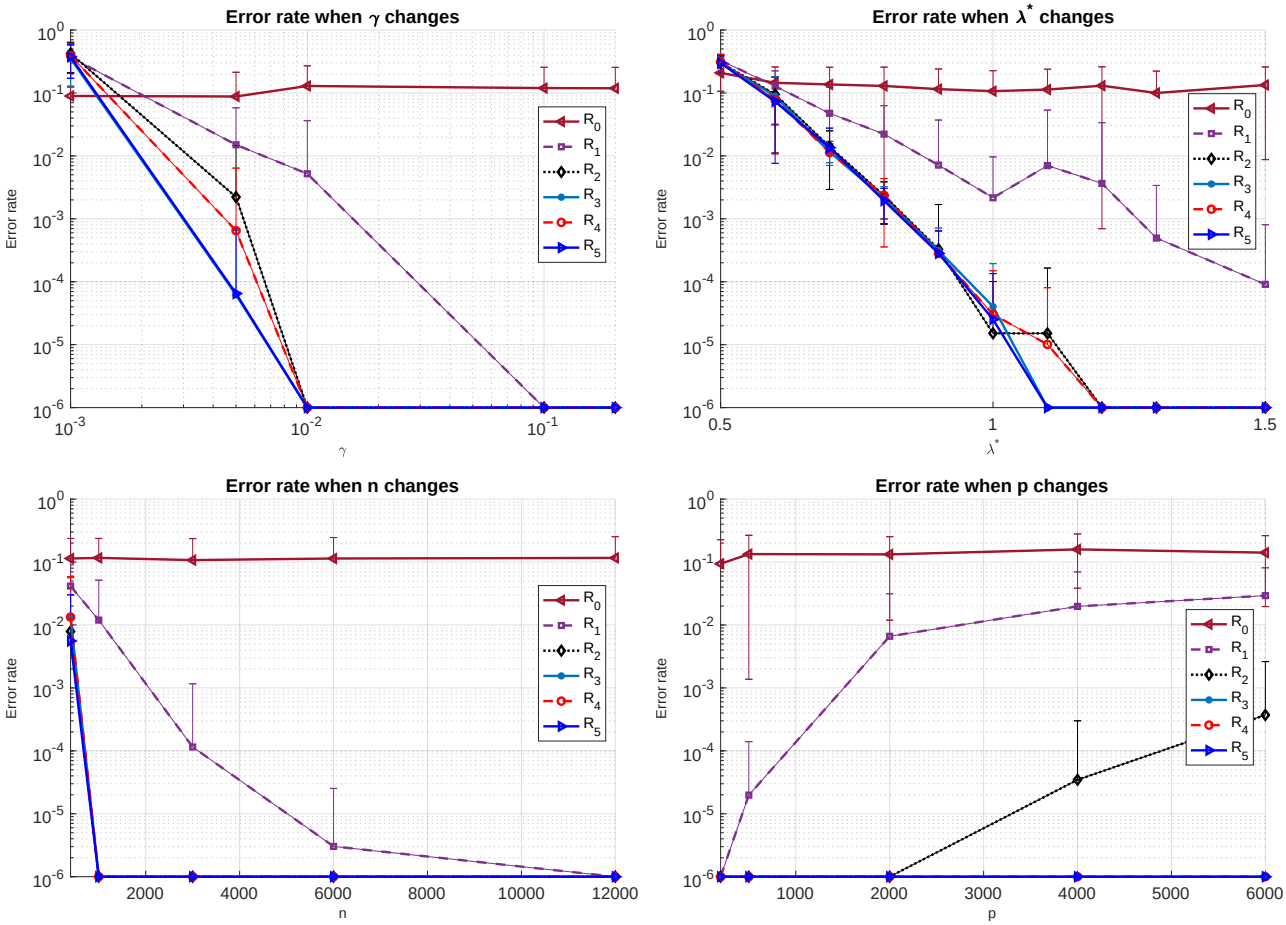

Figure 18: Error rates of MR-WSL. R_0 stands for the initial K-means++ as the warm start for our MR-WSL and R_i stands for the i-th round of MR-WSL. Zero error is displayed as 10^{-6} in the log-scale plot. In particular, we take the log-scale for x axis when γ changes. The settings for Figure 18 are the same as Figure 17. From the plots we can see that after 3-4 rounds, the performance gets stable and achieved optimal.

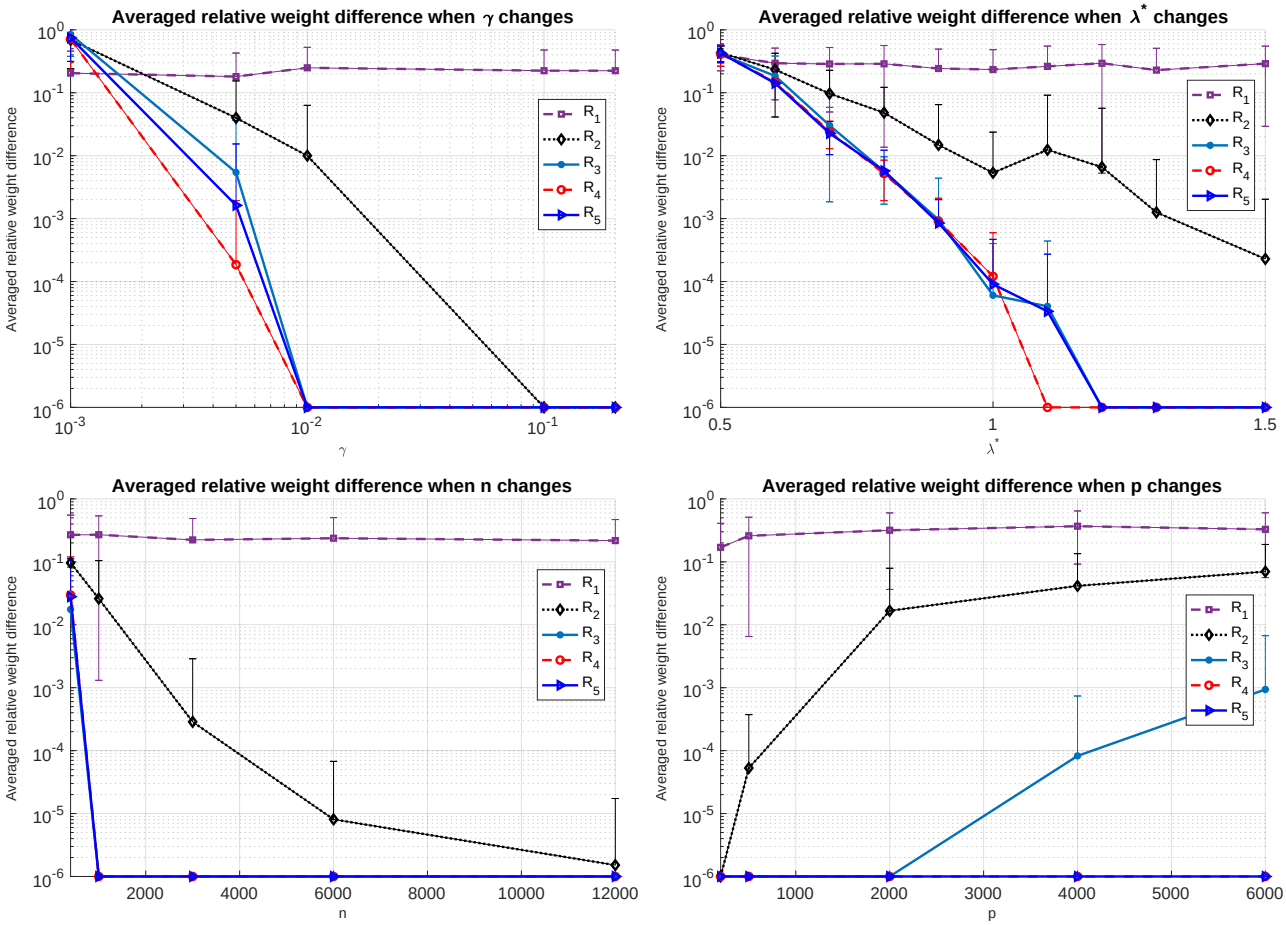

Figure 19: Averaged relative weight difference of MR-WSL. Averaged relative weight difference is defined by $\frac{1}{n}\sum_{j\in[n]}|w_j - w_j^*|/w_j^*$. where w_j, w_j^* is defined in Section 3.3. R_i stands for the i-th round of MR-WSL corresponding to Figure 18. Zero error is displayed as 10^{-6} in the log-scale plot. In particular, we take the log-scale for x axis when γ changes. Figure 18 and Figure 19 have same settings.

Table 1: Fraction of (ϵ, δ)-weights initialized by K-means++ algorithm (1-st round of WSL, R_1 for Figure 18).

ϵ	0	0.2	0.4	0.6	0.8	1.0
$p = 200$	0.2602	0.25373	0.21828	0.10665	0.071097	0.054791
$p = 500$	0.2725	0.25598	0.18602	0.10836	0.06278	0.049165
$p = 2000$	0.44123	0.37208	0.27103	0.12647	0.085495	0.065855
$p = 4000$	0.40125	0.33282	0.24355	0.10086	0.068535	0.05453
$p = 6000$	0.48375	0.38929	0.25716	0.13754	0.07653	0.06671

We fix a grid of ϵ and report the estimated values of δ satisfying (6) for each ϵ. Each δ is averaged over 100 simulations. The setting for Table 1 is the same as Figure 18.

Table 2: Fraction of (ϵ, δ)-weights for the 2-nd round of WSL (R_2 for Figure 18).

ϵ	0	0.2	0.4	0.6	0.8	1.0
$p = 200$	0.005	0.000005	0.000005	0.000005	0.000005	0.000005
$p = 500$	0.02625	0.00006	0.00006	0.00006	0.00006	0.00006
$p = 2000$	0.13125	0.028845	0.01318	0.012105	0.012105	0.00967
$p = 4000$	0.3725	0.088965	0.04746	0.02566	0.02266	0.0198
$p = 6000$	0.54875	0.15886	0.095185	0.057355	0.044775	0.03841

We fix a grid of ϵ and report the estimated values of δ satisfying (6) for each ϵ. Each δ is averaged over 100 simulations. The setting for Table 2 is the same as Figure 18.

Table 3: Fraction of (ϵ, δ)-weights for the 3-rd round of WSL (R_3 for Figure 18).

ϵ	0	0.2	0.4	0.6	0.8	1.0
$p = 200$	0	0	0	0	0	0
$p = 500$	0	0	0	0	0	0
$p = 2000$	0	0	0	0	0	0
$p = 4000$	0.00875	0.005	0.00125	0.00125	0.00125	0.00125
$p = 6000$	0.02125	0.00005	0.00005	0.00005	0.00005	0.00005

We fix a grid of ϵ and report the estimated values of δ satisfying (6) for each ϵ. Each δ is averaged over 100 simulations. The setting for Table 3 is the same as Figure 18.

Table 4: Fraction of (ϵ, δ)-weights for the 4-th round of WSL (R_4 for Figure 18).

ϵ	0	0.2	0.4	0.6	0.8	1.0
$p = 200$	0	0	0	0	0	0
$p = 500$	0	0	0	0	0	0
$p = 2000$	0	0	0	0	0	0
$p = 4000$	0	0	0	0	0	0
$p = 6000$	0	0	0	0	0	0

We fix a grid of ϵ and report the estimated values of δ satisfying (6) for each ϵ. Each δ is averaged over 100 simulations. The setting for Table 4 is the same as Figure 18.