The Dissertation Committee for Bowei Yan
certifies that this is the approved version of the following dissertation:

Theoretical Analysis for Convex and Non-convex Clustering Algorithms

Committee:

Purnamrita Sarkar, Supervisor

Constantine Caramanis

Peter Müller

Stephen Walker

Theoretical Analysis for Convex and Non-convex Clustering Algorithms

by

Bowei Yan

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my beloved parents.

Acknowledgments

I have enjoyed my time at UT immensely and there are many people who have shown me the greatest kindness and given me tremendous help and guidance. I would like to thank everyone who helped me along the journey, as well as those whose presence made it so enjoyable. Below, I highlight a few people who have played especially important roles.

I would like to start by thanking my advisor, Purnamrita Sarkar, who has been an incredible mentor and unwavering advocate. Her wisdom and dedication have shown me what it is like to be a true researcher and a generous mentor. She is not only a great advisor but also a supporting friend, someone who I can talk to whenever I was having hard time in work or life. Purna encouraged me to pursue my own interests and guided me with her insightful questions and suggestions during our conversations, which always led me to a deeper understanding of my work. Her perceptive feedback has not only improved the quality of my research, but also my ability to describe it to others. I am beyond grateful for having the opportunity to learn and work with her.

I would also like to thank Peter Mueller, Stephen Walker, and Constantine Caramanis for serving on my doctoral committee along with Purna. What I learned from their courses built the foundation for my research, and

their questions and feedbacks have been invaluable to improve my thesis.

I would not have completed these work without my collaborators: Mingzhang Yin, who was always there for discussions and conversations; Xiuyuan Cheng, a long time friend who often provided a different perspective to our discussion. I am indebted to Arash Amini, who has given me a lot of guidance and provided many insightful ideas during our meetings on non-convex optimization. I also want to thank Soledad Villar for fruitful communications on semi-definite programming and clustering, and Soumendu Mukherjee and Rachel Wang for our long-distance collaboration on variational inference. I am grateful for having Xueyu Mao and Prateek Srivastava as a team for reading seminars and related discussions.

I also owe an enormous debt of gratitude to Oluwasanmi Koyejo and Pradeep Ravikumar, for their mentorship and collaboration on supervised side of machine learning. I thank Wesley Tansey for his kindly support on his package for graph fused lasso. I want to thank Ying Liu, Haochang Shou, Lei Huang and Xuebin Yang for relating my research on real world problems. My thanks also go to my Master advisor, Yuan Yao, who encouraged my pursuit of a PhD and led me to the research of statistical machine learning theory.

My thanks go to the Statistics Department at UT for its superb supporting system. I would like to thank our amazing staff members for all the support and warmness that I have received from them. More broadly, I would like to thank all the students of SDS for their friendship over the years.

I am grateful for my mentors in my internships: Ken Terao, Weifeng Liu, Eddie Xu, and Jian Zhang. They showed my how machine learning can be used in quantitative finance and helped my through my first steps in industry. I would like to thank Yang Zhao, Junchi Li, Tianyang Li, who have given me a lot of advices on my career path.

I am also very grateful to have the support from many of my other friends: Lingyuan Gao, Chenguang Liu, Yumeng Ou, Min Yang, Bo Zhang, Rayman Zheng, Ye Li, Tina Tang. Their companion made my years of PhD so much more enjoyable, and the hard time easier to pass. Thanks to all those I have missed in these acknowledgements, and I am sure I have missed many.

My mother, Lizhen Gao, and father, Ming Yan, shower me with love and trust throughout my life, unconditionally support my interests and ambitions. I will remain forever grateful. This thesis is dedicated to them.

Finally, to Kai Zhong. Thank you for being by my side through all the ups and downs.

Theoretical Analysis for Convex and Non-convex Clustering Algorithms

Publication No. _____

Bowei Yan, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Purnamrita Sarkar

Clustering is one of the most important unsupervised learning problem in the machine learning and statistics community. Given a set of observations, the goal is to find the latent cluster assignment of the data points. The observations can be either some covariates corresponding to each data point, or the relational networks representing the affinity between pair of nodes. We study the problem of community detection in stochastic block models and clustering mixture models. The two kinds of problems bear a lot of resemblance, and similar techniques can be applied to solve them.

It is common practice to assume some underlying model for the data generating process in order to analyze it properly. With some pre-defined partitions of all data points, generative models can be defined to represent those two types of data observations. For the covariates, the mixture model is one of the most flexible and widely-used models, where each cluster i comes

from some distribution $\mathcal{D}_i$, and the entire distribution is a convex sum over all distributions $\sum_{i=1}^{r} \pi_i \mathcal{D}_i$. We assume that the data is Gaussian or sub-gaussian, and analyze two algorithms: 1) Expectation-Maximization algorithm, which is notoriously non-convex and sensitive to local optima, and 2) Convex relaxation of the k-means algorithm. We show both methods are consistent under certain conditions when the signal to noise ratio is relatively high. And we obtain the upper bounds for error rate if the signal to noise ration is low. When there are outliers in the data set, we show that the semi-definite relaxation exhibits more robust result compared to spectral methods.

For the networks, we consider the Stochastic Block Model (SBM), in which the probability of edge presence is fully determined by the cluster assignments of the pair of nodes. We use a semi-definite programming (SDP) relaxation to learn the clustering matrix, and discuss the role of model parameters. In most SDP relaxations of SBM, the number of communities is required for the algorithm, which is a strong requirement for many real-world applications. In this thesis, we propose to introduce a regularization to the nuclear norm, which is shown to be able to exactly recover both the number of communities and cluster memberships even when the number of communities is unknown.

In many real-world networks, it is more common to see both network structure and node covariates simultaneously. In this case, we present a regularization based method to effectively combine the two sources of information. The proposed method works especially well when the covariates and network

contain complementary information.

Attribution The research presented in this dissertation was the product of truly collaborative work that was completed only through the repeated key insights of everyone involved. Chapter 2 pertains to analysis of EM algorithm and robustness analysis for kernel clustering. The former is the result of collaboration with Mingzhang Yin and Purnamrita Sarkar, and was published at Neural Information Proceeding System (NIPS) 2017. The latter collaboration with Purnamrita Sarkar was published at NIPS 2016.

Chapter 3 pertains to the analysis for semi-definite relaxation for community detection in dense and sparse stochastic block models. Part of the work is in collaboration with Xiuyuan Cheng and Purnamrita Sarkar, and is presented at International Conference of Artificial Intelligence and Statistics (AISTATS) 2018.

Chapter 4 discuss the combination of network and covariates for clustering inference. This work is completed with Purnamrita Sarkar. The first version was published in arXiv in 2014, and is currently under revision for Journal of American Statistics Association.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

It would seem that mythological worlds have been built up only to be shattered again, and that new worlds were built from the fragments.

Franz Boas, in Introduction to James Teitś Traditions of the Thompson River Indians of British Columbia, Memoirs of the American Folklore Society, VI (1898), 18

Identifying patterns is one of the most fundamental cognitive skills human beings possess and it has been crucial in statistics and machine learning. For unsupervised learning problems, the true value of the response (e.g. label) is not available.

Clustering is one of the most fundamental tasks in unsupervised learning. It helps us to organize the data and often serve as an exploratory step for more sophisticated tasks. The motivation of this thesis is to find latent clusters in unlabeled data. There are two common types of observations in real world: one is defined as features of a given object; the other is defined via the relationship between pairs of objects, and is often collected in the form of

networks. These two sources provide different aspects of the problem, yet a systematic understanding on how the two sources can be combined to provide better clustering is not theoretically well understood.

Take for example the Mexican political elites network (described in detail in Chapter 4). This dataset comprises of 35 politicians (military or civilian) and their connections. The associated covariate for each politician is the year when one came into power. After the military coup in 1913, the political arena was dominated by the military. In 1946, the first civilian president since the coup was elected. Hence those who came into power later are more likely to be civilians. Politicians who have similar number of connections to the military and civilian groups are hard to classify from the network alone. Here the temporal covariate is crucial in resolving which group they belong to. On the other hand, politicians who came into power around 1940s, are ambiguous to classify using covariates. Hence the number of connections to the two groups in the network helps in classifying these nodes. In this work, our goal is to provide a solution for such problems, and to effectively combine networks and covariates for an accurate community detection algorithm with theoretical guarantees under broad parameter regimes.

We study the problem under some generative models on both covariates and networks, and assume there exists a ground truth cluster structure, where our goal is to correctly recover this true labeling. For the covariates, we assume the data comes from a mixture model. Formally, assume there are r non-overlapping clusters for the n observations, and there are latent variables

$Z_i \in [r], i \in [n]$ indicating the membership of each observation. With slight mis-use of notation, we sometimes also use Z as a $n \times r$ binary matrix to represent the one-hot encoding of the memberships of all nodes. Each observation is a d-dimensional vector, representing d different covariates. We further assume there exist a collection of distribution $\mathcal{D}_1, \cdots, \mathcal{D}_r$, with mixing weights $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_r)$, such that given latent variable $Z_i = k$, the covariates X is generated from some distribution $\mathcal{D}_k$:

$$Z \sim Categorical(\boldsymbol{\pi})$$
$$X|Z = z \sim \mathcal{D}_z$$

For the convenience of analysis, we also make distributional assumptions on the distribution $\mathcal{D}_z$. One of the most commonly used one is Multivariate Gaussian distribution. Under this setting, we analyze the Expectation-Maximization algorithm and show that with a proper initialization, the mean of each Gaussian distribution can be recovered with error scaling as $O(\sqrt{d/n})$.

In Section 2.8-2.10, we generalize this assumption to sub-gaussian distributions, whose tail decays no slower than a Gaussian. More specifically we look at the model proposed in [38] where the dimension goes to infinity. For these sub-gaussian mixtures, we propose and analyze a kernel-based convex relaxation, and turn the problem into solving a semi-definite programming (SDP). We prove that this method works well even at presence of outliers.

For the network, we study the Stochastic Block Model (SBM)[51]. SBM has drawn much attention among theoretical statisticians and theoretical com-

puter scientists due to its simplicity and flexibility. The model assumes the network is undirected and unweighted, and the existence of edge between a pair of nodes only depends on the membership of both nodes. The assortativity assumption assumes that the connecting probability between nodes in the same cluster is higher than that between nodes in different clusters. Formally, if we still follow the notation where $Z_i \in [r]$ represent the latent membership of the i-th node, then

$$A_{ij}|Z_i = a, Z_j = b \sim Ber(B_{ab})$$

where $B \in [0,1]^{r\times r}$ is a parameter matrix.

The inference for SBM gets harder as the network gets sparser, due to the fact that the number of edges observed decreases. It is shown in [122] that when the average degree is of order $\Theta(1)$, there is no consistent algorithm for clustering the nodes. One can at most recover a proportion of memberships correctly. In the sequel, we will refer this regime as "sparse" regime, and refer the regime where average degree is $\Omega(\log n)$ as "dense" regime. In Chapter 3 we analyze a SDP relaxation which does not require the knowledge of number of clusters in the dense regime, and show that one can achieve exact recovery for both the number of clusters and all cluster memberships. For sparse networks, we show that the proposed SDP can outperform random guess and achieve a constant error rate, which decays as the signal increases.

In Chapter 4, we study the problem where both network and covariates are available. Our analysis states a bound combining both resources, and we

experimentally show that the combined problem outperforms that achieved by merely using a single source of information.

Now we present some common notations for assymptotics and matrix norms which will be used throughout.

1.0.1 Notations

Several matrix norms are considered in this manuscript. For a matrix $M \in \mathbb{R}^{n\times n}$, we use $\|M\|_F$ and $\|M\|$ to denote the Frobenius and operator norms of M respectively. Let the eigenvalues of M be denoted by $\lambda_1 \geq \cdots \geq \lambda_n$. The operator norm $\|M\|$ is simply the largest eigenvalue of M, i.e. λ_1. For a symmetric matrix, it is the magnitude of the largest eigenvalue. The nuclear norm is $\|M\|_* = \sum_{i=1}^{n} \sigma_i$. The ℓ_1 and ℓ_∞ norm are defined the same as the vector ℓ_1 and ℓ_∞ norm $\|M\|_1 = \sum_{ij} |M_{ij}|$, $\|M\|_\infty = \max_{i,j} |M_{ij}|$. For two matrices $M, Q \in \mathbb{R}^{m\times n}$, their inner product is $\langle M, Q\rangle = \text{trace}(M^T Q)$. The $\ell_\infty \to \ell_1$ norm of a matrix M is defined as $\|M\|_{\ell_\infty\to\ell_1} = \max_{\|s\|_\infty\leq 1} \|Ms\|_1$.

Throughout the manuscript, we use $\mathbf{1}_n$ to represent the all one $n \times 1$ vector and $E_n, E_{n,k}$ to represent the all one matrix with size $n \times n$ and $n \times k$. The subscript will be dropped when it is clear from context. We use $\otimes$ to represent the kronecker product.

For the asymptotic analysis, we use the following standard notations for approximated rate of convergence. $T(n)$ is $O(f(n))$ if and only if for some constant c and n_0, $T(n) \leq cf(n)$ for all $n \geq n_0$; $T(n)$ is $\Omega(f(n))$ if for some constant c and n_0, $T(n) \geq cf(n)$ for all $n \geq n_0$; $T(n)$ is $\Theta(f(n))$ if $T(n)$ is

$O(f(n))$ and $\Omega(f(n))$; $T(n)$ is $o(f(n))$ if $T(n)$ is $O(f(n))$ but not $\Omega(f(n))$. $T(n)$ is $o_P(f(n))$ (or $O_P(f(n))$) if it is $o(f(n))$ (or $O(f(n))$) with high probability. $f(n) = \tilde{\Omega}(g(n))$ is short for $\Omega(g(n))$ ignoring logarithmic factors, equivalent to $f(n) \geq Cg(n)\log^k(g(n))$, similar for others.

Chapter 2

Covariate Clustering - Non-Convex and Convex approaches

One of the most natural choices for generative model is a mixture of Gaussians. In this chapter, we assume the data is generated from a collection of distributions $\{\mathcal{D}_1, \cdots, \mathcal{D}_r\}$, where r is the number of clusters. Each distribution comes from a multivariate Gaussian with mean $\boldsymbol{\mu}_i$ and some covariance matrix. There are two questions to be asked when dealing with a Gaussian Mixture Model (GMM), first is how do we estimate the model parameters, and the second is whether we can label all points with high accuracy. In this chapter, we will first discuss the first question and analyze a decades-old algorithm, expectation-maximization (EM) [35] algorithm, and provide theoretical guarantees on the recovery of the parameters. Our result weakens the convergence

The content in this chapter was published in [1] Yan, Bowei, Mingzhang Yin, and Purnamrita Sarkar. "Convergence of Gradient EM on Multi-component Mixture of Gaussians." In Advances in Neural Information Processing Systems, pp. 6959-6969. 2017. and [2] Yan, Bowei, and Purnamrita Sarkar. "On robustness of kernel clustering." In Advances in Neural Information Processing Systems, pp. 3098-3106. 2016. For [1], I participated in posing the problem. I and the second author developed the population analysis with a little help from Prof. Sarkar. Prof. Sarkar and I developed the sample analysis. I wrote the entire paper, both the second author and I conducted the experiments, and Prof. Sarkar helped in revising and rewriting. For [2], Prof. Sarkar proposed the problem of robustness analysis. We jointly formulated the problem, and developed the theory. I implemented and conducted the experimental analysis, and wrote the manuscript. Prof. Sarkar helped revise and rewrite the draft.

criterion in previous work [9], and shows that under a fairly mild initialization condition, the EM algorithm converges linearly to the global optimum.

Then we discuss a Gaussian mixture model in the high dimensional space [38], where the kernel matrix concentrates when the dimension goes to infinity. We use the SDP relaxation proposed in [92] and show that the clustering matrix can be exactly recovered under certain separation conditions. The SDP also enjoys robustness properties when the data is contaminated by arbitrarily distributed outliers. We compare the robustness behavior of the SDP and other commonly-used methods such as kernel PCA and conclude that SDP has higher tolerance to outliers.

2.1 Convergence Analysis for EM Algorithm

Proposed by [35] in 1977, the Expectation-Maximization (EM) algorithm is a powerful tool for statistical inference in latent variable models. A famous example is the parameter estimation problem under parametric mixture models. In such models, data is generated from a mixture of a known family of parametric distributions. The mixture component from which a datapoint is generated from can be thought of as a latent variable.

Typically the marginal data log-likelihood (which integrates the latent variables out) is hard to optimize, and hence EM iteratively optimizes a lower bound of it and obtains a sequence of estimators. This consists of two steps. In the expectation step (E-step) one computes the expectation of the complete data likelihood with respect to the posterior distribution of the unobserved

mixture memberships evaluated at the current parameter estimates. In the maximization step (M-step) one this expectation is maximized to obtain new estimators. EM always improves the objective function. While it is established in [27] that the true parameter vector is the global maximizer of the log-likelihood function, there has been much effort to understand the behavior of the local optima obtained via EM.

When the exact M-step is burdensome, a popular variant of EM, named Gradient EM is widely used. The idea here is to take a gradient step towards the maxima of the expectation computed in the E-step. [64] introduces a gradient algorithm using one iteration of Newton's method and shows the local properties of the gradient EM are almost identical with those of the EM.

Early literature [109, 111] mostly focuses on the convergence to the stationary points or local optima. In [109] it is proven that the sequence of estimators in EM converges to stationary point when the lower bound function from E-step is continuous. In addition, some conditions are derived under which EM converges to local maxima instead of saddle points; but these are typically hard to check. A link between EM and gradient methods is forged in [111] via a projection matrix and the local convergence rate of EM is obtained. In particular, it is shown that for GMM with well-separated centers, the EM achieves faster convergence rates comparable to a quasi-Newton algorithm. While the convergence of EM deteriorates under worse separations, it is observed in [94] that the mixture density determined by estimator sequence of EM reflects the sample data well.

In recent years, there has been a renewed wave of interest in studying the behavior of EM especially in GMMs. The global convergence of EM for a mixture of two equal-proportion Gaussian distributions is fully characterized in [110]. For more than two clusters, a negative result on EM and gradient EM being trapped in local minima arbitrarily far away from the global optimum is shown in [54].

For high dimensional GMMs with r components, the parameters are learned via reducing the dimensionality via a random projection in [28]. In [30] the two-round method is proposed, where one first initializes with more than r points, then prune to get one point in every cluster. It is pointed out in this paper that in high dimensional space, when the clusters are well separated, the mixing weight will go to either 0 or 1 after one single update. It is showed in [114, 79] that one can cluster high dimensional sub-gaussian mixtures by semi-definite programming relaxations.

For the convergence rate of EM algorithm, it is observed in [84] that a very small mixing proportion for one mixture component compared to others leads to slow convergence. In [9] the authors give non-asymptotic convergence guarantees in isotropic, balanced, two-component GMM; their result proves the linear convergence of EM if the center is initialized in a small neighborhood of the true parameters. The local convergence result in this paper has a sub-optimal contraction region.

K-means clustering is another widely used clustering method. Lloyd's algorithm for k-means clustering has a similar flavor as EM. At each step,

it recomputes the centroids of each cluster and updates the membership assignments alternatively. While EM does soft clustering at each step, Lloyd's algorithm obtains hard clustering. The clustering error of Lloyd's algorithm for arbitrary number of clusters is studied in [72]. The authors also show local convergence results where the contraction region is less restrictive than [9].

We would like to point out that there are many notable algorithms [63, 8, 103] with provable guarantees for estimating mixture models. In [75, 40] the authors propose polynomial time algorithms which achieve epsilon approximation to the k-means loss. A spectral algorithm for learning mixtures of gaussians is proposed in [103]. We want to point out that our aim is not to come up with a new algorithm for mixture models, but to understand the interplay of model parameters in the convergence of gradient EM for a mixture of Gaussians with r components. As we discuss later, our work also immediately leads to convergence guarantees of Stochastic Gradient EM. Another important difference is that the aim of these works is recovering the hidden mixture component memberships, whereas our goal is completely different: we are interested in understanding how well EM can estimate the mean parameters under a good initialization.

In this chapter, we study the convergence rate and local contraction radius of gradient EM under GMM with arbitrary number of clusters and mixing weights which are assumed to be known. For simplicity, we assume that the components share the same covariance matrix, which is known. Thus it suffices to carry out our analysis for isotropic GMMs with identity as the shared

covariance matrix. We obtain a near-optimal condition on the contraction region in contrast to [9]'s contraction radius for the mixture of two equal weight Gaussians. We want to point out that, while the authors of [9] provide a general set of conditions to establish local convergence for a broad class of mixture models, the derivation of specific results and conditions on local convergence are tailored to the balance and symmetry of the model.

We follow the same general route: first we obtain conditions for population gradient EM, where all sample averages are replaced by their expected counterpart. Then we translate the population version to the sample one. While the first part is conceptually similar, the general setting calls for more involved analysis. The second step typically makes use of concepts from empirical processes, by pairing up Ledoux-Talagrand contraction type arguments with well established symmetrization results. However, in our case, the function is not a contraction like in the symmetric two component case, since it involves the cluster estimates of all r components. Furthermore, the standard analysis of concentration inequalities by McDiarmid's inequality gets complicated because the bounded difference condition is not satisfied in our setting. We overcome these difficulties by taking advantage of recent tools in Rademacher averaging for vector valued function classes, and variants of McDiarmid type inequalities for functions which have bounded difference with high probability.

The rest of this chapter is organized as follows. In Section 4.2, we state the problem and the notations. In Section 3, we provide the main results

in local convergence rate and region for both population and sample-based gradient EM in GMMs. Section 2.4 and 2.5 provide the proof sketches of population and sample-based theoretical results, followed by the numerical result in Section 4.4.

2.2 Problem Setup and Notations

Consider a GMM with r clusters in d dimensional space, with weights $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_r)$. Let $\boldsymbol{\mu}_i \in \mathbb{R}^d$ be the mean of cluster i. Without loss of generality, we assume $\mathbb{E}X = \sum_i \pi_i \boldsymbol{\mu}_i = 0$ and the known covariance matrix for all components is I_d. Let $\boldsymbol{\mu} \in \mathbb{R}^{rd}$ be the vector stacking the $\boldsymbol{\mu}_i$s vertically. We represent the mixture as $X \sim \mathrm{GMM}(\pi, \boldsymbol{\mu}, I_d)$, which has the density function $p(x|\boldsymbol{\mu}) = \sum_{i=1}^r \pi_i \phi(x|\boldsymbol{\mu}_i, I_d)$. where $\phi(x; \boldsymbol{\mu}, \Sigma)$ is the PDF of $N(\boldsymbol{\mu}, \Sigma)$. Then the population log-likelihood function as $\mathcal{L}(\boldsymbol{\mu}) = \mathbb{E}_X \log\left(\sum_{i=1}^r \pi_i \phi(X|\boldsymbol{\mu}_i, I_d)\right)$. The Maximum Likelihood Estimator is then defined as $\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \arg\max p(X|\boldsymbol{\mu})$. EM algorithm is based on using an auxiliary function to lower bound the log likelihood. Define $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) = \mathbb{E}_X\left[\sum_i p(Z = i|X; \boldsymbol{\mu}^t) \log \phi(X; \boldsymbol{\mu}_i, I_d)\right]$, where Z denote the unobserved component membership of data point X. The standard EM update is $\boldsymbol{\mu}^{t+1} = \arg\max_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t)$. Define

$$w_i(X; \boldsymbol{\mu}) = \frac{\pi_i \phi(X|\boldsymbol{\mu}_i, I_d)}{\sum_{j=1}^r \pi_j \phi(X|\boldsymbol{\mu}_j, I_d)} \tag{2.1}$$

The update step for gradient EM, defined via the gradient operator $G(\boldsymbol{\mu}^t) : \mathbb{R}^{Md} \to \mathbb{R}^{Md}$, is

$$G(\boldsymbol{\mu}^t)^{(i)} := \boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^t + s[\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)]_i = \boldsymbol{\mu}_i^t + s\mathbb{E}_X\left[\pi_i w_i(X; \boldsymbol{\mu}^t)(X - \boldsymbol{\mu}_i^t)\right]. \tag{2.2}$$

where $s > 0$ is the step size and $(.)^{(i)}$ denotes the part of the stacked vector corresponding to the i^{th} mixture component. We will also use $G_n(\boldsymbol{\mu})$ to denote the empirical counterpart of the population gradient operator $G(\mu)$ defined in Eq (2.2). We assume we are given an initialization $\boldsymbol{\mu}_i^0$ and the true mixing weight π_i for each component.

2.2.1 Notations

Define $R_{\max}$ and $R_{\min}$ as the largest and smallest distance between cluster centers i.e., $R_{\max} = \max_{i \neq j} \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|, R_{\min} = \min_{i \neq j} \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|$. Let $\pi_{\max}$ and $\pi_{\min}$ be the maximal and minimal cluster weights, and define κ as $\kappa = \frac{\pi_{\max}}{\pi_{\min}}$.

2.3 Main Results

Despite being a non-convex problem, EM and gradient EM algorithms have been shown to exhibit good convergence behavior in practice, especially with good initializations. However, existing local convergence theory only applies for two-cluster equal-weight GMM. In this section, we present our main result in two parts. First we show the convergence rate and present a near-optimal radius for contraction region for population gradient EM. Then in the second part we connect the population version to finite sample results using concepts from empirical processes and learning theory.

2.3.1 Local contraction for population gradient EM

Intuitively, when $\boldsymbol{\mu}^t$ equals the ground truth $\boldsymbol{\mu}^*$, then the $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^*)$ function will be well-behaved. This function is a key ingredient in [9], where the curvature of the $Q(\cdot|\boldsymbol{\mu})$ function is shown to be close to the curvature of $Q(\cdot|\boldsymbol{\mu}^*)$ when the $\boldsymbol{\mu}$ is close to $\boldsymbol{\mu}^*$. This is a local property that only requires the gradient to be stable at one point.

Definition 2.1 (Gradient Stability)**.** The Gradient Stability (GS) condition, denoted by GS(γ, a), is satisfied if there exists $\gamma > 0$, such that for $\boldsymbol{\mu}_i^t \in \mathbb{B}(\boldsymbol{\mu}_i^*, a)$ with some $a > 0$, for $\forall i \in [r]$.

$$\|\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^*) - \nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)\| \leq \gamma \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$$

The GS condition is used to prove contraction of the sequence of estimators produced by population gradient EM. However, for most latent variable models, it is typically challenging to verify the GS condition and obtain a tight bound on the parameter γ. We derive the GS condition under milder conditions (see Theorem 2.5 in Section 2.4), which bounds the deviation of the partial gradient evaluated at $\boldsymbol{\mu}_i^t$ uniformly over all $i \in [r]$. This immediately implies the global GS condition defined in Definition 2.1. Equipped with this result, we achieve a nearly optimal local convergence radius for general GMMs in Theorem 2.1. The proof of this theorem can be found in Appendix 6.2.2.

Theorem 2.1 (Convergence for Population gradient EM)**.** *Let* $d_0 := \min\{d, r\}$. *If* $R_{\min} = \tilde{\Omega}(\sqrt{d_0})$, *with initialization* $\boldsymbol{\mu}^0$ *satisfying,* $\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\| \leq a, \forall i \in [r]$, *where*

$$a \leq \frac{R_{\min}}{2} - \sqrt{d_0} O\left(\sqrt{\log\left(\max\left\{\frac{r^2\kappa}{\pi_{\min}}, R_{\max}, d_0\right\}\right)}\right)$$

then the Population EM converges:

$$\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\| \leq \zeta^t \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}^*\|, \qquad \zeta = \frac{\pi_{\max} - \pi_{\min} + 2\gamma}{\pi_{\max} + \pi_{\min}} < 1$$

where $\gamma = r^2(2\kappa + 4)\left(2R_{\max} + d_0\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{d_0}/8\right) < \pi_{\min}$.

Remark 2.1. The local contraction radius is largely improved compared to that in [9], which has $R_{\min}/8$ in the two equal sized symmetric GMM setting. It can be seen that in Theorem 2.1, $a/R_{\min}$ goes to $\frac{1}{2}$ as the signal to noise ratio goes to infinity. We will show in simulations that when initialized from some point that lies $R_{\min}/2$ away from the true center, gradient EM only converges to a stationary point which is not a global optimum. More discussion can be found in Section 4.4.

2.3.2 Finite sample bound for gradient EM

In the finite sample setting, as long as the deviation of the sample gradient from the population gradient is uniformly bounded, the convergence in the population setting implies the convergence in finite sample scenario. Thus the key ingredient in the proof is to get this uniform bound over all parameters in the contraction region $\mathbb{A}$, i.e. bound $\sup_{\boldsymbol{\mu} \in \mathbb{A}} \|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\|$, where G and G_n are defined in Section 4.2.

To prove the result, we expand the difference and define the following function for $i \in [r]$, where u is a unit vector on a d dimensional sphere $\mathcal{S}^{d-1}$. This appears because we can write the Euclidean norm of any vector B, as $\|B\| = \sup_{u \in \mathcal{S}^{d-1}} \langle B, u \rangle$.

$$g_i^u(X) = \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^{n} w_1(X_i; \boldsymbol{\mu}) \langle X_i - \boldsymbol{\mu}_1, u \rangle - \mathbb{E} w_1(X; \boldsymbol{\mu}) \langle X - \boldsymbol{\mu}_1, u \rangle. \quad (2.3)$$

We will drop the super and subscript and prove results for g_1^u without loss of generality.

The outline of the proof is to show that $g(X)$ is close to its expectation. This expectation can be further bounded via the Rademacher complexity of the corresponding function class (defined below in Eq (2.4)) by the tools like the symmetrization lemma [80].

Consider the following class of functions indexed by $\boldsymbol{\mu}$ and some unit vector on d dimensional sphere $u \in \mathcal{S}^{d-1}$:

$$\mathcal{F}_i^u = \{ f^i : \mathcal{X} \to \mathbb{R} | f^i(X; \boldsymbol{\mu}, u) = w_i(X; \boldsymbol{\mu}) \langle X - \boldsymbol{\mu}_i, u \rangle \} \quad (2.4)$$

We need to bound the r functions classes separately for each mixture. Given a finite n-sample $(X_1, \cdots, X_n)$, for each class, we define the Rademacher complexity as the expectation of empirical Rademacher complexity.

$$\hat{R}_n(\mathcal{F}_i^u) = \mathbb{E}_\epsilon \left[\sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{j=1}^{n} \epsilon_i f^i(X_j; \boldsymbol{\mu}, u) \right]; \qquad R_n(\mathcal{F}_i^u) = \mathbb{E}_X \hat{R}_n(\mathcal{F}_i^u)$$

where ϵ_i's are the i.i.d. Rademacher random variables.

For many function classes, the computation of the empirical Rademacher complexity can be hard. For complicated functions which are Lipschitz w.r.t functions from a simpler function class, one can use Ledoux-Talagrand type contraction results [68]. In order to use the Ledoux-Talagrand contraction, one needs a 1-Lipschitz function, which we do not have, because our function involves $\boldsymbol{\mu}_i$, $i \in [r]$. Also, the weight functions w_i are not separable in terms of the $\boldsymbol{\mu}_i$'s. Therefore the classical contraction lemma does not apply. In our analysis, we need to introduce a vector-valued function, with each element involving only one $\boldsymbol{\mu}_i$, and apply a recent result of vector-versioned contraction lemma [76]. With some careful analysis, we get the following. The details are deferred to Section 2.5.

Theorem 2.2. *Let $\mathcal{F}_i^u$ be as in Eq.* (2.4) *for $\forall i \in [r]$, then for some universal constant c,*

$$R_n(\mathcal{F}_i^u) \leq \frac{cr^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}}{\sqrt{n}}$$

After getting the Rademacher complexity, one needs to use concentration results like McDiarmid's inequality [78] to achieve the finite-sample bound. Unfortunately for the functions defined in Eq. (2.4), the martingale difference sequence does not have bounded differences. Hence it is difficult to apply McDiarmid's inequality in its classical form. To resolve this, we instead use an extension of McDiarmid's inequality which can accommodate sequences which have bounded differences with high probability [26].

Theorem 2.3 (Convergence for sample-based gradient EM)**.** *Let ζ be the contraction parameter in Theorem 2.1, and*

$$\epsilon^{unif}(n) = \tilde{O}(\max\{n^{-1/2}r^3(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\},(1+R_{\max})d/\sqrt{n}\}). \tag{2.5}$$

If $\epsilon^{unif}(n) \leq (1-\zeta)a$, then sample-based gradient EM satisfies

$$\left\|\hat{\boldsymbol{\mu}}_i^t - \boldsymbol{\mu}_i^*\right\| \leq \zeta^t \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\|_2 + \frac{1}{1-\zeta}\epsilon^{unif}(n); \quad \forall i \in [r]$$

with probability at least $1 - n^{-cd}$, where c is a positive constant.

Remark 2.2. When data is observed in a streaming fashion, the gradient update can be modified into a stochastic gradient update, where the gradient is evaluated based on a single observation or a small batch. By the GS condition proved in Theorem 2.1, combined with Theorem 6 in [9], we immediately extend the guarantees of gradient EM into the guarantees for the stochastic gradient EM.

2.3.3 Initialization

Appropriate initialization for EM is the key to getting good estimation within fewer restarts in practice. There have been a number of interesting initialization algorithms for estimating mixture models. It is pointed out in [54] that in practice, initializing the centers by uniformly drawing from the data is often more reasonable than drawing from a fixed distribution. Under this initialization strategy, we can bound the number of initializations required to find a "good" initialization that falls in the contraction region in Theorem 2.1.

The exact theorem statement and a discussion of random initialization can be found in Appendix 6.4. More sophisticated strategy includes, an approximate solution to k-means on a projected low-dimensional space used in [8] and [63]. While it would be interesting to study different initialization schemes, that is part of future work.

2.4 Local Convergence of Population Gradient EM

In this section we present the proof sketch for Theorem 2.1. The complete proofs in this section are deferred to Appendix 6.2. To start with, we calculate the closed-form characterization of the gradient of $q(\boldsymbol{\mu})$ as stated in the following lemma.

Lemma 2.1. *Define* $q(\boldsymbol{\mu}) = Q(\boldsymbol{\mu}|\boldsymbol{\mu}^*)$. *The gradient of* $q(\boldsymbol{\mu})$ *is* $\nabla q(\boldsymbol{\mu}) = (diag(\pi) \otimes I_d)(\boldsymbol{\mu}^* - \boldsymbol{\mu})$.

If we know the parameter γ in the gradient stability condition, then the convergence rate depends only on the condition number of the Hessian of $q(\cdot)$ and γ.

Theorem 2.4 (Convergence rate for population gradient EM)**.** *If Q satisfies the GS condition with parameter $0 < \gamma < \pi_{\min}$, denote $d_t := \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|$, then with step size $s = \frac{2}{\pi_{\min}+\pi_{\max}}$, we have:*

$$d_{t+1} \leq \left(\frac{\pi_{\max} - \pi_{\min} + 2\gamma}{\pi_{\max} + \pi_{\min}} \right)^t d_0$$

The proof uses an approximation on gradient and standard techniques in analysis of gradient descent.

Remark 2.3. It can be verified that the convergence rate is equivalent to that shown in [9] when applied to GMMs. The convergence slows down as the proportion imbalance $\kappa = \pi_{\max}/\pi_{\min}$ increases, which matches the observation in [84].

Now to verify the GS condition, we have the following theorem.

Theorem 2.5 (GS condition for general GMM)**.** *Let* $\tilde{d} = \min\{d, r\}$ *be the effective dimension. If* $R_{\min} = \tilde{\Omega}(\sqrt{\tilde{d}})$*, and* $\boldsymbol{\mu}_i \in \mathbb{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [r]$ *where*

$$a \leq \frac{R_{\min}}{2} - \sqrt{\tilde{d}\}} \max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3}),$$

then $\|\nabla_{\boldsymbol{\mu}_i} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu})\| \leq \frac{\gamma}{r}\sum_{i=1}^r \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\| \leq \frac{\gamma}{\sqrt{r}}\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$*,*
where $\gamma = r^2(2\kappa + 4)\left(2R_{\max} + \tilde{d}\}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}\}}/8\right)$.
Furthermore, $\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu})\| \leq \gamma\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$.

Proof sketch of Theorem 2.5. W.l.o.g. we show the proof with the first cluster, consider the difference of the gradient corresponding to $\boldsymbol{\mu}_1$.

$$\nabla_{\boldsymbol{\mu}_1} Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t) - \nabla_{\boldsymbol{\mu}_1} q(\boldsymbol{\mu}^t) = \mathbb{E}(w_1(X;\boldsymbol{\mu}^t) - w_1(X;\boldsymbol{\mu}^*))(X - \boldsymbol{\mu}_1^t) \tag{2.6}$$

For any given X, consider the function $\boldsymbol{\mu} \to w_1(X;\boldsymbol{\mu})$, we have

$$\nabla_{\boldsymbol{\mu}} w_1(X;\boldsymbol{\mu}) = \begin{pmatrix} w_1(X;\boldsymbol{\mu})(1 - w_1(X;\boldsymbol{\mu}))(X - \boldsymbol{\mu}_1)^T \\ -w_1(X;\boldsymbol{\mu})w_2(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_2)^T \\ \vdots \\ -w_1(X;\boldsymbol{\mu})w_r(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_r)^T \end{pmatrix} \tag{2.7}$$

Let $\boldsymbol{\mu}^u = \boldsymbol{\mu}^* + u(\boldsymbol{\mu}^t - \boldsymbol{\mu}^*), \forall u \in [0,1]$, obviously $\boldsymbol{\mu}^u \in \Pi_{i=1}^r \mathbb{B}(\boldsymbol{\mu}_i^*, \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|) \subset \Pi_{i=1}^r \mathbb{B}(\boldsymbol{\mu}_i^*, a)$. By Taylor's theorem,

$$\begin{aligned}
&\|\mathbb{E}(w_1(X;\boldsymbol{\mu}_1^t) - w_1(X;\boldsymbol{\mu}_1^*))(X - \boldsymbol{\mu}_1^t)\| = \left\|\mathbb{E}\left[\int_{u=0}^1 \nabla_u w_1(X;\boldsymbol{\mu}^u) du (X - \boldsymbol{\mu}_1^t)\right]\right\| \\
\leq & U_1 \|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\|_2 + \sum_{i \neq 1} U_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2 \leq \max_{i \in [r]}\{U_i\} \sum_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2
\end{aligned} \tag{2.8}$$

where

$$\begin{aligned}
U_1 &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
U_i &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X;\boldsymbol{\mu}^u) w_i(X;\boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_2^u)^T\|_{op}
\end{aligned}$$

Bounding them with careful analysis on Gaussian distribution yields the result. The technical details are deferred to Appendix 6.2. □

2.5 Sample-based Convergence

In this section we present the proof sketch for sample-based convergence of gradient EM. The full proofs in this section are deferred in Appendix 6.3. The main ingredient in proving Theorem 2.3 is the result of the following theorem, which develops an uniform upper bound for the differences between sample-based gradient and population gradient on each cluster center.

Theorem 2.6 (Sample-based EM guarantee)**.** *Denote* $\mathbb{A}$ *as the contraction region* $\Pi_{i=1}^r \mathbb{B}(\boldsymbol{\mu}_i^*, a)$*. Under the condition of Theorem 2.1, with probability at least* $1 - \exp(-cd\log n)$*,*

$$\sup_{\boldsymbol{\mu} \in \mathbb{A}} \left\|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\right\| < \epsilon^{unif}(n); \qquad \forall i \in [r]$$

where

$$\epsilon^{unif}(n) = cr^{3/2}(1 + 3R_{\max})^3 \max\{1, \log(\kappa)\}\sqrt{\frac{d \log n}{n}}. \tag{2.9}$$

Plugging in the expression of G and G_n we recognize the left hand side as $g_i^u(X)$ defined in Eq. (2.3). The quantity $g_1^u(X)$ depends on the sample, the idea for proving Theorem 2.6 is to show it concentrates around its expectation when sample size is large. And its expectation is bounded by the Radamacher complexity. Note that when the function class has bounded differences (changing one data point changes the function by a bounded amount almost surely), as in the case in many risk minimization problems in supervised learning, the McDiarmid's inequality can be used to achieve concentration. However the function class we define in Eq. (2.4) is not bounded almost everywhere, but with high probability, hence the classical result does not apply. Here we prove a concentration inequality following the classical Azuma-Hoeffding / McDiarmid martingale procedure, but with a more careful treatment for the conditional difference utilizing the gaussian tail properties. The proof uses similar techniques as in Theorem 1 of [60]. The following bound improves upon the one shown in [116] and have an optimal rate for dimension.

Theorem 2.7. *Let $g(X)$ be defined in Eq.* (2.3) *with $i = 1$ and some fixed u, then*

$$P\left(g(X) - \mathbb{E}g(X) > 2(1+3R_{\max})\sqrt{\frac{d\log n}{n}}\right) \leq n^{-d}$$

Now it remains to derive the Rademacher complexity under the given function class. Note that when the function class is a contraction, or Lipschitz with respect to another function (usually of a simpler form), one can use the Ledoux-Talagrand contraction lemma [68] to reduce the Rademacher complexity of the original function class to the Rademacher complexity of the simpler function class. This is essential in getting the Rademacher complexities for complicated function classes. As we mention in Section 3.4, our function class in Eq. (2.4) is unfortunately not Lipschitz due to the fact that it involves all cluster centers even for the gradient on one cluster. We get around this problem by introducing a vector valued function, and show that the functions in Eq. (2.4) are Lipschitz in terms of the vector-valued function. In other words, the absolute difference in the function when the parameter changes is upper bounded by the norm of the vector difference of the vector-valued function. Then we build upon the recent vector-contraction result from [76], and prove the following lemma under our setting.

Lemma 2.2. *Let X be nontrivial, symmetric and sub-gaussian. Then there exists a constant $C < \infty$, depending only on the distribution of X, such that for any subset $\mathcal{S}$ of a separable Banach space and function $h_i : \mathcal{S} \to \mathbb{R}$, $f_i : \mathcal{S} \to \mathbb{R}^k$,*

$i \in [n]$ *satisfying* $\forall s, s' \in \mathcal{S}, |h_i(s) - h_i(s')| \leq L\|f(s) - f(s')\|$. *If* ϵ_{ik} *is an independent doubly indexed Rademacher sequence, we have,*

$$\mathbb{E} \sup_{s \in \mathcal{S}} \sum_i \epsilon_i h_i(s) \leq \mathbb{E}\sqrt{2}L \sup_{s \in \mathcal{S}} \sum_{i,k} \epsilon_{ik} f_i(s)_k,$$

where $f_i(s)_k$ *is the* k*-th component of* $f_i(s)$.

Remark 2.4. In contrast to the original form in [76], we have a $\mathcal{S}$ as a subset of a separable Banach Space. The proof uses standard tools from measure theory, and is to be found in Appendix 6.3.

This equips us to prove Theorem 2.2.

Proof sketch of Theorem 2.2. For any unit vector u, the Rademacher complexity of $\mathcal{F}_1^u$ is

$$\begin{aligned} R_n(\mathcal{F}_1^u) =& \mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle X_i - \boldsymbol{\mu}_1, u \rangle \\ \leq & \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle X_i, u \rangle}_{(D)} + \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle \boldsymbol{\mu}_1, u \rangle}_{(E)} \end{aligned} \tag{2.10}$$

We bound the two terms separately. Define $\eta_j(\boldsymbol{\mu}) : \mathbb{R}^{rd} \to \mathbb{R}^r$ to be a vector valued function with the k-th coordinate

$$[\eta_j(\boldsymbol{\mu})]_k = \frac{\|\boldsymbol{\mu}_1\|^2}{2} - \frac{\|\boldsymbol{\mu}_k\|^2}{2} + \langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1 \rangle + \log\left(\frac{\pi_k}{\pi_1}\right)$$

$$\text{It can be shown that} \qquad |w_1(X_j; \boldsymbol{\mu}) - w_1(X_j; \boldsymbol{\mu}')| \leq \frac{\sqrt{r}}{4} \|\eta_j(\boldsymbol{\mu}) - \eta_j(\boldsymbol{\mu}')\| \tag{2.11}$$

Now let $\psi_1(X_j;\boldsymbol{\mu}) = w_1(X_j;\boldsymbol{\mu})\langle X_j, u\rangle$. With Lipschitz property (6.22) and Lemma 6.11, we have

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{j=1}^{n} \epsilon_j w_i(X_j;\boldsymbol{\mu})\langle X_j, u\rangle\right] \leq \mathbb{E}\left[\frac{\sqrt{2}\sqrt{r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}[\eta_j(\boldsymbol{\mu})]_k\right]$$

The right hand side can be bounded with tools regarding independent sum of sub-gaussian random variables. Similar techniques apply to the (E) term. Adding things up we get the final bound. □

Proof of Theorem 2.6. Combining Theorem 2.2, Lemma 6.12 and Theorem 2.7, we have for any d-dimensional unit vector u, with probability at least $1 - n^{-d}$,

$$\begin{aligned} g_i^u(X) \leq & |g_i^u(X) - \mathbb{E}g_i^u(X)| + \mathbb{E}g_i^u(X) \\ \leq & 2(1+3R_{\max})\sqrt{\frac{d\log n}{n}} + 2\mathcal{R}_n(\mathcal{F}_i^u) \\ \leq & cr^{3/2}(1+3R_{\max})^3\max\{1,\log(\kappa)\}\sqrt{\frac{d\log n}{n}} \end{aligned}$$

By standard covering arguments, we have

$$\sup_{\boldsymbol{\mu}\in\mathbb{A}}\left\|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\right\| \leq 2\max_{j=1,\cdots,K} g_i^{u^{(j)}}(X)$$

Using $K \leq e^{2d}$ from Lemma 6.1 along with union bound, we have

$$\sup_{\boldsymbol{\mu}\in\mathbb{A}}\left\|G^{(i)}(\boldsymbol{\mu}) - G_n^{(i)}(\boldsymbol{\mu})\right\| \leq cr^{3/2}(1+3R_{\max})^3\max\{1,\log(\kappa)\}\sqrt{\frac{d\log n}{n}}$$

with probability at least $1 - (ne^{-2})^{-d}$.

□

Combining the pieces we can now prove Theorem 2.3.

Proof of Theorem 2.3. We show the result by induction. When $t = 1$,

$$\begin{aligned}\left\|\boldsymbol{\mu}^1 - \boldsymbol{\mu}^*\right\|_2 = \left\|G_n(\boldsymbol{\mu}^0) - \boldsymbol{\mu}^*\right\| &\leq \left\|G(\boldsymbol{\mu}^0) - \boldsymbol{\mu}^*\right\| + \left\|G_n(\boldsymbol{\mu}^0) - G(\boldsymbol{\mu}^0)\right\| \\ &\leq \zeta \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \epsilon^{\text{unif}}(n)\end{aligned}$$

If $\|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\| < a$ and $\epsilon^{\text{unif}}(n) \leq (1-\zeta)a$, we have $\left\|\boldsymbol{\mu}_i^{t+1} - \boldsymbol{\mu}_i^*\right\| \leq a$. So $\boldsymbol{\mu}^t$ lies in the contraction region for $\forall t \geq 0$.

Then iteratively we get

$$\begin{aligned}\left\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\right\| &\leq \zeta \left\|\boldsymbol{\mu}^{t-1} - \boldsymbol{\mu}^*\right\| + \epsilon^{\text{unif}}(n) \\ &\leq \zeta^t \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \sum_{i=0}^{t-1} \zeta^i \epsilon^{\text{unif}}(n) \\ &\leq \zeta^t \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\| + \frac{1}{1-\zeta}\epsilon^{\text{unif}}(n)\end{aligned}$$

with probability at least $1 - \delta$. □

2.6 Experiments

In this section we collect some numerical results. In all experiments we set the covariance matrix for each mixture component as identity matrix I_d and define signal-to-noise ratio (SNR) as $R_{\min}$.

Convergence Rate We first evaluate the convergence rate and compare with those given in Theorem 2.4 and Theorem 2.5. For this set of experiments, we use a mixture of 3 Gaussians in 2 dimensions. In both experiments $R_{\max}/R_{\min} = 1.5$. In different settings of $\boldsymbol{\pi}$, we apply gradient EM with varying SNR from 1 to 5. For each choice of SNR, we perform 10 independent

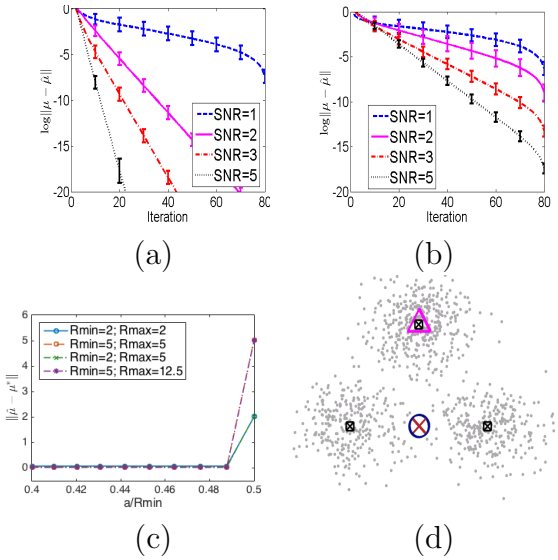


Figure 2.1: (a, b): The influence of SNR on optimization error in different settings. The figures represent the influence of SNR when the GMMs have different cluster centers and weights: (a) $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. (b) $\boldsymbol{\pi} = (0.6, 0.3, 0.1)$. (c) plots statistical error with different initializations arbitrarily close to the boundary of the contraction region. (d) shows the suboptimal stationary point when two centers are initialized from the midpoint of the respective true cluster centers.

trials with $N = 12,000$ data points. The average of $\log \|\boldsymbol{\mu}^t - \hat{\boldsymbol{\mu}}\|$ and the standard deviation are plotted versus iterations. In Figure 2.1 (a) and (b) we plot balanced $\boldsymbol{\pi}$ ($\kappa = 1$) and unbalanced $\boldsymbol{\pi}$ ($\kappa > 1$) respectively.

All settings indicate the linear convergence rate as shown in Theorem 2.4. As SNR grows, the parameter γ in GS condition decreases and thus yields faster convergence rate. Comparing left two panels in Figure 2.1, increasing imbalance of cluster weights κ slows down the local convergence rate as shown in Theorem 2.4.

Contraction Region To show the tightness of the contraction region, we generate a mixture with $r = 3, d = 2$, and initialize the clusters as follows. We use $\boldsymbol{\mu}_2^0 = \frac{\boldsymbol{\mu}_2^* + \boldsymbol{\mu}_3^*}{2} - \epsilon$, $\boldsymbol{\mu}_3^0 = \frac{\boldsymbol{\mu}_2^* + \boldsymbol{\mu}_3^*}{2} + \epsilon$, for shrinking ϵ, i.e. increasing $a/R_{\min}$ and plot the error on the Y axis. Figure 2.1-(c) shows that gradient EM converges when initialized arbitrarily close to the boundary, thus confirming our near optimal contraction region. Figure 2.1-(d) shows that when $\epsilon = 0$, i.e. $a = \frac{R_{\min}}{2}$, gradient EM can be trapped at a sub-optimal stationary point.

2.6.1 Conclusion for analysis of EM algorithm

In previous sections, we have stated population and finite-sample based local convergence results for the non-convex EM algorithm. In the following sections, we study the convex relaxation of k-means in a high dimensional model, and study its properties in the presence of outliers.

2.7 Robust Convex Relaxation for Covariate Clustering

The EM algorithm is one of the oldest clustering algorithm. Despite its popularity in practitioners, it is non-convex and is sensitive to initialization. Now we discuss some convex relaxations of a well-known clustering loss. K-means, named by James MacQueen [73], was proposed by Hugo Steinhaus [101] before. Despite being half a century old, k-means has been widely used and analyzed under various settings.

One major drawback of k-means is its incapability to separate clusters that are non-linearly separated, and that the loss is non-convex. This can be

alleviated by mapping the data to a high dimensional feature space and do clustering on top of the feature space [98, 36, 58], which is generally called kernel-based methods. For instance, the widely-used spectral clustering [99, 86] is an algorithm to calculate top eigenvectors of a kernel matrix of affinities, followed by a k-means on the top r eigenvectors. The consistency of spectral clustering is analyzed by [106]. [36] shows that spectral clustering is essentially equivalent to a weighted version of kernel k-means.

The performance guarantee for clustering is often studied under distributional assumptions; usually a mixture model with well-separated centers suffices to show consistency. In [29], the authors use a Gaussian mixture model, and proposes a variant of EM algorithm that provably recovers the center of each Gaussian when the minimum distance between clusters is greater than some multiple of the square root of dimension. In [8], the authors work with a projection based algorithm and shows the separation needs to be greater than the operator norm and the Frobenius norm of difference between data matrix and its corresponding center matrix, up to a constant.

The non-convexity is often handled by using convex relaxations [79]. For example, SDP relaxations for k-means typed clustering were proposed in [61, 92]. In a very recent work, it is shown in [79] that the effectiveness of SDP relaxation with k-means clustering for subgaussian mixtures, provided the minimum distance between centers is greater than the variance of the sub-gaussian times the square of the number of clusters r.

On a related note, SDP relaxations have been shown to be consistent

for community detection in networks [6, 17]. In particular, Cai et al. [17] consider "inlier" (these are generated from the underlying clustering model, to be specific, a blockmodel) and "outlier" nodes. The authors show that SDP is weakly consistent in terms of clustering the inlier nodes as long as the number of outliers m is a vanishing fraction of the number of nodes.

In contrast, among the numerous work on clustering, not much focus has been on robustness of different kernel k-means algorithms in presence of arbitrary outliers. Yang et al. [118] illustrate the robustness of Gaussian kernel based clustering, where no explicit upper bound is given. Debruyne et al. [34] detect the influential points in kernel PCA by looking at an influence function. In data mining community, many find clustering can be used to detect outliers, with often heuristic but effective procedures [89, 37]. On the other hand, kernel based methods have been shown to be robust for many machine learning tasks. For supervised learning, it is shown in [112] that the robustness of SVM by introducing an outlier indicator and relaxing the problem to a SDP. [32, 33, 25] develop the robustness for kernel regression. For unsupervised learning, [59] proposes a robust kernel density estimation algorithms.

In the remaining part of this chapter, we ask the question: how robust are SVD type algorithms and SDP relaxations when outliers are present. In the process we also present results which compare these two methods. To be specific, we show that without outliers, SVD is weakly consistent, i.e. the *fraction* of misclassified nodes vanishes with high probability, whereas SDP is strongly consistent, i.e. the *number* of misclassified nodes vanishes with

high probability. We also prove that both methods are robust to arbitrary outliers as long as the number of outliers is growing at a slower rate than the number of nodes. Surprisingly our results also indicate that SDP relaxations are more resilient to outliers than K-SVD methods. In Section 2.8 we set up the problem and the data generating model. We present the main results in Section 2.9. Proof sketch and more technical details are introduced in Section 2.10. Numerical experiments in Section 2.11 illustrate and support our theoretical analysis.

2.8 Problem Setup for High-dimensional Sub-Gaussian Mixture

We denote by $Y = [Y_1, \cdots, Y_n]^T$ the $n \times p$ data matrix. Among the n observations, m outliers are distributed arbitrarily, and $n - m$ inliers form r equal-sized clusters, denoted by $C_1, \cdots, C_r$. Let us denote the index set of inliers by $\mathcal{I}$ and index set of outliers by $\mathcal{O}$, $\mathcal{I} \cup \mathcal{O} = [n]$. Also denote by $\mathcal{R} = \{(i, j) : i \in \mathcal{O} \text{ or } j \in \mathcal{O}\}$.

The problem is to recover the true and unknown data partition. With a slight mis-use of notation, we use X as the clustering matrix and Y as the input data matrix. We will also use Z to denote a binary membership matrix, where $Z = \{0, 1\}^{n \times r}$, $Z_{ik} = 1$ if i belongs to the k-th cluster and 0 otherwise. For convenience we assume the outliers are also arbitrarily equally assigned to r clusters, so that each extended cluster, denoted by $\widetilde{C}_i, i \in [r]$ has exactly n/r points. A ground truth clustering matrix $X_0 \in \mathbb{R}^{n \times n}$ can be achieved by $X_0 =$

ZZ^T. It can be seen that $X_0(i,j) = \begin{cases} 1 & \text{if } i,j \text{ belong to the same cluster;} \\ 0 & \text{otherwise.} \end{cases}$

For the inliers, we assume the following mixture distribution model.

$$\text{Conditioned on } Z_{ia} = 1, \quad Y_i = \mu_a + \frac{W_i}{\sqrt{d}},\ \mathbb{E}[W_i] = 0,\ Cov[W_i] = \sigma_a^2 I_d,$$
$$W_i \text{ are independent sub-gaussian random vectors.}$$

Background materials on sub-gaussian random variables and random vectors can be found in Appendix 7.1. We treat Y as a low dimensional signal hidden in high dimensional noise. More concretely μ_a is sparse and $\|\mu_a\|_0$ does not depend on n or d; as $n \to \infty$, $d \to \infty$. W_i's for $i \in [n]$ are independent. For simplicity, we assume the noise is isotropic and the covariance only depends on the cluster. The sub-gaussian assumption is non-parametric and includes most of the commonly used distribution such as Gaussian and bounded distributions. We include some background materials on sub-gaussian random variables in Appendix 7.1. This general setting for inliers is common and also motivated by many practical problems where the data lies on a low dimensional manifold, but is obscured by high-dimensional noise [38].

We use the kernel matrix based on Euclidean distances between covariates. Our analysis can be extended to inner product kernels as well. From now onwards, we will assume that the function generating the kernel is bounded and Lipschitz.

Assumption 2.1. *For n observations $Y_1, \cdots, Y_n$, the kernel matrix (sometimes also called Gram matrix) K is induced by $K(i,j) = f(\|Y_i - Y_j\|_2^2)$, where*

f satisfies:

$$|f(x)| \leq 1, \ \forall x \quad \textit{and} \quad \exists C_0 > 0, s.t. \ \sup_{x,y} |f(x) - f(y)| \leq C_0 |x - y|.$$

A widely used example that satisfies the above condition is the Gaussian kernel. For simplicity, we will without loss of generality assume

$$K(x, y) = f(\|x - y\|^2) = \exp(-\eta \|x - y\|^2). \tag{2.12}$$

2.8.1 Two kernel clustering algorithms

Kernel clustering algorithms can be broadly divided into two categories; one is based on semidefinite relaxation of the k-means objective function and the other is eigen-decomposition based, like kernel PCA, spectral clustering, etc. In this section we describe these two settings.

SDP relaxation for kernel clustering It is well known [36] that kernel k-means could be achieved by maximizing trace($Z^T K Z$) where Z is the $n \times r$ matrix of cluster memberships. However due to the non-convexity of the constraints, the problem is NP-hard. Several convex relaxations for k-means type loss are proposed in the literature (see [92, 79, 114] for more references). In particular in these settings one maximizes $\langle W, X \rangle$, for some positive semidefinite matrix X, where W is a matrix of similarities between pairwise data points. For classical k-means W_{ij} can be $Y_i^T Y_j$ whereas for k-means in the

kernel space one uses a suitably defined kernel similarity function between the ith and jth covariates.

We analyze the following semidefinite programming relaxation. The same relaxation has been used in stochastic block models [6] but to the best of our knowledge, this is the first time it is used to solve kernel clustering problems and shown to be consistent.

$$\max_X \operatorname{trace}(KX) \qquad \text{(SDP-1)}$$
$$\text{s.t., } X \succeq 0, X \geq 0,\ X\mathbf{1} = \frac{n}{r}\mathbf{1},\ \operatorname{diag}(X) = \mathbf{1}$$

While we use the SDP for equal-sized clusters for ease of exposition, in Chapter 4 we analyze SDP relaxations for unequal cluster sizes.

The clustering procedure is listed in Algorithm 1.

Algorithm 1 SDP relaxation for kernel clustering

Require: Observations $Y_1, \cdots, Y_n$, kernel function f.

1: Compute kernel matrix K where $K(i,j) = f(\|Y_j - Y_j\|_2^2)$;
2: Solve SDP-1 and let $\hat{X}$ be the optimal solution;
3: Do k-means on the r leading eigenvectors U of $\hat{X}$.

Kernel singular value decomposition Kernel singular value decomposition (K-SVD) is a spectral based clustering approach. One first does SVD on the kernel matrix, then applies k-means on first r eigenvectors. Different variants include K-PCA [98], which uses singular vectors of centered kernel matrix and spectral clustering [86], which uses singular vectors of normalized graph

laplacian of the kernel matrix. The detailed algorithm is shown in Algorithm 2.

Algorithm 2 K-SVD (K-PCA, spectral clustering)

Require: Observations $Y_1, \cdots, Y_n$, kernel function f.

1: Compute kernel matrix K where $K(i,j) = f(\|Y_j - Y_j\|_2^2)$;
2: **if** K-PCA **then**
3: $\quad K \leftarrow K - K11^T/n - 11^T K/n + 11^T K11^T/n^2$;
4: **else if** spectral clustering **then**
5: $\quad K \leftarrow D^{-1/2}KD^{-1/2}$ where $D = \text{diag}(K1_n)$;
6: **end if**
7: Do k-means on the r leading singular vectors V of K.

2.9 Main Results on Robustness of Kernel Clustering

In this section we summarize our main results in analyzing SDP relaxation of kernel k-means and K-SVD type methods. Our main contribution is two-fold. First, we show that SDP relaxation produces strongly consistent results, i.e. the number of misclustered nodes goes to zero with high probability when there are no outliers, without rounding. On the other hand, K-SVD is weakly consistent, i.e. fraction of misclassified nodes goes to zero when there are no outliers.

In presence of outliers, we see an interesting dichotomy in the behaviors of these two methods. We present upper bounds on the number of outliers, such that the output does not contain clusters that are purely consist of outliers. We see that SDP can tolerate more outliers than K-SVD. When the number of outliers is controlled, both methods can be proven to be weakly

consistent in terms of misclassification error. However, SDP is more resilient to the effect of outliers than K-SVD, if the number of clusters grows or if the separation between the cluster means decays.

Our analysis is organized as follows. First we present a result on the concentration of kernel matrix around its population counterpart. The population kernel matrix for inliers is blockwise constant with r blocks (except the diagonal, which is one). Next we prove that as n increases, the optima $\hat{X}$ of (SDP-1) converges strongly to X_0, when there are no outliers and weakly if the number of outliers grows slowly with n. Then we show the eigenvectors of $\hat{X}$ and K are close to those of their reference matrices, which are piecewise constant aligned with the true clustering structure. We further analyze the k-means step with the eigenvectors as input, to present the conditions on the number of outliers, under which the inliers are clustered into exactly r clusters. Finally we show the mis-clustering error of the clustering returned by Algorithm 1 goes to zero with probability tending to one as $n \to \infty$ when there are no outliers; and when the number of outliers is growing slowly with n, the fraction of mis-clustered nodes from algorithms 1 and 2 converges to zero.

We will start with the concentration of the kernel matrix. We show that under our data model Eq. (2.12) the empirical kernel matrix with the Gaussian kernel restricted on inliers concentrates around a "population" matrix $\tilde{K}^{\mathcal{I}\times\mathcal{I}}$, and the ℓ_∞ norm of $K_f^{\mathcal{I}\times\mathcal{I}} - \tilde{K}_f^{\mathcal{I}\times\mathcal{I}}$ goes to zero at the rate of $O(\sqrt{\frac{\log d}{d}})$. We extend the $\tilde{K}$ on the outlier points to be consistent with Z.

Theorem 2.8. *Let $d_{k\ell} = \|\mu_k - \mu_\ell\|$, and $Z_i = k, Z_j = \ell$, define*

$$\tilde{K}_f(i,j) = \begin{cases} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2) & \text{if } i \neq j, \\ f(0) & \text{if } i = j. \end{cases} . \tag{2.13}$$

Then there exists constant $\rho > 0$, such that with probability at least $1 - n^2 d^{-\rho c^2}$,

$$\sup_{i,j \in \mathcal{I}} |K_{ij} - \tilde{K}_{ij}| \leq c\sqrt{\frac{\log d}{d}}.$$

Remark 2.5. Setting $c = \sqrt{\frac{3 \log n}{d \log d}}$, there exists constant $\rho > 0$, such that $P\left(\|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_\infty \geq \sqrt{\frac{3\log n}{\rho d}}\right) \leq \frac{1}{n}$. The error probability goes to zero for a suitably chosen constant as long as d is growing faster than $\log n$.

While our analysis is inspired by [38], there are two main differences. First we have a mixture model where the population kernel is blockwise constant. Second, we obtain $\sqrt{\frac{\log d}{d}}$ *rates* of convergence by carefully bounding the tail probabilities. In order to attain this we further assume that the noise is sub-gaussian and isotropic. From now on we will drop the subscript f and refer to the kernel matrix as K.

By definition, $\tilde{K}$ is blockwise constant with r unique rows (except the diagonal elements which are ones). Let B be‘ the $r \times r$ Gaussian kernel matrix generated by the centers. An important property of $\tilde{K}$ is that $\lambda_r - \lambda_{r+1}$ (where λ_i is the i^{th} largest eigenvalue of $\tilde{K}$) will be $\Omega(n\lambda_{\min}(B)/r)$.

Lemma 2.3. *If the scale parameter in Gaussian kernel is non-zero, and none of the clusters shares a same center, let B be the $r \times r$ matrix where $B_{k\ell} = f(\|\mu_k - \mu_\ell\|)$, then*

$$\lambda_r(\tilde{K}) - \lambda_{r+1}(\tilde{K}) \geq \frac{n}{r}\lambda_{\min}(B) \cdot \min_k \left(f(\sigma_k^2)\right)^2 - 2\max_k(1 - f(2\sigma_k^2)) = \Omega(n\lambda_{\min}(B)/r)$$

Now we present our result on the consistency of (SDP-1). To this end, we will upper bound $\|\hat{X} - X_0\|_1$, where $\hat{X}$ is the optima returned by (SDP-1) and X_0 is the true clustering matrix. We first present a lemma, which is crucial to the proof of the theorem. Before doing this, we define

$$\gamma_{k\ell} := f(2\sigma_k^2) - f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2); \qquad \gamma_{\min} := \min_{\ell \neq k} \ \gamma_{k\ell} \tag{2.14}$$

The first quantity $\gamma_{k\ell}$ measures separation between the two clusters k and ℓ. The second quantity measures the smallest separation possible. We will assume that γ_{min} is positive. This is very similar to the analysis in asymptotic network analysis where strong assortativity is often assumed. Our results show that the consistency of clustering deteriorates as γ_{min} decreases.

Lemma 2.4. *Let $\hat{X}$ be the solution to* (SDP-1)*, then*

$$\|X_0 - \hat{X}\|_1 \leq \frac{2\langle K - \tilde{K}, \hat{X} - X_0 \rangle}{\gamma_{\min}} \tag{2.15}$$

Combining the above with the concentration of K from Theorem 2.8 we have the following result:

Theorem 2.9. *When $d_{k\ell}^2 > |\sigma_k^2 - \sigma_\ell^2|, \forall k \neq \ell$, and $\gamma_{\min} = \Omega\left(\sqrt{\frac{\log d}{d}}\right)$ then for some absolute constant $c > 0$, $\|X_0 - \hat{X}\|_1 \leq \max\left\{o_P(1), o_P\left(\frac{mn}{r\gamma_{\min}}\right)\right\}$.*

Remark 2.6. When there's no outlier in the data, i.e., $m = 0$, $\hat{X} = X_0$ with high probability and SDP-1 is strongly consistent without rounding. When $m > 0$, the right hand side of the inequality is dominated by mn/r. Note that $\|X_0\|_1 = \frac{n^2}{r}$, therefore after suitable normalization, the error rate goes to zero with rate $O(m/(n\gamma_{min}))$ when $n \to \infty$.

Although $\hat{X}$ is consistent to the ground truth clustering matrix, in practice one often wants to get the labeling in addition to the X_0. Therefore it is usually needed to carry out the last eigen-decomposition step in Algorithm 1. Since X_0 is the clustering matrix, its principal eigenvectors are blockwise constant. In order to show small mis-clustering error one needs to show that the eigenvectors of $\hat{X}$ are converging (modulo a rotation) to those of X_0. This is achieved by a careful application of Davis-Kahan theorem, a detailed discussion of which is deferred to the analysis in Section 2.10.

The Davis-Kahan theorem lets one bound the deviation of the r principal eigenvectors $\hat{U}$ of a Hermitian matrix $\hat{M}$, from the r principal eigenvectors U of M as : $\|\hat{U} - UO\|_F \leq 2^{3/2}\|M - \hat{M}\|_F/(\lambda_r - \lambda_{r+1})$ [119], where λ_r is the r^{th} largest eigenvalue of M and O is the optimal rotation matrix. For a complete statement of the theorem see Appendix 7.6.

Applying the result to X_0 and $\tilde{K}$ provides us with two different upper bounds on the distance between leading eigenvectors. We will see in Theorem 2.11 that the eigengap derived by two algorithms differ, which results in different tolerance for number of outliers and upper bounds for number

of misclustered nodes. Since the Davis-Kahan bounds are tight up-to a constant [119], despite being upper bounds, this indicates that algorithm 1 is less sensitive to the separation between cluster means than Algorithm 2.

To analyze the k-means step with eigenvectors being the input, note that k-means assigns each row of $\hat{U}$ (input eigenvectors of K or $\hat{X}$) to one of r clusters. One of the common hurdles for clustering with outliers is that one mistakenly takes the outliers as separate clusters and miss out or merge the inlier clusters in the k-means step. Let $c_1 \cdots, c_n \in \mathbb{R}^r$ be defined such that c_i is the centroid corresponding to the i^{th} row of $\hat{U}$, and $\{c_i\}_{i=1}^n$ have exactly r unique vectors. Similarly, for the population eigenvectors U (top r eigenvectors of $\tilde{K}$ or X_0), we define the population centroids as $(Z\nu)_i$, for some $\nu \in \mathbb{R}^{r\times r}$. The following theorem shows that as long as the number of outliers is not too large, then the inliers will not lie in smaller than r clusters.

Theorem 2.10. *Let $\hat{V} \in \mathbb{R}^{n\times r}$ be the input eigenvectors of k-means and V be some eigenvectors of $n \times r$ such that V has r unique rows. Assume there exists rotation matrix O such that $\|VO - \hat{V}\| \leq u_{\hat{V}}$. If $3u_{\hat{V}}^2 + 2\frac{mr}{n} < 1$, then each cluster will have at least one inlier.*

The upper bound $u_{\hat{V}}^2$ can vary for different algorithms, and it is a function of m and the eigengap of the population matrix. When we apply the upper bound generated from the Davis-Kahan Theorem, we can get some explicit sufficient condition for m, as stated in the following corollary.

Corollary 2.1. *1. Algorithm 1 returns exactly r inlier clusters if $m < \frac{C_1 n \gamma_{\min}}{r}$;*

2. Assume $\frac{d}{\log d} > 2r + \frac{Cn^2}{(\lambda_r(\tilde{K}) - (\lambda_{r+1}(\tilde{K}))}$, then Algorithm 2 returns exactly r inlier clusters as long as $m < \frac{C_2 n}{\frac{n^2}{(\lambda_r - \lambda_{r+1})^2} + C'r}$. In particular, when all clusters share the same variance, all clusters returned by Algorithm 2 contain inliers if $m < \frac{C_3 n \gamma_{\min}^2}{r^2}$.

Theorem 2.10 and Corollary 2.1 are proved in Appendix 7.7.

We now show that when the empirical centroids are close to the population centroids with a rotation, then the node will be correctly clustered.

We give a general definition of a superset of the misclustered nodes applicable both to K-SVD and SDP:

$$\mathcal{M} = \{i : \|c_i - Z_i \nu O\| \geq 1/\sqrt{2n/r}\} \tag{2.16}$$

Theorem 2.11. *Let $\mathcal{M}_{sdp}$ and $\mathcal{M}_{ksvd}$ be defined as Eq. 2.16, where c_i's are generated from Algorithm 1 and 2 respectively. Let λ_r be the r^{th} largest eigenvalue value of $\tilde{K}'$. We have:*

$$|\mathcal{M}_{sdp}| \leq \max\left\{o_P(1), O_P\left(\frac{m}{\gamma_{\min}}\right)\right\}$$
$$|\mathcal{M}_{ksvd}| \leq O_P \max\left\{\frac{mn^2}{r(\lambda_r - \lambda_{r+1})^2}, \frac{n^3 \log d}{rd(\lambda_r - \lambda_{r+1})^2}\right\}$$

Remark 2.7. Getting a bound for λ_r in terms of $\gamma_{\min}$ for general blockwise constant matrices is difficult. But as shown in Lemma 2.3, the eigengap is

$\Omega(n/r\lambda_{min}(B))$. Plugging this back in we have,

$$|\mathcal{M}_{ksvd}| \leq \max\left\{O_P\left(\frac{mr}{\lambda_{min}(B)^2}\right), O_P\left(\frac{nr\log d/d}{\lambda_{\min}(B)^2}\right)\right\}$$

.

In some simple cases one can get explicit bounds for λ_r, and we have the following.

Corollary 2.2. *Consider the special case when all clusters share the same variance σ^2 and $d_{k\ell}$ are identical for all pairs of clusters. The number of mis-clustered nodes of K-SVD is upper bounded by:*

$$|\mathcal{M}_{ksvd}| \leq \max\left(O_P\left(\frac{mr}{\gamma_{\min}^2}\right), O_P\left(\frac{nr\log d/d}{\gamma_{\min}^2}\right)\right) \tag{2.17}$$

Corollary 2.2 is proved in Appendix 7.9.

Remark 2.8. The situation may happen if cluster center for a is of the form ce_a where e_a is a binary vector with $e_a(i) = \mathbf{1}_{a=i}$. In this case, the algorithm is weakly consistent (fraction of misclassified nodes vanish) when $\gamma_{\min} = \Omega\left(\max\{\sqrt{\frac{r\log d}{d}}, \sqrt{\frac{mr}{n}}\}\right)$. Compared to $|\mathcal{M}_{sdp}|$, $|\mathcal{M}_{ksvd}|$ an additional factor of $\frac{r}{\gamma_{\min}}$. With same m, n, the algorithm has worse upper bound of errors and is more sensitive to $\gamma_{\min}$, which depends both on the data distribution and the scale parameter of the kernel. The proposed SDP can be seen as a denoising procedure which enlarges the separation. It succeeds as long as the denoising is faithful, which requires much weaker assumptions.

2.10 Proof of the main results

In this section, we show the proof sketch of the main theorems. The full proofs are deferred to supplementary materials.

2.10.1 Proof of Theorem 2.8

In Theorem 2.8, we show that if the data distribution is subgaussian, the ℓ_∞ norm of $K - \tilde{K}$ concentrate with rate $O(\sqrt{\frac{\log d}{d}})$.

Proof sketch. With the Lipschitz condition, it suffices to show $\|Y_i - Y_j\|_2^2$ concentrates to $d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2$. To do this, we decompose $\|Y_i - Y_j\|_2^2 = \|\mu_k - \mu_\ell\|_2^2 + 2\frac{(W_i - W_j)^T}{\sqrt{d}}(\mu_k - \mu_\ell) + \frac{\|W_i - W_j\|_2^2}{d}$. Now it suffices to show the third term concentrates to $\sigma_k^2 + \sigma_\ell^2$ and the second term concentrates around 0. Note the fact that $W_i - W_j$ is sub-gaussian, its square is sub-exponential. With sub-gaussian tail bound and a Bernstein type inequality for sub-exponential random variables, we prove the result. □

With the elementwise bound, the Frobenius norm of the matrix difference is just one more factor of n.

Corollary 2.3. *With probability at least* $1 - n^2 d^{-\rho c^2}$, $\|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_F \leq cn\sqrt{\log d/d}$.

2.10.2 Proof of Theorem 2.9

Lemma 2.4 is proved in Appendix 7.4, where we make use of the optimality condition and the constraints in SDP-1. Equipped with Lemma 2.4

we're ready to prove Theorem 2.9.

Proof sketch. In the outlier-free ideal scenario, Lemma 2.4 along with the dualtiy of ℓ_1 and ℓ_∞ norms we get $\|\hat{X} - X_0\|_1 \leq \frac{2\|K-\tilde{K}\|_\infty \|\hat{X}-X_0\|_1}{\gamma_{\min}}$. Then by Theorem 2.8, we get the strong consistency result. When outliers are present, we have to derive a slightly different upper bound. The main idea is to divide the matrices into two parts, one corresponding to the rows and columns of inliers, and the other corresponding to those of the outliers. Now by the concentration result (Theorem 2.8) on K along with the fact that both the kernel function and $X_0, \hat{X}$ are bounded by 1; and the rows of $\hat{X}$ sums to n/r because of the constraint in SDP-1, we obtain the proof. The full proof is deferred to Appendix 7.5. □

2.10.3 Proof of Theorem 2.11

Although Theorem 2.9 provides insights on how close the recovered matrix $\hat{X}$ is to the ground truth, it remains unclear how the final clustering result behaves. In this section, we bound the number of misclassified points by bounding the distance in eigenvectors of $\hat{X}$ and X_0. We start by presenting a lemma that provides a bound for k-means step.

K-means is a non-convex procedure and is usually hard to analyze directly. However, when the centroids are well-separated, it is possible to come up with sufficient conditions for a node to be correctly clustered. When the set of misclustered nodes is defined as Eq. 2.16, the cardinality of $\mathcal{M}$ is directly upper bounded by the distance between eigenvectors. To be explicit, we have

the following lemma. Here $\hat{U}$ denotes top r eigenvectors of K for K-SVD and $\hat{X}$ for SDP. U denotes the top r eigenvectors of $\tilde{K}$ for K-SVD and X_0 for SDP. O denotes the corresponding rotation that aligns the empirical eigenvectors to their population counterpart.

Lemma 2.5. *$\mathcal{M}$ is defined as Eq. (2.16), then $|\mathcal{M}| \leq \frac{8n}{r}\|\hat{U} - UO\|_F^2$.*

Lemma 2.5 is proved in Appendix 7.8.

Analysis of $|\mathcal{M}_{sdp}|$: In order to get the deviation in eigenvectors, note the r^{th} eigenvalue of X_0 is n/r, and $r + 1^{th}$ is 0, let $U \in \mathbb{R}^{n\times r}$ be top r eigenvectors of X and $\hat{U}$ be eigenvectors of X_0. By applying Davis-Kahan Theorem, we have

$$\exists O, \|\hat{U} - UO\|_F \leq \frac{2^{3/2}\|\hat{X} - X_0\|_F}{n/r} \leq \frac{\sqrt{8\|\hat{X} - X_0\|_1}}{n/r} = O_P\left(\sqrt{\frac{mr}{n\gamma_{\min}}}\right) \tag{2.18}$$

Applying Lemma 2.5,

$$|\mathcal{M}_{sdp}| \leq \frac{8n}{r}\left(\frac{2^{3/2}\|\hat{X} - X_0\|_F}{n/r}\right)^2 \leq \frac{cn}{r}\left(\sqrt{\frac{mr}{n\gamma_{\min}}}\right)^2 \leq O_P\left(\frac{m}{\gamma_{\min}}\right)$$

Analysis of $|\mathcal{M}_{ksvd}|$: In the outlier-present kernel scenario, by Corollary 2.3,

$$\|K - \tilde{K}\|_F \leq \|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_F + \|K^{\mathcal{R}} - \tilde{K}^{\mathcal{R}}\|_F = O_P(n\sqrt{\log d/d}) + O_P(\sqrt{mn})$$

Again by Davis-Kahan theorem, and the eigengap between λ_r and λ_{r+1} of $\tilde{K}$ from Lemma 2.3, let U be the matrix with rows as the top r eigenvectors

of $\tilde{K}$. Let $\hat{U}$ be its empirical counterpart.

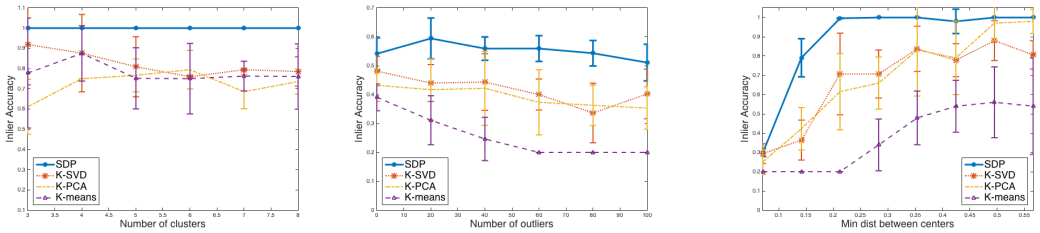
$$\exists O, \|\hat{U} - UO\|_F \leq \frac{2^{3/2}\|K - \tilde{K}\|_F}{\lambda_r - \lambda_{r+1}} \leq O_P\left(\frac{\max\{\sqrt{mn}, n\sqrt{\log d/d}\}}{\lambda_r - \lambda_{r+1}}\right) \quad (2.19)$$

Now we apply Lemma 2.5 and get the upper bound for number of misclustered nodes for K-SVD.

$$\begin{aligned}
|\mathcal{M}_{ksvd}| \leq & \frac{8n}{r}\left(\frac{2^{3/2}C\max\{\sqrt{mn}, n\sqrt{\log d/d}\}}{\lambda_r(\tilde{K}) - \lambda_{r+1}(\tilde{K})}\right)^2 \\
\leq & \frac{Cn}{r}\max\left\{\left(\frac{\sqrt{mn}}{\lambda_r - \lambda_{r+1}}\right)^2, \frac{n^2\log d}{d(\lambda_r - \lambda_{r+1})}\right\} \\
\leq & O_P\max\left\{\frac{mn^2}{r(\lambda_r - \lambda_{r+1})^2}, \frac{n^3\log d}{rd(\lambda_r - \lambda_{r+1})^2}\right\}
\end{aligned}$$

2.11 Experiments for Robustness of Kernel Clustering

In this section, we collect some numerical results. For implementation of the proposed SDP, we use Alternating Direction Method of Multipliers that is used in [6]. In each synthetic experiment, we generate $n - m$ inliers with r equal-sized clusters. The centers of the clusters are sparse and hidden in a p-dim noise. For each generated data matrix, we add in m observations of outliers. To capture the arbitrary nature of the outliers, we generate half the outliers by a random Gaussian with large variance (3 times of the signal), and the other half by a uniform distribution that scatters across all clusters. We compare Algorithm 1 with 1) k-means by Lloyd's algorithms; 2) kernel SVD and 3) kernel PCA by [98]. For all methods, we assume the number of clusters r is known. In practice when dealing with outliers, it is natural to assume



(a) # clusters (b) # outliers (c) Separation

Figure 2.2: Performance vs parameters: (a) Inlier accuracy vs number of cluster ($n = p = 1500, m = 10, d^2 = 0.125, \sigma = 1$); (b) Inlier accuracy vs number of outliers ($n = 1000, r = 5, d^2 = 0.02, \sigma = 1, p = 500$); (c) Inlier accuracy vs separation ($n = 1000, r = 5, m = 50, \sigma = 1, p = 1000$).

there is an extra cluster accounting for outliers, so we cluster both K-SVD and K-PCA with r clusters and $r + 1$ clusters.

The evaluating metrics are accuracy of inliers, i.e., number of correctly clustered nodes divided by the total number of inliers. To avoid the identification problem, we search for all permutations mapping predicted labels to ground truth labels and record the best accuracy. Each set of parameter is run 10 replicates and the mean accuracy and standard deviation (shown as error bars) are reported. For all k-means used in the experiments we do 10 restarts and choose the one with largest objective.

For each experiment, we change only one parameter and fix all the others. Figure 2.2 shows how the performance of different clustering algorithms change when (a) number of clusters (b) number of outliers (c) minimum distance between clusters increases. The value of all parameters used are specified in the caption of the figure. Setting number of clusters as $r + 1$ doesn't help with clustering the inliers, which is observed in all experiments, the curve is

then not shown here.

Panel (a) shows the inlier accuracy for various methods as we increase number of clusters. It can be seen that as we increase number of clusters in presence of outliers, the performance of all methods deteriorate except for the SDP, which matches the rate presented in Theorem 2.11. We also examine the ℓ_1 norm of $X_0 - \hat{X}$, which remains stable as the number of clusters increases. Note that the decrease in accuracy for K-SVD might result from the fact that it fails to meet the condition in Corollary 2.1, which is stronger than the condition for SDP. Panel (b) describes the trend with respect to number of outliers. The accuracy of SDP on inliers is almost unaffected by the number of outliers while other methods suffer with large m. Panel (c) compares the performance as the minimum distance between cluster centers changes. Both SDP and K-SVD are consistent as the distance increases. Compared to K-SVD, SDP concentrates faster and with smaller variation across random runs, which matches the analysis given in Section 3.4.

2.12 Discussion

In this chapter, we first analyze the EM algorithm, which optimizes a non-convex loss. We then investigate the consistency and robustness of two kernel-based clustering algorithms. In the first part, we give a tight contraction bound for local convergence of EM, and propose novel techniques in handling finite-sample analysis. In the second half, we show the semidefinite programming relaxation is strongly consistent without outliers and weakly consistent

in presence of arbitrary outliers. We also show that K-SVD is also weakly consistent in that the mis-clustering rate is going to zero as the observation grows and the outliers are of a small fraction of inliers. By comparing two methods, we conclude that although both are robust to outliers, the proposed SDP is less sensitive to the minimum separation between clusters. The experimental result also supports the theoretical analysis.

While we obtain error bounds for SDP for high-dimensional sub-gaussian mixtures [38] in the regime where the kernel matrix concentrates, it is interesting to consider cases where the signal to noise ratio is low and such concentration does not hold. For ease of exposition, we defer this to Chapter 4 where we obtain error rates for general mixture of sub-gaussians and connect the problem with community detection in sparse graphs.

In next chapter, we will look at community detection problems in networks and we will see the semi-definite relaxations can also be used in those problems with provable guarantees.

Chapter 3

Community Detection in Stochastic Block Models

Community detection in networks is a fundamental problem in machine learning and statistics. A variety of important practical problems like analyzing socio-political ties among leading politicians [41], understanding brain graphs arising from diffusion MRI data [12], investigating ecological relationships between different tiers of the food chain [53] can be framed as community detection problems. Much attention has been focused on developing models and methodology to recover latent community memberships. Among generative models, the stochastic block model [51] and its variants ([3] etc.) have attracted a lot of attention, since their simplicity facilitates efficient algorithms and asymptotic analysis [96, 5, 24].

In this chapter, we focus on the widely-used Stochastic Block Model

The content in this chapter was published in Yan, Bowei, Purnamrita Sarkar, and Xiuyuan Cheng. "Provable Estimation of the Number of Blocks in Block Models." In International Conference on Artificial Intelligence and Statistics, pp. 1185-1194. 2018. Prof. Sarkar proposed the problem of finding the number of blocks in a block model automatically. I mostly did the theoretical analysis independently with a little help from the other authors. I wrote the paper, and implemented the methodology. Prof. Sarkar helped in rewriting and revising the draft and brainstormed about experimental settings. Xiuyuan Cheng participated the discussions, and provided feedbacks on the revision of the draft.

(SBM) [51]. The model assumes the probability of an edge between two nodes are completely determined by the unknown cluster memberships of the nodes. Essentially, this imposes stochastic equivalence, i.e. all nodes in the same cluster behave identically in a probabilistic sense. Despite its simplicity, the SBM is used as a building block in more sophisticated models like the Degree Corrected Block Models [57] and Mixed Membership Block Models [4] and has been applied successfully for clustering real world networks.

We use a convex relaxation proposed in [92] and analyze the error for the recovery. We consider two degree regimes in the analysis of random networks. Let n be the number of nodes in the network. When the average degree is $\Omega(\log n)$, we can find consistent algorithm and exact recovery is possible; whereas when the average degree is $\Theta(1)$, no algorithm can find a consistency solution [122]. When the signal is strong enough, one can find a solution that has a non-decreasing error rate which is better than random guess [67, 43, 83].

In the first section, we analyze SBM in the dense regime, whereas the second section is dedicated to the sparse regime. As we will show below, the technique used in both regimes vary significantly. For the dense regime, one could expect exact recovery when the graph is large enough. For semi-definite programming based methods, a common proof technique is construction of primal-dual pairs [17, 21, 44]. In contrast, for sparse regime, a constant fraction of nodes will always be mis-clustered. In this case, the primal-dual witness method will not work, and we use the Grothendieck's inequality to carry out the upper bound.

3.1 Community detection for dense graphs

Although community detection has drawn much attention from both theorists and practitioners, most existing methods require the prior knowledge of the true number of clusters, which is often unavailable in real data applications. In this chapter, we mainly focus on provably estimating the number of clusters in a network.

While it is tempting to use a two-stage procedure [23] where the number of clusters is estimated first and then used as an input for clustering, an erroneous estimation on the number of clusters can deteriorate the clustering accuracy. Instead, we design an algorithm which estimates the true number of clusters and recovers the cluster memberships simultaneously, with provable guarantees.

Semi-definite programming (SDP) relaxations for network clustering have been widely studied and many different formulations have been proposed. It has been empirically observed that these methods have better clustering performance compared to spectral methods [6, 114, 23]. As shown by [22, 6], SDPs arise naturally when the likelihood of a SBM with equal cluster sizes is relaxed. SDP returns a relaxation of the clustering matrix, which is a $n \times n$ (n being the number of nodes) symmetric matrix whose ij^{th} element is one if nodes i and j belong to the same cluster and zero otherwise. We present a detailed discussion on related work in Section 3.3. In this work, we use the SDP formulation proposed by [92], which uses a normalized variant of the clustering matrix. Similar relaxations have been used to study k-means

clustering for sub-gaussian mixtures [79] and SBMs [113].

For community detection in SBM, an algorithm is considered effective if it is asymptotically consistent. There are two types of consistency in the literature. When the number of nodes in the graph is large enough, the network is sufficiently dense, and the signal (usually defined by the separation between intra-cluster probability and inter-cluster probability) is strong enough, *strongly consistent* methods recover the ground truth labels exactly, while the *weakly consistent* methods recover a fraction of labels correctly where the fraction approaches one as n goes to infinity.

There have been a number of SDP relaxations for general unbalanced cluster sizes which have been shown to be strongly consistent [93, 44, 17]. One can argue that these methods readily render themselves to estimation of the number of blocks r. The idea would be to run the SDP with different values of r, and for the correct one the clustering matrix will be the true clustering matrix with high probability. However, all these methods require the knowledge of model parameters. Furthermore, they work in the unequal cluster size setting by introducing an additional penalty term, which requires further tuning. Hence each run with a different choice of r would have an internal tuning step adding to the already expensive computation of the SDP. We propose a formulation that is a) entirely tuning free when the number of clusters is known, and b) when it is unknown, is able to recover the number of clusters and the clustering matrix in one shot.

Furthermore, our method provably works in the weakly assortative set-

ting, whereas the usual necessary separation condition for recovery is that the maximal inter-cluster connecting probability (think of this as noise) is smaller than the minimal intra-cluster connecting probability (the signal) by a certain margin. This separation condition is known as strong assortativity. In contrast, our work only requires that for each node, the probability of connecting to the nodes in its own cluster is greater by a margin than the largest probability of connecting with nodes in other clusters. This property is called weak assortativity. It is not hard to see that weakly assortative models are a superset of strongly assortative models. Weak assortativity was first introduced in [6], who establish exact recovery under this weaker condition for SDPs for blockmodels with *equal sized* communities.

In Sec 4.4 we sketch a rather interesting empirical property of our algorithm (also pointed out in [93]); namely it can identify different granularities of separations as a byproduct. The tuning phase, which we sketch in Section 4.4, finds different substructures of the network as it searches over different tuning parameters. For example, if there are K meta clusters which are more well separated than the rest, then as we tune, we will first find these meta-clusters, and then finer substructures within them. While this is not the main goal of this approach, it indeed makes our approach ideal for exploratory analysis of networks. We also leave the theoretical analysis of finding multi-resolution clusterings for future work.

We will formalize these concepts in Section 4.2 and discuss the related work in more detail in Section 3.3. Section 3.4 contains our main theoretical

contributions and finally, in Section 3.5 we demonstrate the efficacy of our algorithm compared to existing methods on a variety of simulated and real networks.

3.2 Problem Setup and Notations

Assume $(S_1, \cdots, S_r)$ represent a r-partition for n nodes $\{1, \cdots, n\}$. Let $m_i = |S_i|$ be the size of each cluster, and let $m_{\min}$ and $m_{\max}$ be the minimum and maximum cluster sizes respectively. We denote by A the $n \times n$ binary adjacency matrix with the true and unknown membership matrix $Z = \{0,1\}^{n \times r}$,

$$P(A_{ij} = 1|Z) = Z_i^T B Z_j \quad \forall i \neq j, \qquad \text{(SBM}(B, Z)\text{)}$$

$$P(A_{ii} = 0) = 1, \quad Z^T Z = diag(\boldsymbol{m}), \qquad (3.1)$$

where B is a $r \times r$ matrix of within and across cluster connection probabilities and $\boldsymbol{m}$ is a length r vector of cluster sizes. The elements of B can decay with graph size n. From this section to Section 3.5, we focus on the regime where the average expected degree grows faster than logarithm of n. In this regime, it is possible to obtain strong or weak consistency.

Given any block model, the goal for community detection is to recover the column space of Z. For example if we can solve ZZ^T or its normalized variant $Zdiag(\boldsymbol{m})^{-1}Z^T$, then the labels can be recovered from the eigenvectors of the clustering matrix.

The normalized clustering matrix: In this formulation we focus on recovering the following normalized version:

$$X_0 = Z\text{diag}(\boldsymbol{m})^{-1}Z^T \tag{3.2}$$

It can be easily checked that $X_0\mathbf{1}_n = \mathbf{1}_n$, since $Z\mathbf{1}_k = \mathbf{1}_n$. Furthermore, X_0 is positive semi-definite and its trace (which equals its nuclear norm as well) equals the number of clusters r.

Assortativity (strong vs. weak): Assortativity is a condition usually required in membership recovery. The strong assortativity (see Eq. (3.3)) requires the smallest diagonal entry to be greater than the largest off-diagonal entry.

$$\min_k B_{kk} - \max_{k\neq\ell} B_{k\ell} > 0 \tag{3.3}$$

$$\min_k \left(B_{kk} - \max_{\ell\neq k} B_{k\ell} \right) > 0. \tag{3.4}$$

[6] first introduces an SDP that provably achieves exact recovery for weakly assortative models (Eq. (3.4)) with *equal cluster sizes*, i.e., compared with (3.3), weak assortativity only compares the probability within the same row and column; it requires that any given cluster k, should have a larger probability of connecting within itself than with nodes in any other cluster. It is easy to check that strong assortativity indicates weak assortativity and not vice versa.

For any matrix $X \in \mathbb{R}^{n\times n}$, denote $X_{S_kS_\ell}$ as the submatrix of X on indices $S_k \times S_\ell$, and $X_{S_k} := X_{S_k\times S_k}$. Let $\mathbf{1}$ be all one vector, and $\mathbf{1}_{S_k} \in \mathbb{R}^n$

be the indicator vector of S_k, equal to one on S_k and zero elsewhere. The inner product of two matrices is defined as $\langle A, B\rangle = \text{trace}(A^T B)$. We use $\circ$ to denote the Schur (elementwise) product of two matrices. Standard notations for complexity analysis o, O, Θ, Ω will be used. And those with a tilde are to represent the same order ignoring log factors.

3.3 Prior Work on Estimating Number of Communities in a Network and Community Detection with Convex Relaxations

While most community detection methods assume that the number of communities (r) is given apriori, there has been much empirical and some theoretical work on estimating r from networks.

Methods for estimating r: A large class of methods chooses r by maximizing some likelihood-based criterion. While there are notable methods for estimating r for *non*-network structured data from mixture models [91, 47, 11, 90], we will not discuss them here.

Many likelihood-based methods use variants or approximations of Bayesian Information Criterion (BIC); BIC, while a popular choice for model selection, can be computationally expensive since it depends on the likelihood of the observed data. Variants of the Integrated Classification Likelihood (ICL, originally proposed by [11]) were proposed in [31, 65]. Other BIC type criteria are studied in [74, 97, 77].

In [50] a computationally efficient variational Bayes technique is pro-

posed to estimate r. This method is empirically shown to be more accurate than BIC and ICL and faster than Cross Validation based approaches [20]. A Bayesian approach with a new prior and an efficient sampling scheme is used to estimate r in [95]. While the above methods are not provable, a provably consistent likelihood ratio test is proposed to estimate r in [107].

Another class of methods is based on the spectral approach. The idea is to estimate r by the number of "leading eigenvalues" of a suitably normalized adjacency matrix [88, 56, 18, 39]. Of these the USVT estimator [18] uses random matrix theory to estimate r simply by thresholding the empirical eigenvalues of the adjacency matrix appropriately. In [13] it is shown that the informative eigenvalues of the non-backtracking matrix are real-valued and separated from the bulk under the SBM. In [66], the spectrum of the non-backtracking matrix and the Bethe-Hessian operator are used to estimate r, the later being shown to work better for sparse graphs.

Abbe et. al. [1] propose a degree-profiling method achieving the optimal information theoretical limit for exact recovery. This agnostic algorithm first learns a preliminary classification based on a subsample of edges, then adjust the classification for each node based on the degree-profiling from the preliminary classification. However it involves a highly-tuned and hard to implement spectral clustering step (also noted by [93]). It also requires specific modifications when applied to real world networks (as pointed out by the authors) .

In [124], communities are sequentially extracted from a network; the

stopping criterion uses a bootstrapped approximation of the null distribution of the statistic of choice. In [10], the null distribution of a spectral test statistic is derived, which is used to test $r = 1$ vs $r > 1$ at each step of a recursive bipartitioning algorithm. A generalization of this approach for testing a null hypothesis for r blocks can be found in [69]. While the algorithm in [10] often produces over-estimates of r, hypothesis test in [69] depends on a preliminary fitting with an algorithm which exactly recovers the parameters. The final accuracy heavily depends on the accuracy of this fit. Network cross-validation based methods have also been used for selecting r. The cross-validation can be carried out either via node splitting [4], or node-pair splitting [49, 20]; the asymptotic consistency of these methods are shown in [20]. We conclude with a comparison of our approach to other convex relaxations.

Comparison to other convex relaxations In recent years, SDP has drawn much attention in handling community detection problems with Stochastic Block Models. Various of relaxations have been shown to possess strong theoretical guarantees in recovering the true clustering structure without rounding [6, 44, 46, 17, 93, 81, 43]. Most of them aim at recovering a binary clustering matrix, and show that the relaxed SDP will have the ground truth clustering matrix as its unique optimal solution. For unbalanced cluster sizes, an extra penalization is often introduced which requires additional tuning [17, 44, 93]. While one can try different choices of r for these SDPs until achieving exact recovery, the procedure is slower since each run would need another internal

tuning step.

SDP with a normalized clustering matrix was introduced by [92]. They have been used for network clustering [113] and for the relaxation of k-means clustering of non-network structured data [92, 79] .

$$\begin{aligned} \max \quad & \langle A, X \rangle \\ s.t. \quad & X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}, \mathrm{trace}(X) = r \end{aligned} \tag{SDP-PW}$$

However the formulation in [113] requires an additional parameter as an lower bound on the minimum size of the clusters; loose lower bounds can empirically deteriorate the performance. Also the authors only establish weak consistency of the solution.

Some of these methods do not require the knowledge of r in the constraints, but instead have the dependency implicitly. In [21], a convexified modularity minimization for Degree-corrected SBM is proposed, which also works for SBMs as a special case of degree corrected models. The authors suggest one over total number of edges as the default value for the tuning parameter, but when dealing with delicate structures of the network, this suggested value can be sub-optimal and further tuning is required. The procedure also requires r for the final clustering of the nodes via Spectral Clustering from the clustering matrix.

A different convex relaxation motivated by low-rank matrix recovery is studied in [23]. Here, first the eigenspectrum of A is used to estimate r, which is subsequently used to estimate tuning parameters required in the main algorithm. We can also tune the tuning parameter with other heuristics, but as

the theorem in that paper implies, the tuning parameter needs to lie between the minimal intra-cluster probabilities and maximal inter-cluster probabilities, which is only feasible for strongly assortative settings. We provide more details in the experimental section.

Hierarchical clustering structures A phenomenon that has been observed [93, 23] is that convex relaxations can be used to find hierarchical structures in the networks by varying the tuning parameter. In the experimental section we demonstrate this with some examples.

Separation conditions In terms of the separation conditions, most aforementioned convex relaxations are consistent in the dense regime under *strong assortativity* except [6] and [113]. However, [6] only prove exact recovery of clusters for equal sized clusters, whereas [113] only show weak consistency and require the knowledge of additional parameters like the minimum cluster size. [93] shows exact recovery while matching the information theoretical lower bound, which is not the goal of this work.

In this section, we compare our algorithm with noted representatives from the related work. From the Spectral methods, we compare with the USVT estimator and the Bethe Hessian based estimator [66], which has been shown to empirically outperform a variety of other provable techniques like [107] and [20]. For these methods, we first estimate r and then use the Regularized Spectral Clustering [5, 67] algorithm to obtain the final clustering. From the

convex relaxation literature, we compare with [23] and [21], neither of which require r for estimating the clustering except for the final clustering step.

3.4 Main result for community detection with unknown number of clusters

In various SDP relaxations for community detection under SBMs, the objective function is taken as the linear inner product of the adjacency matrix A and the target clustering matrix X, some formulations also have some additional penalty terms. The inner product objective can be derived from several different metrics for the opitimality of the clustering, such as likelihood or modularity. The penalty terms vary depending on what kind of a solution the SDP is encouraged to yield. For example, in low-rank matrix recovery literature, it is common practice to use the nuclear norm regularization to encourage low-rank solution. For a positive semi-definite matrix, the nuclear norm is identical to its trace. When the number of clusters r is unknown, we consider the following SDP.

$$\begin{aligned} \max \quad & \text{trace}(AX) - \lambda \text{trace}(X) \\ s.t. \quad & X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}, \end{aligned} \tag{SDP-λ}$$

where λ is a tuning parameter, and $X \geq 0$ is an element-wise non-negativity constraint. The following theorem guarantees the exact recovery of the ground truth solution matrix, when λ lies in the given range for the tuning parameter.

Theorem 3.1. *Let $\hat{X}$ be the optimal solution of* (SDP-λ) *for $A \sim SBM(B, Z)$ where $m_{\min}$ and $m_{\max}$ denote the smallest and largest cluster sizes respectively.*

Define the separation parameter $\delta = \min_k(B_{kk} - \max_{\ell \neq k} B_{k\ell})$. *If*

$$c_1 \max_k \sqrt{m_k B_{kk}} + c_2 \sqrt{n \max_{k \neq \ell} B_{k\ell}} \leq \lambda \leq m_{\min} \left(\delta - \max_{k,\ell} \sqrt{\frac{B_{k\ell} \log m_k}{m_k}} \right)$$

then $\hat{X} = X_0$ *with probability at least* $1 - n^{-1}$ *provided*

$$\delta \geq 2\sqrt{6 \log n} \max_k \sqrt{\frac{B_{kk}}{m_k}} + 6 \max_{\ell \neq k} \sqrt{\frac{B_{k\ell} \log n}{m_{\min}}} + \frac{c\sqrt{n p_{\max}}}{m_{\min}} \tag{3.5}$$

Remark 3.1. The above theorem controls how fast the different parameters can grow or decay as n grows. For ease of exposition, we will discuss these constraints on each parameter by fixing the others. The number of clusters r can increase with n. In the dense setting, when $B_{kk} = \Theta(1)$, $m_{\min} = \omega(\sqrt{n})$ and $r = o(\sqrt{n})$, which matches with the best upper bound on r from existing literature. Finally when $\max_k B_{kk} = \Theta(\log n/n)$, we note that $m_{\min} = \tilde{\Theta}(n)$ and $r = \tilde{\Theta}(1)$.

We can see from the condition in Theorem 3.1 that the tuning parameter should be of the order $\sqrt{d}$ where d is the average degree. In fact, as shown in the following theorem, when the λ is greater than the operator norm, (SDP-λ) returns a degenerating rank-1 solution. This gives an upper bound for λ.

Proposition 3.1. *When* $\lambda \geq \|A\|_{op}$, *then the solution for* (SDP-λ) *is* $\boldsymbol{1}\boldsymbol{1}^T/n$.

The proof of Proposition 3.1 is to be found in Appendix 8.1.2. Recall the properties of the ground truth clustering matrix defined in Eq. (3.2). If the optimal solution recovers the ground truth X_0 exactly, we can estimate r easily from its trace. Therefore we have the following corollary.

Corollary 3.1. *Let $\hat{X}$ be the optimal solution of* (SDP-λ) *with $A \sim SBM(B, Z)$, where $B \in [0,1]^{r\times r}$. Under the condition in Theorem 3.1, trace$(\hat{X}) = r$ with probability at least $1 - n^{-1}$.*

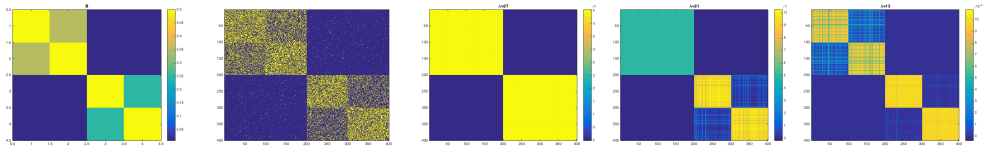
In particular, when r is known, we have the following exact recovery guarantee, which is stronger than the weak consistency result in [113].

Theorem 3.2. *Let $A \sim SBM(B, Z)$, where $B \in [0,1]^{r\times r}$. X_0 is the optimal solution of* (SDP-PW) *with probability at least $1 - n^{-1}$, if the separation condition Eq.* (3.5) *holds true.*

We can see that the two SDPs (SDP-λ) and (SDP-PW) are closely related. In fact, the Lagrangian function of (SDP-PW) is same as the Lagrangian function of (SDP-λ) if we take the lagrangian multiplier for the constraint trace$(X) = r$ as λ. We use this fact in the proof of Theorem 3.1. Both proofs rely on constructing a dual certificate witness, which we elaborate in the following subsection.

3.4.1 Dual Certificate Witness

In this sketch we develop the sufficient conditions with a certain construction of the dual certificate which guarantees X_0 to be the optimal solution.



(a) B (b) Adjacency (c) $\lambda = 27$ (d) $\lambda = 21$ (e) $\lambda = 13$.

Figure 3.1: Solution matrices with various choices of λ.

We derive the main conditions and leave the technical details to the supplementary materials. To start with, the KKT conditions of (SDP-PW) can be written as below.

First Order Stationary

$$-A - \Lambda + (1\alpha^T + \alpha 1^T) + \beta I - \Gamma = 0 \tag{3.6}$$

Primal Feasibility

$$X \succeq 0,\ 0 \leq X \leq 1,\ X\mathbf{1}_n = \mathbf{1}_n,\ \text{trace}(X) = r \tag{3.7}$$

Dual Feasibility

$$\Lambda \succeq 0,\quad \Gamma \geq 0 \tag{3.8}$$

Complementary Slackness

$$\langle \Lambda, X \rangle = 0,\quad \Gamma \circ X = 0 \tag{3.9}$$

For (SDP-λ), we replace β by λ and drop the trace constraint in the primal feasibility. Since we use X_0 as the primal construction, removing one primal feasibility condition has no impact on the other part of the proof.

Consider the following primal-dual construction.

$$X_{S_k} = E_{m_k}/m_k; \quad X_{S_k S_\ell} = 0, \forall k \neq \ell \tag{3.10}$$

$$\Lambda_{S_k} = -A_{S_k} + (1_{m_k}\alpha_{S_k}^T + \alpha_{S_k}1_{m_k}^T) + \beta I_{m_k},$$

$$\Lambda_{S_k S_\ell} = -(I - \frac{E_{m_k}}{m_k})A_{S_k S_\ell}(I - \frac{E_{m_\ell}}{m_\ell}) \tag{3.11}$$

$$\Gamma_{S_k} = 0,$$

$$\Gamma_{S_k,S_\ell} = -A_{S_k,S_\ell} - \Lambda_{S_k,S_\ell} + (1_{m_k}\alpha_{S_\ell}^T + \alpha_{S_k}1_{m_\ell}^T) \tag{3.12}$$

$$\alpha_{S_k} = \frac{1}{m_k}\left(A_{S_k}\mathbf{1}_{m_k} + \phi_k \mathbf{1}_{m_k}\right) \tag{3.13}$$

$$\phi_k = -\frac{1}{2}\left(\beta + \frac{\mathbf{1}_{m_k}^T A_{S_k}\mathbf{1}_{m_k}}{m_k}\right) \tag{3.14}$$

The first order condition Eq. (3.6) is satisfied by construction. By Eq. (3.13) and (3.14), it can be seen that

$$\alpha_{S_k}^T \mathbf{1}_{m_k} = \frac{1}{m_k}\left(\mathbf{1}_{m_k}^T A_{S_k}\mathbf{1}_{m_k}\right) + \phi_k = \frac{\mathbf{1}_{m_k}^T A_{S_k}\mathbf{1}_{m_k}}{2m_k} - \frac{\beta}{2}$$

In view of the fact that both Λ and X are positive semi-definite, $\langle \Lambda, X\rangle = 0$ is equivalent to $\Lambda X = 0$. Now it remains to verify:

$$(a)\ \Lambda X = 0; \qquad (b)\ \Lambda \succeq 0; \qquad (c)\ \Gamma_{uv} \geq 0, \forall u, v$$

And it can be seen that (a) holds by construction.

Positive Semidefiniteness of Λ For (b), since $\text{span}(\mathbf{1}_{S_k}) \subset \ker(\Lambda)$, it suffices to show that for any $u \in \text{span}(\mathbf{1}_{S_k})^\perp$, $u^T\Lambda u \geq \epsilon\|u\|^2$. Consider the

decomposition $u = \sum_k u_{S_k}$, where $u_{S_k} := u \circ \mathbf{1}_{S_k}$, and $u_{S_k} \perp \mathbf{1}_{m_k}$.

$$\begin{aligned} u^T \Lambda u &= \sum_k u_{S_k}^T \Lambda_{S_k} u_{S_k} + \sum_{k \neq \ell} u_{S_k}^T \Lambda_{S_k S_\ell} u_{S_\ell} \\ &= -\sum_k u_{S_k}^T A_{S_k} u_{S_k} + \beta \sum_k u_{S_k}^T u_{S_k} - \sum_{k \neq \ell} u_{S_k}^T A_{S_k S_\ell} u_{S_\ell} \\ &= -\sum_k u_{S_k}^T (A - P)_{S_k} u_{S_k} \\ &\qquad - \sum_{k \neq \ell} u_{S_k}^T (A - P)_{S_k S_\ell} u_{S_\ell} + \beta \|u\|_2^2 \\ &= -u^T A u + \beta \|u\|_2^2 \geq \epsilon \|u\|^2 \end{aligned}$$

In order to obtain a sufficient condition on β, we will use the following lemma from Theorem 5.2 of [70], which provides a tight bound for the spectral norm $\|A - \mathbb{E}A\|$ for stochastic block models.

Lemma 3.1 ([70] Theorem 5.2)**.** *Let A be the adjacency matrix of a random graph on n nodes in which edges occur independently. Set $\mathbb{E}A = P = (p_{ij})$ and assume that $n \max_{ij} p_{ij} \leq d$ for $d \geq c_0 \log n$ and $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that $\|A - P\| \leq C\sqrt{d}$, with probability at least $1 - n^{-r}$.*

By Lemma 3.1, a sufficient condition is to have

$$\beta = \Omega(\sqrt{np_{\max}}) \geq \|A - P\|_2 \tag{3.15}$$

Positiveness of Γ For (c), denote $d_i(S_k) = \sum_{j \in S_k} A_{i,j}$, which is the number of edges from node i to cluster k, and $\bar{d}_i(S_k) = \frac{d_i(S_k)}{m_k}$. Define the average degree

between two clusters as $\bar{d}(S_k S_\ell) = \frac{\sum_{i \in S_\ell} d_i(S_k)}{m_\ell}$. For $k \neq \ell$, $u \in C_k, v \in C_\ell$, we have $\Gamma_{uv} \geq 0$ equivalent to

$$\begin{aligned} &\bar{d}_u(S_k) - \bar{d}_u(S_\ell) + \frac{1}{2}\left(\bar{d}(S_k S_\ell) - \bar{d}(S_k S_k)\right) \\ &+ \bar{d}_v(S_\ell) - \bar{d}_v(S_k) + \frac{1}{2}\left(\bar{d}(S_k S_\ell) - \bar{d}(S_\ell S_\ell)\right) \\ &- \frac{\beta}{2m_\ell} - \frac{\beta}{2m_k} \geq 0 \end{aligned} \tag{3.16}$$

By Chernoff bound and union bound, we have a sufficient condition of $\Gamma_{uv} \geq 0$ for all pairs of (u, v):

$$\delta \geq 2\sqrt{6 \log n} \max_k \sqrt{\frac{B_{kk}}{m_k}} + \max_{\ell \neq k} 6\sqrt{\frac{B_{k\ell} \log n}{m_{\min}}} + c\frac{np_{\max}}{m_{\min}}$$

A complete proof could be found in Appendix 8.1.

3.5 Experiments on Estimating Number of Clusters in Block Models

First, we present a procedure for tuning λ in (SDP-λ) in subsection 4.4.2. Then, in subsection 3.5.2 and 3.5.3 we present results on simulated and real data.

3.5.1 Tuning and substructure finding

As shown in Proposition 3.1, choice of λ should not exceed the operator norm of the observed network. Therefore we do a grid search for λ from 0 to $\|A\|_{op}$ in log scale. For each candidate λ, we solve (SDP-λ) and get the corresponding solution $\hat{X}_\lambda$. The estimated number of clusters is defined as

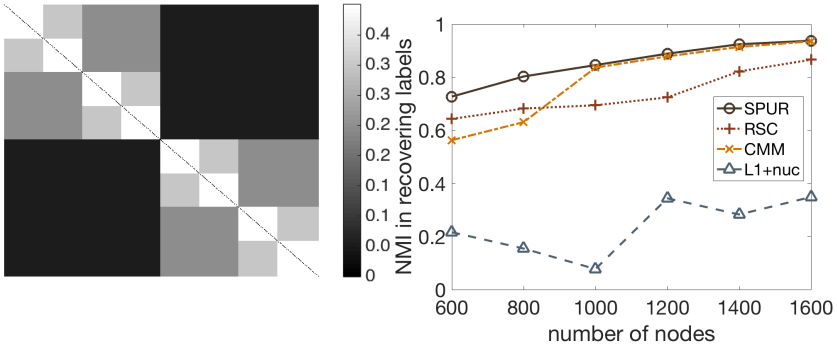
Algorithm 3 Semidefinite Program with Unknown r (SPUR)

Input: graph A, number of candidates T;
for i = 0:T-1 **do**
  $\lambda = \exp(\frac{i}{T}\log(1 + \|A\|_{op})) - 1$;
  $\hat{X}_\lambda$ = solution of (SDP-λ).
  $\theta(\lambda) = \frac{\sum_{i \le r_\lambda} \sigma_i(X_\lambda)}{\text{trace}(\hat{X}_\lambda)}$;
end for
$\hat{\lambda} = \arg\max_\lambda \theta(\lambda)$;
Output: $\hat{X}_{\hat{\lambda}}$, $\hat{r} = [\text{trace}(\hat{X}_{\hat{\lambda}})]$;

$r_\lambda = [\text{trace}(\hat{X}_\lambda)]$, where $[\cdot]$ represent the rounding operator. Let $\sigma_i(X)$ be the i-th eigenvalue of X. We then pick the solution which maximizes the proportion of leading eigenvalues $\hat{\lambda} = \arg\max_\lambda \sum_{i \le r_\lambda} \sigma_i(\hat{X}_\lambda)/\text{trace}(\hat{X}_\lambda)$. This fraction calculates the proportion of leading eigenvalues in the entire spectrum. If it equals to one, then the solution is low rank. The algorithm is summarized in Algorithm 3. In the experiments, for scalability concerns we fix a smaller range and search over the range $0.1\sqrt{\bar{d}}$ to $2\sqrt{\bar{d}}$, where $\bar{d}$ denotes the average degree.

In theory, when λ lies in the interval specified by Corollary 3.1 exact recovery is possible. Yet, in practice, solutions with different choices of λ, even outside of the theoretical range, still gives us some useful information about the sub-structures of the network. Figure 3.1 shows a probability matrix which has large separation into two big clusters and each further splits into two smaller clusters with different separations. With a larger λ it returns an under estimated r, but consistent to the hierarchical structure in the original network. In this vein, the tuning method provides a great way to do exploratory analysis

of the network.



(a) Expectation of network: (b) NMI;

Figure 3.2: The expectation matrix and NMI used for the known r setting.

3.5.2 Synthetic data

We present our simulation results in three parts - known r, increasing r and unknown r. We report the normalized mutual information (NMI) of predicted label and ground truth membership, and the accuracy of estimating r. For each experiment, the average over 10 replicates is reported.

Known number of clusters We compare the NMI of SPUR against some state-of-the-art methods, including Regularized Spectral Clustering (RSC) [5], and two convex relaxations which do not require r as input to the optimization: convexified modularity maximization (CMM) in [21]; and the ℓ_1 plus nuclear norm penalty method proposed in [23] (L1+nuc). In this setting, we use (SDP-PW) directly which does not involve any tuning. In contracst, due to the hierarchical structure of the network, the default values for the tuning

parameters in both methods would only be able to recover the lowest level of hierarchy, which consists of two clusters. Hence for a fair comparison, we try a grid search for those tuning parameters and choose the one that gives largest eigengap between the r-th and $(r+1)$-th eigenvalues of the clustering matrices. The expectation of the network generated is shown in the left panel of Figure 3.2. The right panel shows that the proposed method outperforms the competing methods.

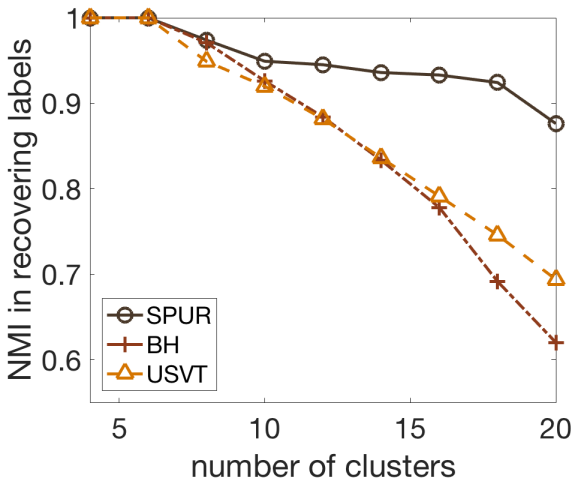
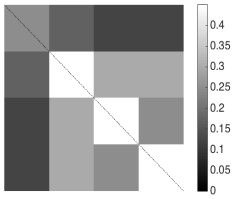
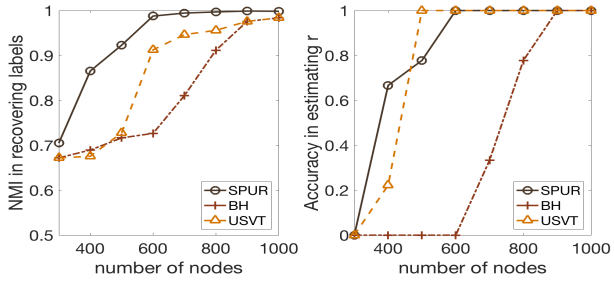


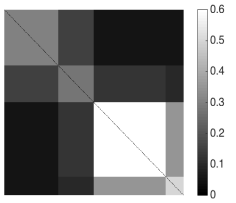
Figure 3.3: NMI under planted partition model with increasing (unknown) number of clusters.



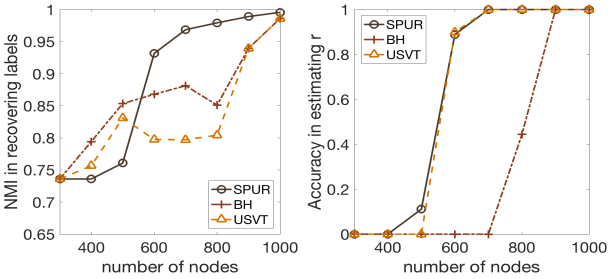
(a) balanced setting



NMI and Accuracy of r.



(b) unbalanced setting:



NMI and Accuracy of r.

Figure 3.4: The first row shows weakly assortative models with balanced cluster sizes and the corresponding NMI and accuracy in estimating r; the second row shows those for unbalanced cluster sizes.

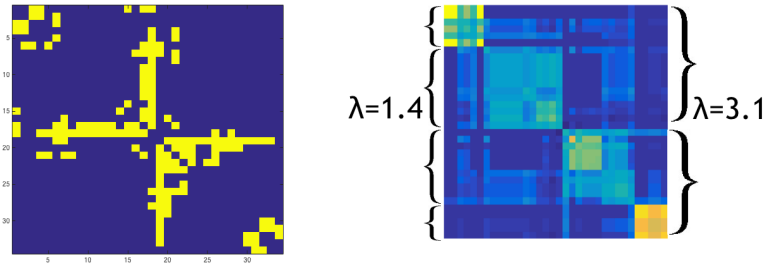


Figure 3.5: Adjacency matrix and predicted X for karate club dataset; ordered by predicted labels.

Increasing number of clusters In this experiment, we fix the number of nodes as 400 and increase the number of clusters from 4 to 20. With each given r we generate the graph with $B_{kk} = 0.6, B_{k\ell} = 0.1, \forall k \neq \ell$ and $m_{\max}/m_{\min} = 4$, then run the various estimation algorithms same as in previous experiment to estimate both r and the cluster memberships. It is shown in 3.3 that as number of clusters increases, all methods deteriorate, but the performance for SPUR declines slower than the others.

Unknown number of clusters In this experiment, we carry out two synthetic experiments for weakly assortative graphs for both balanced and unbalanced cluster sizes. We generate the network with expectation matrices shown in the leftmost column of Figure 3.4, and show the NMI of predicted labels with ground truth labels, and the fraction of returning the correct r, for both balanced (Figure 3.4-(a)) and unbalanced (Figure 3.4-(b)) settings. We run SPUR, and compare the result with 1) the Bethe-Hessian estimator (BH) in [66], in particular BHac (which has been shown to perform better for unbalanced settings), 2) USVT in [18]. For all competing methods, we run spectral clustering with the estimated r to estimate the cluster memberships. As we can see here, SPUR has a better accuracy in label recovery than competing methods. SPUR also achieves accurate cluster number faster than competing methods.

Datasets	Truth	SDP	BH	USVT	CMM
College Football	12	13	10	10	10
Political Books	3	3	4	4	2
Political Blogs	2	3	8	3	2
Dolphins	2	5	2	4	7
Karate	2	2	2	2	2

Table 3.1: Estimated number of clusters for real networks.

3.5.3 Real Datasets

We apply the proposed method on several real world data sets[1]: the college football dataset [42], the political books, political blogs [2], dolphins and karate club [120] datasets. We compare the performance of SPUR with BH, CMM and USVT in Table 4.1. As seen from [66], most algorithm correctly finds r for about 2 or 3 of these networks. It is also worth pointing out that this typically happens because different techniques finds different clusterings of the hidden substructures [10]. We will now show one such substructure we found in the Karate club data.

Figure 3.5 shows the adjacency matrix and $\hat{X}$ for the Karate club data set. For $\lambda = 3.1$, we find two clusters, whereas for $\lambda = 1.4$, we find 4 clusters, which are further subdivisions of the first level. While our tuning method picks up $\lambda = 3.1$ ($r = 2$) based on the scoring, we show the substructure for $\lambda = 1.4, r = 4$ in Figure 3.5. The left panel shows the adjacency matrix of

[1]All datasets used here are available at `http://www-personal.umich.edu/~mejn/netdata/`.

the Karate club data ordered according to the clusters obtained with $\lambda = 1.4$. The right panel of Figure 3.5 shows finer substructure of $\hat{X}$; as suggested by the adjacency matrix, within each group there are two small clique like groups at the two corners, and the hubs from each group.

In the above sections, we present SPUR, a SDP-based algorithm which provably learns the number of clusters r in a SBM under the weakly assortative setting. Our approach does not require the knowledge of model parameters, and foregoes the added tuning step used by existing SDP approaches for unequal size clusters when r is known. For unknown r, the tuning in the objective provides guidance in exploring the finer sub-structure in the network. Simulated and real data experiments show that SPUR performs comparably or better than state-of-the-art approaches.

While most dense network-based community detection schemes give perfect clustering in the limit [5, 6, 17, 24, 115], in the sparse case no algorithm is consistent; however semidefinite relaxations (among others) can achieve an error rate governed by the within and across cluster probabilities [43, 82]. In the following, we present the analysis for sparse graphs.

3.6 Community Detection for sparse networks

In this section, we discuss community detection in the sparse stochastic block model.

There are many available semidefinite programming (SDP) relaxations

for clustering blockmodels [6, 17, 24]. The common element in all of these is maximizing the inner product between A and X, for a positive semidefinite matrix X. Here X is a stand-in for the clustering matrix ZZ^T. Unequal-sized clusters is usually tackled with an extra regularization term added to the objective function (see [45, 93, 17] among others). While the above consistency results are for dense graphs, it is shown in [43, 82] that in the sparse regime one can use this method to obtain an error rate which is a constant w.r.t n and depends on the gap between the within and across cluster probabilities.

There have been several papers talking about using SDP for clustering in sparse graphs [81, 43]. The key ingredient in their analysis is the Grothendieck's inequality, which uses the sub-optimality of the ground truth matrix to turn the norm of the difference between optimal clustering matrix and the ground truth clustering matrix. Below we present a key technical lemma bounding $\|X_M - X_0\|_F$. The main goal of this lemma is to establish an upper bound on the Frobenius norm difference between the solution to an SDP with input matrix M to the ideal clustering matrix.

Lemma 3.2. *Let X_M be the solution of the following SDP for some input matrix M.*

$$\begin{aligned} \max \quad & \langle M, X \rangle, \\ s.t. \quad & X \succeq 0, 0 \leq X \leq \frac{1}{m_{\min}}, X\mathbf{1} = \mathbf{1}, trace(X) = r. \end{aligned}$$

Also let Q be a reference matrix where $Q_{ij} = \beta_k^{(in)}, \forall i, j \in C_k$, and $\beta_k^{(out)} \geq$

$Q_{ij} \geq 0, \forall i \in C_k, j \in C_\ell, k \neq \ell$. *If* $\min_k(\beta_k^{(in)} - \beta_k^{(out)}) \geq 0$, *then*

$$\|X_M - X_0\|_F^2 \leq 2\frac{\langle M - Q, X_M - X_0\rangle}{m_{\min} \min_k(\beta_k^{(in)} - \beta_k^{(out)})} \tag{3.17}$$

Remark 3.2. The key to the above lemma is to find a suitable reference matrix Q which satisfies some separation conditions between the blocks. The deviation between X_M and X_0 is small if $M - Q$ is small, and large if the separation between blocks in Q is small. While the proof technique is inspired by [43], the details are different because of our use of different constraints and because our reference matrix Q does not have to be blockwise constant and can be weakly assortative instead of strongly assortative.

The following Proposition shows the main result for SDP on sparse graphs.

Proposition 3.2. *Let* a_k, b_k *defined as in Theorem 4.1 are positive constants and* $g \geq 9$. *Then with probability tending to 1,*

$$\frac{\|X_A - X_0\|_F}{\|X_0\|_F} \leq \epsilon,$$

if $\min_k(a_k - b_k) \geq \frac{23\alpha^2 r\sqrt{g}}{\epsilon^2}$ *where* $\alpha := m_{\max}/m_{\min}$.

Note that in the above result, in order to have the error rate ϵ to go to zero, one would require $a_k - b_k$ to go to infinity, whereas by definition a_k, b_k are constants. Therefore one can only hope for a small albeit constant ϵ. In addition, in order to have a small ϵ, one needs r and α to be constants w.r.t n.

Remark 3.3 (Comparison with prior work). In contrast to having $\min_k a_k - \max_k b_k$ (strong assortativity) in the denominator like [43], we have $\min_k(a_k - b_k)$ (weak assortativity), which allows for a much broader parameter regime.

3.7 Conclusion for network community detection

In this chapter, we presented the theoretical results obtained for both dense and sparse graphs under the stochastic block model with SDP relaxations. When the number of communities is not known, we propose a new SDP framework that is able to recover the memberships with proper tuning, if the graph is relatively dense and well-separated. We have shown different proof techniques that are used in both proofs, and experimental evidences of the superior performance for SDP relaxations.

So far we have established the theoretical behavior of SDP relaxations for both networks and covariates, in the next chapter, we will derive bounds for k-means loss for sub-gaussian mixtures with low signal to noise ratio, and combine the techniques used in these problems to investigate the inference on graphs with node covariates.

Chapter 4

Networks with Covariates

In this chapter, we investigate community detection in networks in the presence of node covariates. In many instances, covariates and networks individually only give a partial view of the cluster structure. One needs to jointly infer the full cluster structure by considering both. In Statistics, an emerging body of work has been focused on combining information from both the edges in the network and the node covariates to infer community memberships. However, so far the theoretical guarantees have been established in the dense regime, where the network can lead to perfect clustering under a broad parameter regime, and hence the role of covariates is often not clear. In this chapter, we examine sparse networks in conjunction with finite dimensional sub-gaussian mixtures as covariates under moderate separation conditions. In this setting each individual source can only cluster a non-vanishing fraction of nodes correctly. We propose a simple optimization framework which provably improves clustering accuracy when the two sources carry partial information

The content in this chapter was conducted in collaboration with Purnamrita Sarkar, which is now available on arXiv (Yan, Bowei, and Purnamrita Sarkar. "Convex Relaxation for Community Detection with Covariates." arXiv preprint arXiv:1607.02675 (2016).). Purnamrita Sarkar proposed the initial idea and all technical proofs were shown jointly by both authors. The experimental part was mainly performed by Bowei Yan, and the writing was mainly by Purnamrita Sarkar.

about the cluster memberships, and hence perform poorly on their own. Our experiments show that combining the two sources requires weaker separation conditions for each individual source. Our optimization problem can be solved using scalable convex optimization algorithms. Using a variety of simulated and real data examples, we show that the proposed method outperforms other existing methodology.

4.1 Background

Although most real world network datasets come with covariate information associated with nodes, existing approaches are primarily focused on using the network for inferring the hidden community memberships or labels. Take for example the Mexican political elites network (described in detail in Section 4.4). This dataset comprises of 35 politicians (military or civilian) and their connections. The associated covariate for each politician is the year when one came into power. After the military coup in 1913, the political arena was dominated by the military. In 1946, the first civilian president since the coup was elected. Hence those who came into power later are more likely to be civilians. Politicians who have similar number of connections to the military and civilian groups are hard to classify from the network alone. Here the temporal covariate is crucial in resolving which group they belong to. On the other hand, politicians who came into power around 1940's, are ambiguous to classify using covariates. Hence the number of connections to the two groups in the network helps in classifying these nodes. Our method can successfully

classify these politicians and has higher classification accuracy than existing methods [12, 123].

In Statistics literature, there has been some interesting work on combining covariates and dense networks (average degree growing faster than logarithm of the number of nodes). In [12], the authors present assortative covariate-assisted spectral clustering (ACASC) where one does Spectral Clustering on the the gram matrix of the covariates plus the regularized graph Laplacian weighted by a tuning parameter. A joint criterion for community detection (JCDC) with covariates is proposed by [123], which could be seen as a covariate reweighted Newman-Girvan modularity. This approach enables learning different influence on each covariate. In concurrent work [108] provide a variational approach for community detection.

All of the above works are carried out in the dense regime with strong separability conditions on the linkage probabilities. ACASC also requires the number of dimensions of covariates to grow with the number of nodes for establishing consistency.

In contrast to the above, we prove our result for sparse graphs where the average degree is constant and the the covariates are finite dimensional sub-gaussian mixtures with moderate separability conditions. In our setting, neither source can yield consistent clustering in the limit. We show that combining the two sources leads to improved clustering accuracy under weaker conditions on separability on each individual source.

Widely known as multi-view clustering, leveraging information from multiple sources have been long studied in Machine learning and Data mining. In [62], the authors use a regularization framework so that the clustering adheres to the dissimilarity of clustering from each view. In [71], the authors optimize the nonnegative matrix factorization loss function on each view, plus a regularization forcing the factors from each view to be close to each other. The only provable method is by [19], where the authors obtain guarantees where the two views are mixtures of Log-concave distributions. This algorithm does not apply to networks.

In this chapter, we propose a penalized optimization framework for community detection when node covariates are present. We take the sparse degree regime of Stochastic Blockmodels, where one can only correctly cluster a non-vanishing fraction of nodes. Similarly, for covariates, we assume that the covariates are generated from a finite dimensional sub-gaussian mixture with moderate separability conditions. We prove that our method leads to an improved clustering accuracy under weaker conditions on the separation between clusters from each source. As byproducts of our theoretical analysis we obtain new asymptotic results for sparse networks under weak separability conditions and kernel clustering of finite dimensional mixture of sub-gaussians.

Using a variety of real world and simulated data examples, we show that our method has improved performance over existing methods. We also illustrate in the simulation that if the two sources only have partial and in some sense orthogonal information about the clusterings, then combining them leads

to better clustering than using the individual sources.

In Section 4.2, we introduce relevant notation and present our optimization framework. In Section 4.3, we present our main results, followed by experimental results on simulations and real world networks in Section 4.4. Majority of the proofs are presented in Appendix 9.

4.2 Problem Setup

In this section, we introduce our model and set up the convex relaxation framework. For the covariates, we define,

$$\text{(Covariate Model)} \qquad Y_i = \sum_{a=1}^{r} Z_{ia}\mu_a + W_i \tag{4.1}$$

W_i are mean zero d dimensional sub-gaussian vectors with spherical covariance matrices $\sigma_k^2 I_d$ and sub-gaussian norm ψ_k (for $i \in C_k$). Compare with Eq. (2.12) the key difference is that the noise does not scale with the square root of the dimension, which makes the signal to noise ratio lower for high-dimensional problems. We define the distance between clusters C_k and C_ℓ as $d_{k\ell} = \|\mu_k - \mu_\ell\|$ and the separation as $d_{\min} = \min_{k\neq\ell} d_{k\ell}$.

4.2.1 Optimization Framework

We now present our optimization framework. We have talked about many SDP relaxations for networks in Chapter 3. Yet

We analyze the widely-used Gaussian kernel defined in Eq. (2.12) to allow for non-linear boundaries between clusters. This kernel function is upper

bounded by 1 and is Lipschitz continuous w.r.t. the distance between two observations. Same as in Chapter 3, we use X as a stand in for the normalized variant of the clustering matrix ZZ^T, i.e. the desired solution X_0 is as defined in Eq. (3.2). It can be seen that $\|X_0\|_F^2 = r$.

We have already shown in Chapter 2 that SDP can be used as a convex relaxations of the k-means loss. In our optimization framework, we propose to add a k-means type regularization term to the network objective, which enforces that the estimated clusters are consistent with the latent memberships in the covariate space.

$$X = \arg\max_X \langle A + \lambda K, X \rangle \quad s.t. \quad X \in \mathcal{F}, \tag{4.2}$$

where λ is a tuning parameter and the constraint set $\mathcal{F} = \{X \succeq 0, \quad 0 \leq X \leq \frac{1}{m_{\min}}, \quad X\mathbf{1}_n = \mathbf{1}_n, \quad \text{trace}(X) = r\}$ is similar to [92]. Compared with Eq. (SDP-PW), the only different is that the element-wise upper bound of X. The $m_{\min}$ in the constraint can be replaced by any lower bound on the smallest cluster size, and is mainly of convenience for the analysis. In the implementation, it suffices to enforce the element-wise positivity constraints, and other linear constraints. For ease of exposition, we define

$$X_M = \arg\max_X \langle M, X \rangle \quad s.t. \quad X \in \mathcal{F}, \tag{4.3}$$

When $K(i,j) = Y_i^T Y_j$, then the non-convex variant of the objective function naturally assumes a form similar to the work of ACASC (modulo normalization of A).

4.3 Main Results

Typically in existing SDP literature for sparse networks or sub-gaussian mixtures [43, 79], one obtains a relative error bound of the deviation of X_M (the solution of the SDP) from the ideal clustering matrix X_0. This relative error is typically proportional to the ratio of the observed matrix with a suitably defined reference matrix, and some quantity which measures the separation between the different clusters. Our theoretical result shows that the relative error of the solution to the combined SDP is proportional to the ratio of the observed $A + \lambda K$ matrix to a suitably defined reference matrix to a quantity which measures separation between clusters. This quantity is a non-linear combination of the separations stemming from the two sources. We first present an informal version of the main result. *Main theorem (informal): Let $X_{A+\lambda K}$ be the solution of SDP* (4.3). *Let s_G^k and s_C^k be constants denoting the separations of cluster k from the other clusters defined in terms of the model parameters of the network and the covariates respectively. Then*

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq \frac{c_G + \ell c_C}{\min_k \left(s_G^k + \ell s_C^k\right)},$$

where c_G and c_C are constants representing the error corresponding to the graph and the covariates, and ℓ is a tuning parameter.

Note that in SBM, the separation is well-defined, i.e. when $M = A$, a natural choice of the reference matrix is $\mathbb{E}[A|Z]$ which is blockwise constant. In this case, the separation is given by $\min_k(B_{kk} - \max_\ell B_{k\ell})$, and leads to a result on weakly assortative sparse block models which we present in more

details in Section 3.6. However, for the kernel matrix K, the main difficulty is that one cannot achieve element-wise or operator norm concentration of K (also discussed in [106]). This makes the choice of the reference matrix difficult.

The results on networks, covariates and the combination of the two essentially reduces to identifying good reference matrices (Q) for the input matrices A, K, and $A + \lambda K$, which

1. Satisfies the properties of Q in the above lemma.

2. Has a large separation $\min_k(\beta_k^{(in)} - \beta_k^{(out)})$ increasing the denominator of Eq. (3.17).

3. Has a small deviation from M, thereby reducing the numerator of Eq (3.17).

Now the main work is to choose the reference matrix Q for $A + \lambda K$. As pointed out before, a common choice for reference matrix of A is $\mathbb{E}[A|Z]$. For the covariates, we divide the nodes into "good" nodes $\mathcal{S}_k := \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}$ and the rest. Also define $\mathcal{S} = \cup_{k=1}^r \mathcal{S}_k$. Δ_k will be defined such that the kernel matrix induced by the rows and columns in $\mathcal{S}$ is weakly assortative, and $3\Delta_k + \Delta_\ell \leq d_{k\ell}$. Define

$$r_k := f(2\Delta_k), \quad s_k := \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell), \qquad \nu_k = r_k - s_k \tag{4.4}$$

A simple use of triangle inequality gives $\min_{i,j \in \mathcal{S}_k} K_{ij} \geq r_k$ and $\max_{i \in \mathcal{S}_k, j \in S_\ell, \ell \neq k} K_{ij} \leq s_k$. Hence the separation for cluster k is $\nu_k := r_k - s_k$. We define the reference

matrix K_I as:

$$(K_I)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i,j \in C_k \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \neq \ell \end{cases} \tag{4.5}$$

The choice of Δ_k is crucial. A large Δ_k makes the size of non-separable nodes $\mathcal{S}^c$ small, but drives down the separation ν_k.

We are now ready to present our main result. As we will show in the proof, the new separation is $\gamma = \min_k((p_k - q_k) + \lambda\nu_k)$. Typically, in the general case with unequal sub-gaussian norms, one should benefit from using different Δ_k's for different clusters. For example for a cluster with a large $p_k - q_k$, we can afford to have a small ν_k. To think in terms of Δ_k, for this cluster one can have a large Δ_k, which will make $|\mathcal{S}_k|$ larger than before, but will not affect the separation $(p_k - q_k) + \lambda\nu_k$ of cluster k very detrimentally. We now present our first main theorem.

Theorem 4.1. *Let* $a_k = nB_{kk}, b_k = n\max_{\ell\neq k} B_{k\ell}$, $g := \frac{2}{(n-1)}\sum_{i<j} Var(a_{ij}) \geq 9$. *Take* $\lambda = \ell/n$, $m_k = n\pi_k$, $m_{\min} = n\pi_{\min}$, *and* $\pi_0 := \sum_k (m_k \exp(-\Delta_k^2/5\psi_k^2) + \sqrt{m_k \log m_k/2})/n$. *Let* $X_{A+\lambda K}$ *be defined as in Eq* (4.3). *If* $\pi_{\min} = \Theta(1)$ *and* $\min_k(a_k - b_k + \ell\nu_k) > 0$, *then, with probability tending to one,*

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq 2K_G \frac{6\sqrt{g} + \ell\,(2\pi_0 + \sum_k \pi_k^2(1 - f(2\Delta_k)))}{\pi_{\min}^2 \min_k(a_k - b_k + \ell\nu_k)},$$

where $\nu_k = f(2\Delta_k) - \max_{\ell\neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$ *for some* $\Delta_k, \Delta_\ell \geq 0$ *and* $\max(\Delta_k, \Delta_\ell) \leq d_{k\ell}/4$.

Here K_G is the Grothendieck's constant. The best value of K_G is still unknown, and the best known bound is $K_G \leq 1.783$ [16]. First note that in the

sparse case, we take $\lambda = \ell/n$ for some constant ℓ. In general the upper bound depends on several parameters such as λ and the scale parameter η in the gaussian kernel. We provide procedures for tuning λ and η in Section 4.4. The Δ_k's show up in the numerator as well as the denominator. Finding the optimal Δ_k is cumbersome in the general case with unequal ψ_k's. In Section 4.3.1 we derive an upper bound for equal Δ_k's for concreteness.

Remark 4.1. Ideally one would want to show that the upper bound obtained in the above theorem is smaller than the Bayes Error rate for clustering with either source alone. However, for clustering in Blockmodels finding a polynomial time algorithm which achieves the Bayes error rate is still an open problem.

Now we present a natural byproduct of our analysis, namely the result on covariate clustering i.e. bounds on $\|X_0 - X_K\|_F$.

4.3.1 Result on Covariates

We present a result for covariates analogous to the sparse graph setting, which establishes that, while SDP with covariates is not consistent with finite signal-to-noise ratio, it achieves a small error rate if the cluster centers are further apart. But before delving into our analysis, we provide a brief overview of existing work.

For covariate clustering, it is common to make distributional assumptions; usually a mixture model with well-separated centers suffices to show consistency. The most well-studied model is Gaussian mixture models, which

can be inferred by Expectation-Maximization algorithm [116] and its variant [29]. The condition required for provable recovery on the separation is usually the minimum distance between clusters is greater than some multiple of the square root of dimension (or effective dimension).

Another popular technique is based on SDP relaxations. For example, it is proposed in [92, 79] a SDP relaxation for k-means type clustering. To make the analysis concrete, for Proposition 4.1, we use $\Delta_k = \Delta$.

Proposition 4.1 (Analysis for Covariates)**.** *Let K be the kernel matrix generated from kernel function f. Denote ν_k as in Eq* (4.4)*. If $\frac{d_{\min}}{\psi_{\max}} > \max\left\{\sqrt{d}, \frac{180}{\sqrt{d}}\right\}$, then with properly chosen η, with probability at least $1 - \sum_k \frac{1}{m_k}$,*

$$\frac{\|X_K - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 d \frac{\psi_{\max}^2}{d_{\min}^2} \max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right), r\right\}$$

Remark 4.2 (Comparison with prior work). In recent work [79], it is shown the effectiveness of SDP relaxation with k-means clustering for sub-gaussian mixtures, provided the minimum distance between centers is greater than the standard deviation of the sub-gaussian times the number of clusters r. We provide a dimensionality reduction scheme, which also shows that the separation condition requires that $d_{\min} = \Omega(\sqrt{\min(r, d)})$. Our proof technique is new and involves carefully constructing a reference matrix for Lemma 3.2.

Compare with EM algorithm, it is worth pointing out that SDP recovers the membership and by de-noising the data using the SDP solution matrix [79]

could give an estimate of the cluster centers, but it is not asymptotically consistent. EM, on the other hand, can give estimates that converge to the global optimum under mild initialization conditions for isotropic Gaussian mixture models.

4.3.2 Analysis of Covariate Clustering when $d \gg r$

In high dimensional statistical problems, the signal is often assumed to lie in a low dimensional subspace or manifold. This is why much of Gaussian Mixture modeling literature first computes some projection of the data onto a low dimensional subspace [103]. To reduce the dimensionality of the raw data, one could do a feature selection for the covariates (e.g. [55, 105]). In contrast, here we propose a much simpler dimensionality reduction step, which does not distort the pairwise distances between cluster means too much. The intuition is that, for clustering a subgaussian mixture, if $d \gg r$, the effective dimensionality of the data is r since the cluster means lie in an at most r-dimensional subspace.

Hence we propose the following simple dimensionality reduction algorithm when $d \gg r$ in a spirit similar to [19]. We show the effect of dimensionality reduction on the pairwise distance matrix in the Supplementary material. [ADD]

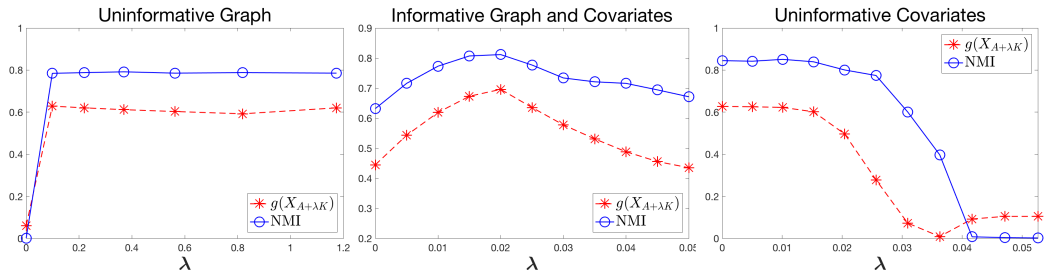
We split up the sample into two random subsets P_1 and P_2 of sizes n_1 and $n - n_1$ and compute the top $r - 1$ eigenvectors U_{r-1} of the matrix $\hat{S} = \frac{\sum_{i \in P_1} (Y_i - \bar{Y})(Y_i - \bar{Y})^T}{n_1} \in \mathbb{R}^{d \times d}$, where $\bar{Y} = \frac{\sum_{i \in P_1} Y_i}{n_1}$. Now we project the

covariates from subset P_2 onto this lower dimensional subspace as $Y_i' = U_{r-1}^T Y_i$ to get the low dimensional projections. We take $n_1 = n/\log n$.

Lemma 4.1. *Let $M := \sum_k \pi_k \mu_k \mu_k^T$. If $\sum_k \pi_k \mu_k = 0$, and $\lambda_{r-1}(M) \geq 5\psi_{\max}^2 + C\sqrt{\frac{d \log^2 n}{n}}$ for some constant C, the projected Y_i' are also independent data points generated from an isotropic sub-gaussian mixture in $r-1$ dimensions. Furthermore the minimum distance between the means in the $r-1$ dimensional space is at least $d_{\min}/2$ with probability at least $1 - \tilde{O}(r^2 n^{-d})$, where $d_{\min}$ is the separation in the original space.*

The proof of this lemma is deferred to Appendix. We believe the proof can be generalized to non-spherical cases as long as the largest eigenvalue of covariance matrix for each cluster is bounded. Typically λ signifies the amount of signal. For example, for the simple case of mixture of two gaussians with $\pi_1 = 1/2$, and $\mu_2 = -\mu_1$, $\lambda = \|\mu_1\|^2$, which is essentially $d_{\min}^2/4$. Hence the condition on λ essentially translates to a lower bound on the signal to noise ratio, i.e. $d_{\min}^2 \geq 48\psi_{\max}^2 + C'\sqrt{\frac{d \log^2 n}{n}}$ for some constant C'. When $d > r$, if one applies Lemma 4.1 on the $r-1$ dimensional space, then as long as $d_{\min}^2 = \Omega(\psi_{\max}^2 r)$, the separation in the low dimensional space also satisfies the separation condition in Proposition 4.1. Thus the dimensionality reduction brings down the separation condition in Proposition 4.1 from $\Omega(\psi_{\max}\sqrt{d})$ to $\Omega(\psi_{\max}\sqrt{\min(r,d)})$.

The sample splitting is merely for theoretical convenience which ensures that the projection matrix and the projected data are independent, resulting



(a) (b) (c)

Figure 4.1: Tuning: (a) $B = 0.005E_3, n = 1000$, $d = 6, d_{\min} = 15\sigma$; (b) $d = 6, d_{\min} = 1.3, \sigma = (1, 1, 5), B = \text{diag}(0.004, 0.024, 0.024) + 0.004E_3$; (c) $d = 6, d_{\min} = 0, B = 0.0144I_3 + 0.0016E_3$.

in the fact that the final projection is also an independent sample from a sub-gaussian mixture. To be concrete, the labels of P_1 do not matter asymptotically, since they incur a relative error in $\|X_0 - X_K\|_F/\|X_0\|_F$ less than $\sqrt{n^2/(m_{\min}^2 \log n)}/\sqrt{r} \leq \sqrt{\alpha^2 r/\log n}$, where α and r are both constants. In our setting, the relative error in Proposition 4.1 is a small but non-vanishing constant, and so this additional vanishing error term does not affect it. However this sample splitting step is not necessary in practice [19], and so we do not pursue this further.

We now present the tuning procedure, and experimental results.

4.4 Experiments

In this section, we present results on real and simulated data. The cluster labels in our method are obtained by spectral clustering of the solution matrix returned by the SDP. We will use SDP-comb, SDP-net, SDP-cov to represent the labels estimated from $X_{A+\lambda K}$, X_A and X_K respectively. Perfor-

mance of the clustering is measured by normalized mutual information (NMI), which is defined as the mutual information of the two distributions divided by square root of the product of their entropies. We have also calculated classification accuracy and they show similar trends, so only NMI is reported in this section. For real and simulated data, we compare: (1) Covariate-assisted spectral clustering (ACASC) [12]; (2) JCDC [123], (3) SDP-comb, (4) SDP-net and (5) SDP-cov. The last two are used as references of graph-only and covariate-only clustering respectively.

4.4.1 Implementation and computational cost

Solving semidefinite programming with linear and non-linear constraints has been a challenging problems in numerical optimization community. Many SDPs proposed in statistical literature [17, 24, 6] are solved by the alternating descent method of multipliers (ADMM) algorithm [15]. Although ADMM is tractable for middle-sized problems and reasonable numerical behavior, whether it convergences in presence of non-negative constraints, which is prevalent in network literatures, remains an open problem. Recently, the authors of [117] propose a majorized semismooth Newton-CG augmented Lagrangian method, called SDPNAL+, which is provably convergent. We solve the SDP using the matlab package of SDPNAL+ in all our experiments[1]. The package provides an efficient implementation of the algorithm. Solving the SDP for

[1]The code used for the experiment can be found at `https://github.com/boweiYan/SDP_SBM_unbalanced_size`.

matrix of size 1000×1000 takes less than a minute on a Macbook with a 1.1 GHz Intel Core M processor.

4.4.2 Choice of Tuning Parameters

As we pointed out earlier, the elementwise upper bound $\frac{1}{m_{\min}}$ is only for convenience of theoretical analysis. In the implementation, we do not enforce this constraint. So the main tuning parameters would be the scale parameter in the kernel matrix η and the tradeoff parameter between graph and covariates λ. In most of our experiments the number of clusters is assumed known. In this section, we also provide a practical way to choose among candidates of r when it is not given.

Choice of η We use the method proposed in [100] to select the scale parameter. The intuition is to keep enough (say 10%) of the data points in the "range" of the kernel for most (say 95%) data points. Given the covariates, we first compute the pairwise distance matrix. Then for each data point Y_i, compute q_i as 10% quantile of $d(Y_i, Y_j), \forall j \in [n]$. The bandwidth is defined as

$$w = \frac{95\% \text{ quantile of } q_i}{\sqrt{95\% \text{ quantile of } \chi^2_d}}$$

and scale parameter $\eta = \frac{1}{2w^2}$.

Note when the data is high-dimensional, we will first conduct dimensionality reduction as in Section 4.3.2, then use the intrinsic dimension to tune the scale parameter.

Choice of λ As λ increases, the resulting $X_{A+\lambda K}$ clustering gradually changes from X_A clustering to X_K clustering. Our theoretical results show that, with the right λ, $X_{A+\lambda K}$ and X_0 should be close, and hence also have similar eigenvalues. Define the eigen gap function for clustering matrices $g(X) := (\lambda_r(X) - \lambda_{r+1}(X))/\lambda_r(X)$. Using Weyl's inequality and the fact that $\|X_{A+\lambda K} - X_0\|_{\text{op}} \leq \|X_{A+\lambda K} - X_0\|_F$, we have: $\lambda_r(X_0) - \|X_{A+\lambda K} - X_0\|_F \leq \lambda_r(X_{A+\lambda K}) \leq \lambda_r(X_0) + \|X_{A+\lambda K} - X_0\|_F$. Since $g(X_0) = 1$, we pick the λ maximizing $g(X_{A+\lambda K})$. In Figure 4.1 (a)-(c), figures from left to right represent the situation where graph is uninformative (Erdős-Rényi), both are informative and covariates are uninformative. We plot $g(X_{A+\lambda K})$ and NMI of the clustering from $X_{A+\lambda K}$ with the true labels against λ. Figure 4.1 shows that $g(X_{A+\lambda K})$ and NMI of the predicted clustering have a similar trend, justifying the effectiveness of the tuning procedure.

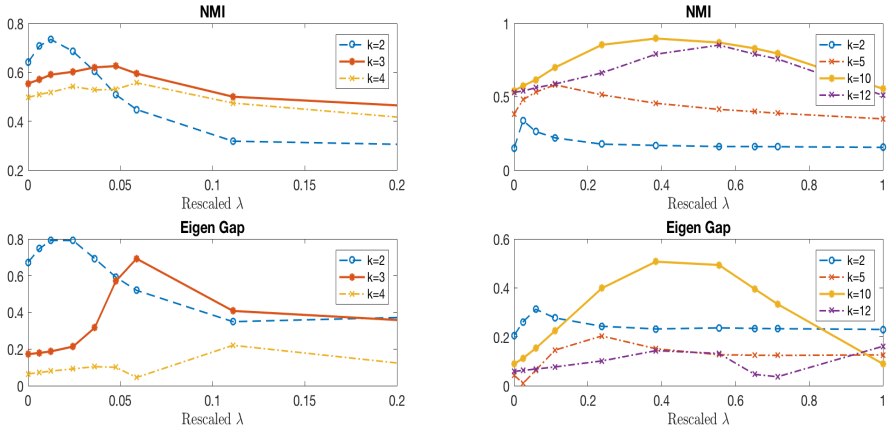
Unknown number of clusters In many real world settings, it is generally hard to possess the knowledge of number of clusters. Methods are proposed for selecting number of blocks under sparse stochastic block models [66], but most of these methods are designed specific for graph adjacency matrix and cannot be generalized to continuous matrix scenarios. We observe that the eigen gap acts as an informative indicator for picking the number of clusters. So when the number of clusters is unknown, we run the SDP over a grid of λ, k, and choose the pair that maximizes the eigen gap. As we show in Figure 4.2, we construct two settings and test the performance of using eigen

gap to select r. In the first setting, the true model has 3 clusterings with proportion $3:4:5$, the probability matrix is $B = 0.01 * \begin{bmatrix} 1.6 & 1.2 & 0.16 \\ 1.2 & 1.6 & 0.02 \\ 0.16 & 0.02 & 1.2 \end{bmatrix}$. And the covariates are high dimensional gaussian centered at $\mu_1 = (0, 2, 0 \cdots, 0)$, $\mu_2 = (-1, -0.8, 0 \cdots, 0)$, $\mu_3 = (1, -0.8, 0 \cdots, 0)$. We sample $n = 800$ data points, and run SDP on top of it with different choice of λ and specified number of clusters k. For each pair of parameter, we compute the NMI and eigengap and plot them on the upper and lower panel of Figure 4.2-(a). As we can see, the eigen gap presents a similar trend as the NMI, hence picking the pair that optimizes eigen gap will have a relatively high NMI as well. Note here the mis-specified $k = 2$ has a higher NMI than that of the true value of r. This tells us even the number of clusters is mis-specified, the SDP is still able to find structure that correlates with the underlying model. This phenomenon is also observed in several other works [115, 93].

In the second scenario, we generate a planted partition model with 10 equal-sized clusters, where $B = 0.046 I_{10} + 0.004 E_{10}$, along with Gaussian covariates centered at $[3 * I_{10} \mid \mathbf{0}_{3,90}]$. We conduct the same type of experiment as above and plot the NMI and eigengap. In this case, the eigen gap succussfully recovered the true number of clusters.

4.4.3 Simulation Studies

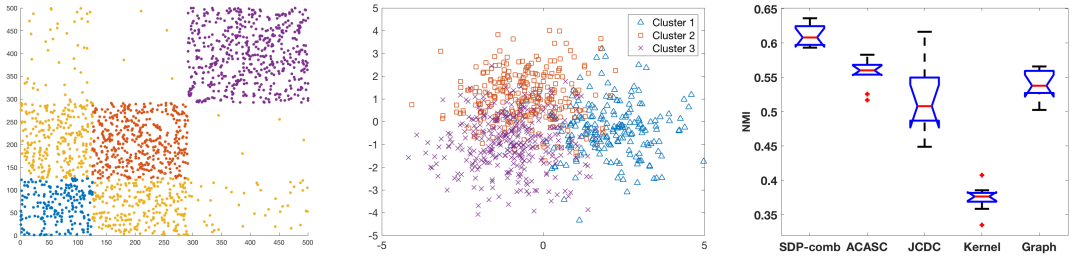
In this part we consider two simulation settings. In the first setting, we generate three clusters with sizes 3:4:5, with $n = 800$. The probability matrix is $B = 0.01 * \begin{bmatrix} 1.6 & 1.2 & 0.16 \\ 1.2 & 1.6 & 0.02 \\ 0.16 & 0.02 & 1.2 \end{bmatrix}$, and the covariates for each cluster are generated



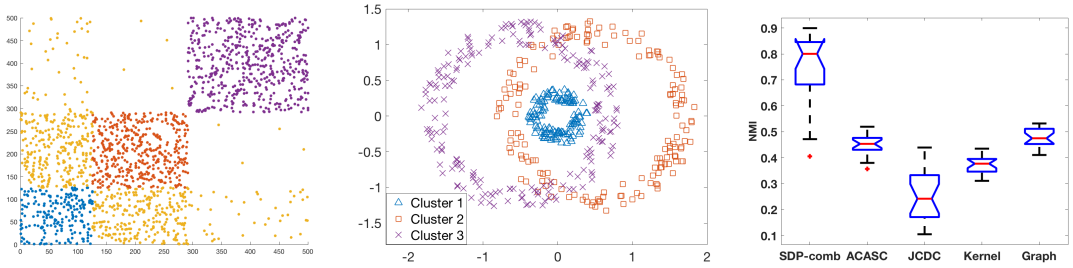
(a) true $r = 3$ (b) true $r = 10$.

Figure 4.2: NMI and eigen gap for various choice of r.

with 100 dimensional unit variance isotropic Gaussians, whose centers are only non-zero on the first two dimensions with $\mu_1 = (0, 2, 0 \cdots, 0)$, $\mu_2 = (-1, -0.8, 0 \cdots, 0)$, $\mu_3 = (1, -0.8, 0 \cdots, 0)$. This is the same setting as in the first simulation for unknown r. In this example, the network cannot separate out clusters one and two well, whereas the covariates can. On the other hand, clusters two and three are not well separated in the covariate space, while they are well separated using the network parameters. The experiments are repeated on 10 independently generated samples and the box plot for NMI is shown as in Figure 4.3(c). In the second row of Figure 4.3, we examine covariates with nonlinear cluster boundaries. The graph used here is the same as above, and the covariates are 2-dimensional, whose scatter plot is shown in Figure 4.3(e). In this case, the kernel matrix is able to pick up local similarities hence performs better than combination via inner product similarity as used



Simulation 1: (a) Graph (b) Covariates - 1 (c) NMI - 1



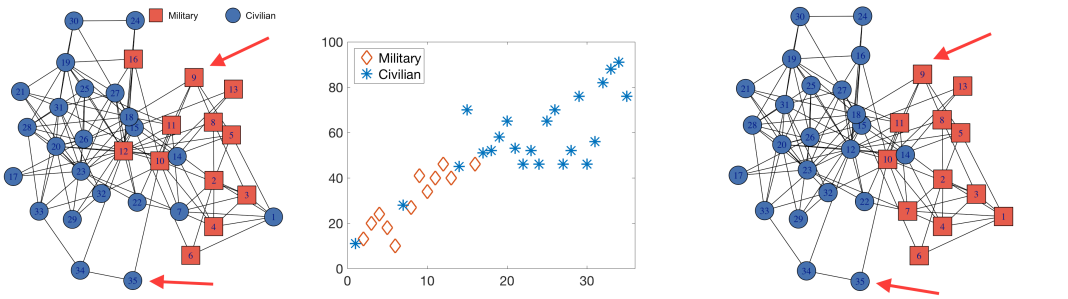
Simulation 2: (d) Graph (e) Covariates - 2 (f) NMI - 2

Figure 4.3: The first and second rows have results for isotropic Gaussian covariates and covariates lies on a nonlinear manifold respectively. We plot the adjacency matrix A in (a) and (b), where blue, red and purple points represent within cluster edges for 3 ground truth clusters respectively and yellow points represent inter-cluster edges. In (b) and (e) we plot covariates ; different shapes and colors imply different clusters. (c) and (f) show the box plots for NMI.

in ACASC. In both simulations, SDP-comb outperforms others.

4.4.4 Real World Networks

Now we present results on a real world social network and an ecological network. The performance of clustering is evaluated by NMI with the ground truth labels.



(a) Ground truth (b) Node feature (c) Predicted by SDP-comb

Figure 4.4: Mexican political network.

Mexican political elites As discussed before, this network [41] depicts the political, kinship, or business interactions between 35 Mexican presidents and close collaborators, etc. The two ground truth clusters consist of the military and the civilians, indicating the background of the politician. The year in which a politician first held a significant governmental position, is used as a covariate. Figure 4.4(b) shows that the covariate gives a good indication of the labels. This is because the military dominated the political arena after the revolution in the beginning of the twentieth century, and were succeeded by the civilians.

Table 4.1 shows the NMI of all methods, where our method outperforms other covariate-assisted approaches. From Figure 4.4(a, c), for example, node 35 has exactly one connection to each of the military and civilian groups, but seized power in the 90s, which strongly indicates a civilian background. On the other hand, node 9 took power in 1940, a year when civilian and military had almost equal presence in politics, making it hard to detect node 9's political affiliation. However, this node has more edges to the military group than

the civilian group. By taking the graph structure into consideration, we can correctly assign the military label to it.

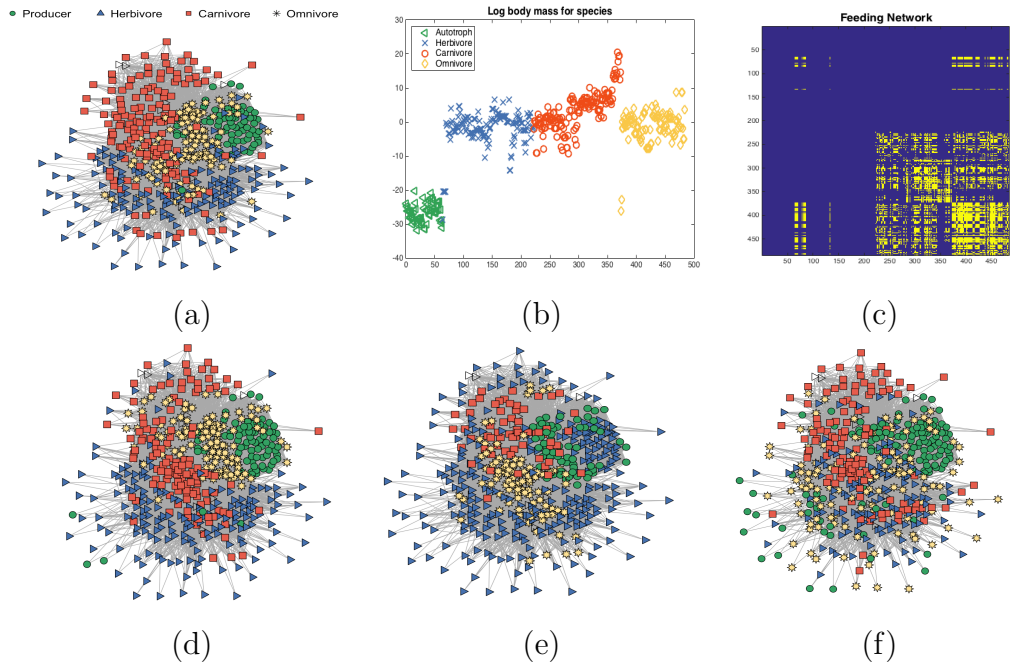


Figure 4.5: Weddell sea network: (a) True labels; (b) Log body mass; (c) Constructed adjacency matrix A_τ; we show labels from (d) SDP-comb; (e) SDP-net; (f) SDP-cov.

Weddell sea trophic dataset The next example we consider is an ecological network collected by [53] describing the marine ecosystem of Weddell Sea, a large bay off the coast of Antarctica. The dataset lists 489 marine species and their directed predator-prey interactions, as well as the average adult body mass for each of the species. We use a thresholded symmetrization of the directed graph as the adjacency matrix. Let G be the directed graph, the $(i, j)^{th}$

Dataset	SDP-net	SDP-cov	SDP-comb	ACASC	JCDC
Mexican politicians	0.37	0.43	**0.46**	0.37	0.25
Weddell Sea	0.36	0.22	**0.50**	0.32	0.42

Table 4.1: NMI with ground truth for various methods

entry of GG^T captures the number of other species which i and j both feed on. We create binary matrices $A_\tau = 1(GG^T \geq \tau)$. Choosing different τ's between 1 to 10 gives similar clustering. We use $\tau = 5$.

All species are labeled into four categories based on their prey types. Autotrophs (e.g. plants) do not feed on anything. Herbivores feed on autotrophs. Carnivores feed on animals that are not autotrophs, and the remaining are omnivores, which feed both on autotrophs and other animals (herbivore, carnivore, or omnivores). Since body masses of species vary largely from nanograms to tons, we work with the normalized logarithm of mass following the convention in [85]. Figure 4.5(b) illustrates the log body mass for species. Without loss of generality, we order the nodes as autotrophs, herbivores, carnivores and omnivores.

In Figures 4.5(c), we plot A_τ. Since the autotrophs do not feed on other species in this dataset, and since herbivores do not have too much overlap in the autotrophs they feed on, the upper left corner of the input network is extremely sparse. On the other side, the body sizes for autotrophs are much smaller than those of other prey types. Therefore the kernel matrix clearly separates them out.

We see that SDP-net (Figure 4.5(e)) heavily misclusters the autotrophs

since it only replies on the network. SDP-net (Figure 4.5(f)) only takes the covariates into account and cannot distinguish herbivores from omnivores, since they possess similar body masses. However, SDP-comb (Figure 4.5(d)) achieves a significantly better NMI by combining both sources. Table 4.1 shows the NMI between predicted labels and the ground truth from SDP-comb, JCDC and ACASC. While JCDC and ACASC can only get as good as the the best of graph or covariates, our method achieves a higher NMI.

4.5 Discussion

In this paper, we propose a regularized convex optimization framework to infer community memberships jointly from sparse networks and finite dimensional covariates. We theoretically show that our framework can improve clustering accuracy of either source under weaker separation conditions. In particular, when each source only has partial information about the clustering, our methodology can lead to high clustering accuracy, when either source fails. We demonstrate the performance of our methodology on simulated and real networks, and show that it in general performs better than other state-of-the-art methods. While for ease of exposition we limit ourselves to two sources, our method can be easily generalized to multiple views or sources. Empirically, we demonstrate that our method works for covariates with non-linear cluster boundaries; we intend to extend our theoretical analysis to this setting and non-isotropic covariates as well.

Chapter 5

Conclusion and Open Problems

In this thesis, I have summarized my work on theoretical analysis for several convex and non-convex optimization problems, EM algorithm for Gaussian mixture models and SDP relaxation for sub-gaussian mixture models, stochastic block models and network with node covariates. We have proved the theoretical upper bounds and exact recovery results for sparse and dense SBM respectively, and have shown the effectiveness of the proposed SDP with experimental evidences.

For future directions, there are still many open problems in the area. In covariate clustering, it is very strong to assume one knows the number of latent clusters beforehand. When the number of clusters is unknown, few methods could find that with provable guarantees, even for isotropic Gaussian mixture models [102].

There are many other non-convex algorithms that uses alternating minimization to find the distribution of latent variables in clustering problems. For example, the Latent Dirichlet Allocation, which is widely used in topic modeling for natural language processing, and the mean field variational inference for stochastic block models, are both methods that optimize a non-convex loss

function. The behavior of these methods and whether they can converge to the global optima is largely unknown. There has been some effort in this direction [7, 121], but more work needs to be done for a complete understanding of these problems.

The power of SDP in community detection under stochastic block models has been extensively studied as we saw in the main part of the thesis. However, for more complicated structured networks, for example those with hierarchical cluster structures or dynamic networks, whether SDP can be used to get theoretical guarantees is still largely unknown.

Identifying the latent structures in unsupervised data is a key ingredient in a diverse set of applications, starting from finding friends on a social network like Facebook to studying drug-drug or protein-protein interactions in medical problems; from viral marketing to image-segmentation; from documents understanding to context-based keyword search in databases. Networked data and relational data are being generated from corporate and public sources every day. Understanding the behavior of clustering algorithms holds the key to effectively utilizing these data sets and helps us define the boundary of our algorithmic conclusion.

Chapter 6

Appendix for EM Algorithm

6.1 Accompanying Lemmas

In this subsection, we collect some lemmas on Gaussian distribution and basic properties of Gaussian mixture model. Most of them can be derived with fundamental analysis techniques. The following lemma from [104] bounds the covering number of a unit sphere.

Lemma 6.1 (Lemma 5.2 [104]). *Let $\mathcal{S}^{n-1}$ be the unit Euclidean sphere equipped with Euclidean metric. Denote $\mathcal{N}(\mathcal{S}^{n-1}, \epsilon)$ as the covering number with ϵ-net, then*

$$\mathcal{N}(\mathcal{S}^{n-1}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n$$

Specifically, when $\epsilon = 1/2$, we have

$$\mathcal{N}(\mathcal{S}^{n-1}, \frac{1}{2}) \leq \exp(2n)$$

The following lemma is useful while carrying out spherical coordinate transformation.

Lemma 6.2. *(1) The volume for a d-dimensional r-ball is $\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} r^d$;*

(2) $\int_0^\pi \sin^k(x)dx = \frac{\sqrt{\pi}\Gamma(\frac{k+1}{2})}{\Gamma\left(\frac{k}{2}+1\right)}$, *and*

$$\int_{\theta_{d-1}=0}^{2\pi}\int_{\theta_{d-2}=0}^{\pi}\cdots\int_{\theta_1=0}^{\pi}\sin^{d-2}(\theta_1)\cdots\sin(\theta_{d-2})d\theta_1\cdots d\theta_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

(3) If $X \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$, *then*

$$\mathbb{E}_X\|X-\boldsymbol{\mu}\|^p = 2^{\frac{p}{2}}\frac{\Gamma\left(\frac{p+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\sigma^p$$

Proof. (1, 2) can be proven by elementary integration. Now we prove (3). By spherical coordinate transformation,

$$\mathbb{E}_X\|X-\boldsymbol{\mu}\|^p = (2\pi\sigma^2)^{-\frac{d}{2}}\int_{u=0}^{\infty}u^{p+d-1}e^{-\frac{u^2}{2\sigma^2}}du\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} = 2^{\frac{p}{2}}\frac{\Gamma\left(\frac{p+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\sigma^p$$

□

Lemma 6.3 (Gamma tail bound [14])**.** *If* $X \sim Gamma(v, c)$, *then* $P(X > \sqrt{2vt}+ct) \le e^{-t}$. *Or equivalently,*

$$P(X > t) \le \exp\left(-\frac{v}{c^2}\left(1+\frac{ct}{v}-\sqrt{1+\frac{2ct}{v}}\right)\right)$$

In particular, if $\frac{ct}{v} \ge 4$,

$$P(X > t) \le \exp\left(-\frac{v}{c^2}\sqrt{\frac{ct}{v}}\right) = \exp\left(-\sqrt{\frac{vt}{c^3}}\right)$$

Lemma 6.4. *For* $\forall d > 0$, *if* $\delta \ge 2\sqrt{d+1}$, *then*

$$\int_\delta^\infty u^d e^{-\frac{u^2}{2}}du \le 2^{\frac{d-1}{2}}\Gamma\left(\frac{d+1}{2}\right)\exp\left(-\frac{\delta}{2}\sqrt{d+1}\right)$$

For $p \in \{0,1,2\}$, when $\delta \geq 2\sqrt{d+p}$,

$$\int_{\delta}^{\infty} (u+x)^p u^{d-1} e^{-\frac{u^2}{2}} du \leq 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)(x+d)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right)$$

Proof. By changing of variables $v = \frac{u^2}{2}$ and integration by parts, we have

$$\begin{aligned}\int_{r}^{\infty} u^d e^{-\frac{u^2}{2}} du =& 2^{\frac{d-1}{2}} \int_{\frac{r^2}{2}}^{\infty} v^{\frac{d-1}{2}} e^{-v} dv \\ =& 2^{\frac{d-1}{2}}\Gamma\left(\frac{d+1}{2}\right) P(V > \frac{r^2}{2})\end{aligned}$$

where $V \sim \text{Gamma}(\frac{d+1}{2}, 1)$. By Lemma 6.3, if $r^2 \geq 4(1+d)$,

$$P\left(V > \frac{r^2}{2}\right) \leq \exp\left(-\frac{r}{2}\sqrt{d+1}\right)$$

Hence we have the first inequality. For the second, when $p = 0$, it follows directly from first part. When $p = 1$,

$$\begin{aligned}\int_{r}^{\infty} (u+x)^p u^{d-1} e^{-\frac{u^2}{2}} du =& \int_{r}^{\infty} u^d e^{-\frac{u^2}{2}} du + x\int_{r}^{\infty} u^{d-1} e^{-\frac{u^2}{2}} du \\ \leq& 2^{\frac{d-1}{2}}\Gamma\left(\frac{d+1}{2}\right)\exp\left(-\frac{r}{2}\sqrt{d+1}\right) + x2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)\exp\left(-\frac{r}{2}\sqrt{d}\right) \\ \leq& 2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)(x+d)\exp\left(-\frac{r}{2}\sqrt{d}\right)\end{aligned}$$

where we use $\Gamma\left(\frac{d+1}{2}\right) < \Gamma\left(\frac{d}{2}+1\right) = \frac{d}{2}\Gamma\left(\frac{d}{2}\right)$, and $\exp\left(-\frac{r}{2}\sqrt{d+1}\right) < \exp\left(-\frac{r}{2}\sqrt{d}\right)$ in the last step.

When $p = 2$,

$$\begin{aligned}
&\int_r^\infty (u+x)^2 u^{d-1} e^{-\frac{u^2}{2}} du = \int_r^\infty u^{d+1} e^{-\frac{u^2}{2}} du + 2x \int_r^\infty u^d e^{-\frac{u^2}{2}} du \\
&+ x^2 \int_r^\infty u^{d-1} e^{-\frac{u^2}{2}} du \\
\leq & 2^{\frac{d}{2}} \Gamma\left(\frac{d}{2}+1\right) \exp\left(-\frac{r}{2}\sqrt{d+2}\right) \\
&\quad + 2x \cdot 2^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d+1}\right) + x^2 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right) \\
\leq & (d + \sqrt{2}dx + x^2) 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right) \\
\leq & (x+d)^2 2^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) \exp\left(-\frac{r}{2}\sqrt{d}\right)
\end{aligned}$$

□

Using Lemma 6.4, we can get an easy to use tail bound for Euclidean norm of a Gaussian vector.

Lemma 6.5. *If $X \sim \mathcal{N}(0, I_d)$, for $r \geq 2\sqrt{d}$, we have*

$$P(\|X\| \geq r) \leq \exp(-\frac{r\sqrt{d}}{2})$$

Proof. By spherical coordinate transformation,

$$\begin{aligned}
P(\|X\| \geq r) = & \int (2\pi)^{-d/2} \exp(-\|x\|^2/2) dx \\
= & (2\pi)^{-d/2} \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_r^\infty r^{d-1} e^{-r^2/2} dr \\
\leq & \exp\left(-\frac{r}{2}\sqrt{d}\right)
\end{aligned}$$

□

Lemma 6.6. *If $X \sim GMM(\pi, \boldsymbol{\mu}^*, \sigma^2 I_d)$, then X is a sub-gaussian random vector with sub-gaussian norm $\sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|$.*

Proof. For any unit vector u, consider the random variable $X_u = \langle X, u \rangle$. By the definition in [104], it suffices to show that X_u has a sub-gaussian norm upper bounded by $\sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|$.

$$\|X_u\|_{\phi_2} = \sup_{p \geq 1} (\mathbb{E}|X_u|^p)^{1/p}$$

For any $p \geq 1$, let Z be the latent variable in the mixture model, we have

$$\begin{aligned}
p^{-1/2} \left(\mathbb{E}|X_u|^p\right)^{1/p} =& p^{-1/2} \left(\sum_{i=1}^M \mathbb{E}[|X_u|^p | Z = i] \cdot P(Z = i) \right)^{1/p} \\
\leq& p^{-1/2} \sum_{i=1}^M \pi_i \left(\mathbb{E}[|X_u|^p | Z = i]\right)^{1/p} \\
\overset{(i)}{\leq}& p^{-1/2} \sum_{i=1}^M \pi_i \left(\mathbb{E}[|X_u - \boldsymbol{\mu}_i^*|^p | Z = i]^{1/p} + \|\boldsymbol{\mu}_i^*\| \right) \\
\leq& p^{-1/2} \left(\sum_{i=1}^M \pi_i p^{1/2} \sigma + \|\boldsymbol{\mu}_i^*\| \right) \leq \sigma + \sum_{i=1}^M \pi_i \|\boldsymbol{\mu}_i^*\|
\end{aligned}$$

where (i) follows from Minkovski's inequality. □

The following lemma characterize the relation between $\|\boldsymbol{\mu}_{\max}^*\|$ and $R_{\max}$.

Lemma 6.7. *If $X \sim GMM(\pi, \boldsymbol{\mu}^*, \sigma^2 I_d)$ with $\mathbb{E}X = 0$, let $\|\boldsymbol{\mu}_{\max}^*\| = \max_i \|\boldsymbol{\mu}_i^*\|$, then*

$$\|\boldsymbol{\mu}_{\max}^*\| \leq R_{\max} \leq 2\|\boldsymbol{\mu}_{\max}^*\|$$

Proof. We first prove $\|\boldsymbol{\mu}^*_{\max}\| \leq R_{\max}$ by contradiction. Assume $\|\boldsymbol{\mu}^*_{\max}\| > R_{\max}$, by definition of $R_{\max}$, all the cluster centers lies in the ball $\mathbb{B}(\|\boldsymbol{\mu}^*_{\max}\|, R_{\max})$, but the origin is outside of the ball, which contradicts the fact that $\mathbb{E}X = \sum_i \pi_i \boldsymbol{\mu}^*_i = 0$.

The second inequality follows from triangle inequality, assume $R_{\max}$ is achieved at R_{ij}, then

$$R_{\max} \leq \|\boldsymbol{\mu}^*_i\| + \|\boldsymbol{\mu}^*_j\| \leq 2\|\boldsymbol{\mu}^*_{\max}\|.$$

□

Lemma 6.8. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\sqrt{n}L$ Lipschitz if there exists a constant L such that the restriction of f on a certain coordinate is L-Lipschitz.*

Proof. We first relax the norm of difference via a chain of triangle inequalities where each pair of terms only vary on one dimension.

$$\begin{aligned}
&|f(x_1, x_2, \cdots, x_n) - f(y_1, y_2, \cdots, x_n)| \\
\leq &\sum_{i=1}^{n} |f(y_1, y_2, \cdots, y_{i-1}, x_i, x_{i+1}, \cdots, x_n) - f(y_1, y_2, \cdots, y_{i-1}, y_i, x_{i+1}, \cdots, x_n)| \\
\leq &\sum_{i=1}^{n} L|x_i - y_i| \leq \sqrt{n}L \|x - y\|
\end{aligned}$$

□

6.2 Proofs in Section 2.4

Proof of Lemma 2.1. By (2.2), $\nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu}) = \mathbb{E}_X w_i(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_i)$. Without loss of generality, we only show the claim for $i = 1$. That is equivalent of saying, if $X \sim \text{GMM}(\pi, \boldsymbol{\mu}^*)$, we have $\mathbb{E}[w_1(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_1^*)] = 0$. Denote $\mathcal{N}(\boldsymbol{\mu}_i^*, \Sigma)$ as $\mathcal{N}_i$ and its distribution as $\phi_i(X)$. Decompose the left hand side with respect to the mixture components, we have

$$\begin{aligned}
\mathbb{E}[w_1(X)X] =& \sum_i \pi_i \mathbb{E}_{X\sim\mathcal{N}_i}[w_1(X)X] \\
=& \sum_i \pi_i \int \phi_i(X) \frac{\pi_1\phi_1(X)}{\sum_k \pi_k \phi_k(X)} X dx \\
=& \pi_1 \mathbb{E}_{X\sim\mathcal{N}_1} X = \pi_1 \boldsymbol{\mu}_1^*
\end{aligned}$$

Similarly $\mathbb{E}[w_1(X)] = \pi_1$. Hence $\nabla_{\boldsymbol{\mu}_1} q(\boldsymbol{\mu}) = \mathbb{E}_X w_1(X; \boldsymbol{\mu}^*)(X - \boldsymbol{\mu}_1) = \pi_1(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_1)$.

This completes the proof. □

Proof of Theorem 2.4. Define By Lemma 2.1, the GS condition is equivalent to

$$\left\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu})\right\| \leq \gamma \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$$

By triangle inequality,

$$\begin{aligned}\left\|\boldsymbol{\mu}_1^{t+1}-\boldsymbol{\mu}_1^*\right\| &= \left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*+s\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t)\right\| \\ &\leq \left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*+s\nabla q(\boldsymbol{\mu})\right\| + s\left\|\nabla Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t)-\nabla q(\boldsymbol{\mu})\right\| \\ &\leq \frac{\pi_{\max}-\pi_{\min}}{\pi_{\max}+\pi_{\min}}\left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*\right\| + \frac{2}{\pi_{\max}+\pi_{\min}}\gamma\left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*\right\| \\ &\leq \frac{\pi_{\max}-\pi_{\min}+2\gamma}{\pi_{\max}+\pi_{\min}}\left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*\right\|\end{aligned}$$

To see why the last inequality hold, notice that $q(\boldsymbol{\mu})$ has largest eigenvalue $-\pi_{\min}$ and smallest eigenvalue $-\pi_{\max}$. Apply the classical result for gradient descent, with step size $s=\frac{2}{\pi_{\max}+\pi_{\min}}$ guarantees

$$\left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*+s\nabla q(\boldsymbol{\mu})\right\| \leq \frac{\pi_{\max}-\pi_{\min}}{\pi_{\max}+\pi_{\min}}\left\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*\right\|$$

□

6.2.1 Proofs of Theorem 2.5

We start with two lemmas.

Lemma 6.9. *For $X\sim GMM(\pi,\boldsymbol{\mu}^*,I_d)$, if $R_{\min}=\tilde{\Omega}(\sqrt{d})$, and $\boldsymbol{\mu}_i\in\mathbb{B}(\boldsymbol{\mu}_i^*,a),\forall i\in[r]$ where*

$$a\leq\frac{R_{\min}}{2}-\sqrt{d}\max(4\sqrt{2[\log(R_{\min}/4)]_+},8\sqrt{3}).$$

Then for $p=0,1,2$ and $\forall i\in[r]$, we have

$$\mathbb{E}_X w_i(X;\boldsymbol{\mu})(1-w_i(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_i\|^p \leq 2r\left(\frac{3}{2}R_{\max}+d\right)^p\exp\left(-\left(\frac{R_{\min}}{2}-a\right)^2\sqrt{d}/8\right).$$

Using the same techniques, for the cross terms, we have the following lemma.

Lemma 6.10. *Assume* $X \sim GMM(\pi, \boldsymbol{\mu}^*, I_d)$, *and* $\boldsymbol{\mu}_i \in \mathbb{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [r]$. *Under the same conditions as in Lemma 6.9, we have for* $\forall i \neq j \in [r]$,

$$\begin{aligned}&\mathbb{E}_X[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_i\|\cdot\|X-\boldsymbol{\mu}_j\|]\\ \leq&(1+2\kappa)\left(\frac{3}{2}R_{\max}+d\right)^2\exp\left(-\left(\frac{R_{\min}}{2}-a\right)^2\sqrt{d}/8\right)\end{aligned}$$

Proof of Lemma 6.9. Without loss of generality, we prove the claim for $i = 1$. Recall the definition of $w_i(X;\boldsymbol{\mu})$ from Equation 2.1. For $p \in \{0, 1, 2\}$,

$$\begin{aligned}&\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p\\ =&\sum_{i\in[r]}\pi_i\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p\\ \leq&\pi_1\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)}w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p+\sum_{i\neq 1}\pi_i\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p\end{aligned} \tag{6.1}$$

First let us look at the first term. Define event $\mathcal{E}_\delta^{(1)} = \{X : X \sim \mathcal{N}(\boldsymbol{\mu}_1^*); \|X - \boldsymbol{\mu}_1^*\| \leq \delta\}$ for some $\delta > 0$. We will see later that we need $\delta < \frac{R_{\min}}{2} - a$. Then for $X \in \mathcal{E}_\delta^{(1)}$ using triangle inequality, we have

$$\|X-\boldsymbol{\mu}_i\|\begin{cases}\leq\|X-\boldsymbol{\mu}_i^*\|+\|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i\|\leq\delta+a & i=1\\ \geq\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_1^*\|-\|X-\boldsymbol{\mu}_1^*\|\geq\|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_1^*\|-\|\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i\|-\delta\geq R_{\min}-\delta-a & i\neq 1\end{cases} \tag{6.2}$$

$$
\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)} w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
=&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_\delta^{(1)}]P(\mathcal{E}_\delta^{(1)}) \\
&\quad + \mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_\delta^{(1)c}]P(\mathcal{E}_\delta^{(1)c})
\end{aligned}
$$

In view of the fact that $w_1(X;\boldsymbol{\mu})$ is monotonically decreasing w.r.t. $\|X-\boldsymbol{\mu}_i\|$ and increasing w.r.t. $\|X-\boldsymbol{\mu}_1\|$, we have

$$
\begin{aligned}
1-w_1(X;\boldsymbol{\mu}) \leq & \frac{(1-\pi_1)\exp\left(-\frac{(R_{\min}-\delta-a)^2}{2}\right)}{\pi_1\exp\left(-\frac{(\delta+a)^2}{2}\right)+(1-\pi_1)\exp\left(-\frac{(R_{\min}-\delta-a)^2}{2}\right)} \\
\leq & \frac{1-\pi_1}{\pi_1}\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2\delta-2a)\right)
\end{aligned}
$$

Also notice that $w_1(X;\boldsymbol{\mu})\leq 1$, we have

$$
\begin{aligned}
&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p|\mathcal{E}_\delta^{(1)}]P(\mathcal{E}_\delta^{(1)}) \\
\leq & \frac{1-\pi_1}{\pi_1}\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2\delta-2a)\right)(\delta+a)^p
\end{aligned}
$$

For $\mathcal{E}_\delta^{(1)c}$, note $w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\leq\frac{1}{4}$, we have for $p=1$,

$$
\begin{aligned}
&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\||\mathcal{E}_\delta^{(1)c}]P(\mathcal{E}_\delta^{(1)c}) \\
\leq & \frac{1}{4}\int_{u=\delta}^{\infty}(u+a)(2\pi)^{-\frac{d}{2}}\exp\left(-\frac{u^2}{2}\right)\cdot\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}u^{d-1}du \\
\leq & \frac{1}{4}(2\pi)^{-\frac{d}{2}}\frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}\int_{u=\delta}^{\infty}(u+a)\exp\left(-\frac{u^2}{2}\right)u^{d-1}du \\
\overset{(i)}{\leq} & \frac{a+d}{4}\exp\left(-\frac{\delta}{2}\sqrt{d}\right)
\end{aligned}
$$

The inequality (i) follows from Lemma 6.4 when $\delta>2\sqrt{d+1}$. Similarly, for

$p = 2$,

$$\begin{aligned}
&\mathbb{E}[w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^2|\mathcal{E}_\delta^{(1)c}]P(\mathcal{E}_\delta^{(1)c}) \\
\leq &\frac{2^{-\frac{d}{2}-1}}{\Gamma\left(\frac{d}{2}\right)}\int_\delta^\infty (u+a)^2 u^{d-1}e^{-\frac{u^2}{2}}du \overset{(ii)}{\leq} \frac{(a+d)^2}{4}\exp\left(-\frac{\delta}{2}\sqrt{d}\right)
\end{aligned}$$

The inequality (ii) follows from Lemma 6.4 when $\delta > 2\sqrt{d+1}$ and $p = 2$. Therefore for the first mixture we have,

$$\begin{aligned}
&\pi_1\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*)}w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
\leq &(1-\pi_1)(\delta+a)^p\exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2\delta-2a)\right)+\pi_1\frac{(a+d)^p}{4}\exp\left(-\frac{\delta}{2}\sqrt{d}\right)
\end{aligned} \tag{6.3}$$

Next we bound $\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p$ for $i \neq 1$. For some $0 < \delta < \frac{R}{2} - a$, we have

$$\begin{aligned}
&\pi_i\mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)}w_1(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_1\|^p \\
=&\int_X \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX \\
=&\underbrace{\int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,\delta)}\frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX}_{I_1^{(p)}} \\
&+\underbrace{\int_{X\notin\mathbb{B}(\boldsymbol{\mu}_i^*,\delta)}\frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\cdot\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j\pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX}_{I_2^{(p)}}
\end{aligned} \tag{6.4}$$

When $\|X-\boldsymbol{\mu}_i^*\| \leq \delta$, since by assumption $\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*\| \leq a$,

$$\begin{aligned}
\frac{\phi(X;\boldsymbol{\mu}_i^*)}{\phi(X;\boldsymbol{\mu}_i)} &= \exp\left(\frac{\|X-\boldsymbol{\mu}_i\|^2}{2}-\frac{\|X-\boldsymbol{\mu}_i^*\|^2}{2}\right) \\
&= \exp\left(\left(X-\frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^*}{2}\right)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*)\right)
\end{aligned} \tag{6.5}$$

Since by Cauchy-Schwarz we have $|(X - \frac{\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^*}{2})^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*)| = |(X - \boldsymbol{\mu}_i^* + \frac{\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_i}{2})^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*)| \leq (\delta + a/2)a$, we have:

$$\exp\left(-(\delta+\frac{a}{2})a\right) \leq \frac{\phi(X;\boldsymbol{\mu}_i^*)}{\phi(X;\boldsymbol{\mu}_i)} \leq \exp\left((\delta+\frac{a}{2})a\right) \tag{6.6}$$

For such X, $\phi(X;\boldsymbol{\mu}_1) \leq (2\pi)^{-\frac{d}{2}}\exp\left(-\frac{(R_{\min}-\delta-a)^2}{2}\right)$, and we have

$$\begin{aligned}
I_1^{(p)} =& \int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,\delta)} \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\pi_i\phi(X;\boldsymbol{\mu}_i^*)}{\sum_j \pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX \\
\leq& \int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,\delta)} \frac{\pi_1\phi(X;\boldsymbol{\mu}_1)\pi_i\phi(X;\boldsymbol{\mu}_i)\exp\left((\delta+\frac{a}{2})a\right)}{\sum_j \pi_j\phi(X;\boldsymbol{\mu}_j)}\|X-\boldsymbol{\mu}_1\|^p dX \\
\leq& \pi_1\exp\left((\delta+\frac{a}{2})a\right)\int_{X\in\mathbb{B}(\boldsymbol{\mu}_i^*,\delta)}\phi(X;\boldsymbol{\mu}_1)\|X-\boldsymbol{\mu}_1\|^p dX \\
\leq& \pi_1(2\pi)^{-d/2}\exp\left((\delta+\frac{a}{2})a\right)(R_{\max}+a+\delta)^p\exp\left(-\frac{(R_{\min}-\delta-a)^2}{2}\right)\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}\delta^d \\
\leq& \frac{\pi_1 2^{-d/2}}{\Gamma(\frac{d}{2}+1)}\exp\left((\delta+\frac{a}{2})a - \frac{(R_{\min}-\delta-a)^2}{2}\right)(R_{\max}+a+\delta)^p\delta^d \\
\leq& \pi_1 2^{1-d}\exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right)\right)(R_{\max}+a+\delta)^p\delta^d
\end{aligned}$$

The last inequality follows from the fact that $\Gamma\left(\frac{d}{2}+1\right) \geq ([\frac{d}{2}])! \geq 2^{\frac{d}{2}-1}$. On the other hand, for I_2, since $w_1(X;\boldsymbol{\mu}) \leq 1$, taking spherical coordinate transformation we have,

$$\begin{aligned}
I_2^{(p)} \leq& \int_{\|X-\boldsymbol{\mu}_i^*\|\geq\delta}\pi_i\phi(X;\boldsymbol{\mu}_i^*)\|X-\boldsymbol{\mu}_1\|^p dX \\
\leq& \pi_i\int_{\|X-\boldsymbol{\mu}_i^*\|\geq\delta}(2\pi)^{-d/2}\exp(-\frac{\|X-\boldsymbol{\mu}_i^*\|^2}{2})\|X-\boldsymbol{\mu}_1\|^p dX \\
\leq& \frac{\pi_i 2^{1-d/2}}{\Gamma(\frac{d}{2})}\int_{u=\delta}^{\infty} u^{d-1}\exp\left(-\frac{u^2}{2}\right)(u+R_{\max}+a)^p du
\end{aligned}$$

Apply Lemma 6.4, when $\delta \geq 2\sqrt{d+2}$, for $p \in \{0,1,2\}$

$$I_2^{(p)} \leq \pi_i \left(R_{\max} + a + d\right)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right) \tag{6.7}$$

Summing up I_1 and I_2, for any $0 < \delta < R_{\min}/2$, from (6.4) we get:

$$\begin{aligned}
&\pi_i \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_i^*)} w_1(X;\boldsymbol{\mu}) \|X - \boldsymbol{\mu}_1\|^p \\
\leq &\pi_1 2^{1-d} \exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right)\right)\left(R_{\max}+a+\delta\right)^p \delta^d &(6.8)\\
&+ \pi_i \left(R_{\max}+a+d\right)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right) &(6.9)
\end{aligned}$$

Now plugging Eq. (6.3) and Eq. (6.9) into Eq. (6.1) gives,

$$\begin{aligned}
&\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))\|X-\boldsymbol{\mu}_1\|^p \\
\leq &(1-\pi_1)(\delta+a)^p \exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2\delta-2a)\right) + \pi_1 \frac{(a+d)^p}{4}\exp\left(-\frac{\delta}{2}\sqrt{d}\right) \\
&+ \pi_1(r-1)2^{1-d}\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right)\right)\left(R_{\max}+a+\delta\right)^p\delta^d \\
&+ (1-\pi_1)\left(R_{\max}+a+d\right)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right) \\
\leq &\underbrace{(1-\pi_1)(r+a)^p \exp\left(-\frac{1}{2}R_{\min}(R_{\min}-2\delta-2a)\right)}_{(A)} + \underbrace{(R_{\max}+a+d)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right)}_{(B)} \\
&+ \underbrace{2\pi_1(r-1)\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right) + d\log(\delta/2)\right)\left(R_{\max}+a+\delta\right)^p}_{(C)}
\end{aligned}$$

Note that in order to have a negative term inside exponential of (A), we require $\delta + a < \frac{R_{\min}}{2}$. In order to ensure the same for (C), we need:

$$a < \frac{R_{\min}}{2}\left(1 - \frac{\delta}{R_{\min}}\right)^2 \tag{6.10}$$

If $\delta^2 \geq 2d\log(\delta/2)$, then we have:

$$\begin{aligned}
&\exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right) + d\log(\delta/2)\right) \\
&\leq \exp\left(R_{\min}\left(a - \frac{R_{\min}}{2}(1-\delta/R_{\min})^2\right) + \delta^2/2\right) \\
&\leq \exp\left(R_{\min}a - \left(\frac{R_{\min}^2}{2} - \delta R_{\min} + \frac{\delta^2}{2}\right) + \frac{\delta^2}{2}\right) \\
&= \exp\left(-\frac{1}{2}R_{\min}(R_{\min} - 2\delta - 2a)\right)
\end{aligned}$$

Therefore, $(A)+(C) \leq (1-\pi_1+2\pi_1(r-1))(R_{\max}+a+\delta)^p \exp\left(-\frac{1}{2}R_{\min}(R_{\min} - 2\delta - 2a)\right)$.

Finally, if $\delta \leq R_{\min}\frac{R_{\min}/2-a}{R_{\min}+\sqrt{d}/2}$, we have:

$$\exp\left(-\frac{1}{2}R_{\min}(R_{\min} - 2\delta - 2a)\right) \leq \exp(-\frac{\delta}{2}\sqrt{d})$$

Hence,

$$\begin{aligned}
(A)+(B)+(C) \leq& (2-\pi_1+2\pi_1(r-1))\left(\frac{3}{2}R_{\max}+d\right)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right) \\
\leq& 2r\left(\frac{3}{2}R_{\max}+d\right)^p \exp\left(-\frac{\delta}{2}\sqrt{d}\right)
\end{aligned}$$

Set

$$\delta = \frac{R_{\min}/2 - a}{4}, \quad a \leq \frac{R_{\min}}{2} \tag{6.11}$$

then Eq (6.10) and $a+\delta \leq \frac{R_{\min}}{2}$ are automatically satisfied. When $R_{\min} \geq \frac{\sqrt{d}}{6}$, we have $\delta \leq R_{\min}\frac{R_{\min}/2-a}{R_{\min}+\sqrt{d}/2}$. Finally in order to meet the constraints

$$\delta \geq 2\sqrt{d+2} \Leftarrow \delta \geq 3\sqrt{d} \tag{6.12}$$

$$\delta^2 \geq 2d\log\delta/2 \tag{6.13}$$

we need

$$\frac{R_{\min}/2 - a}{4} \geq \max(\sqrt{2d[\log(R_{\min}/4)]_+}, 2\sqrt{3}\sqrt{d})$$
$$a \leq \frac{R_{\min}}{2} - \sqrt{d}\max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3})$$

The right hand side of last inequality is non-negative when $R_{\min} = \tilde{\Omega}(\sqrt{d})$. Under these conditions, with Eq. (6.11) plugged in, we have

$$\mathbb{E}_X w_1(X;\boldsymbol{\mu})(1 - w_1(X;\boldsymbol{\mu}))\|X - \boldsymbol{\mu}_1\|^p \leq 2r\left(\frac{3}{2}R_{\max} + d\right)^p \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{d}/8\right)$$

□

Proof of Lemma 6.10. For any $\delta \leq \frac{R_{\min}}{2} - a$, define $\mathcal{E}_0 = \{X : \exists i, \text{ such that } Z_X = i, \|X - \boldsymbol{\mu}_i^*\| > \delta\}$ and $\mathcal{E}_k = \{X : Z_X = k, \|X - \boldsymbol{\mu}_k^*\| \leq \delta\}$.

$$\begin{aligned}
&\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\| \cdot \|X - \boldsymbol{\mu}_j\|\right] \\
\leq &\underbrace{\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\||\mathcal{E}_0\right]P(\mathcal{E}_0)}_{I_0} \\
&+ \sum_{k\in[r]} \underbrace{\pi_k \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_k^*)}\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\|\|X - \boldsymbol{\mu}_k\| \leq \delta\right]}_{I_k}
\end{aligned}$$

First we look at I_0, this again can be decomposed as the sum over mixtures. Similarly as in Eq. (6.7), we have

$$I_0 \leq (R_{\max} + a + d)^2 \exp\left(-\frac{\delta}{2}\sqrt{d}\right)$$

For I_k, by Eq. (6.6),

$$
\begin{aligned}
I_k =& \int_X \frac{\pi_i\phi(X;\boldsymbol{\mu}_i)\pi_j\phi(X;\boldsymbol{\mu}_j)\pi_k\phi(X;\boldsymbol{\mu}_k^*)}{(\sum_\ell \pi_\ell\phi(X;\boldsymbol{\mu}_t))^2}\|X-\boldsymbol{\mu}_i\|\cdot\|X-\boldsymbol{\mu}_j\|dX \\
\leq& \int_X \frac{\pi_i\phi(X;\boldsymbol{\mu}_i)\pi_j\phi(X;\boldsymbol{\mu}_j)\pi_k\phi(X;\boldsymbol{\mu}_k)\exp((\delta+a/2)a)}{(\sum_\ell \pi_\ell\phi(X;\boldsymbol{\mu}_\ell))^2}\|X-\boldsymbol{\mu}_i\|\cdot\|X-\boldsymbol{\mu}_j\|dX \\
\leq& \kappa\pi_k 2\pi^{-\frac{d}{2}}\exp(-\frac{R_{(\min}-\delta-a)^2}{2})\exp((\delta+a/2)a)(R_{\max}+\delta+a)^2\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}\delta^d \\
\leq& \pi_k\kappa 2^{-d/2}\frac{1}{\Gamma\left(\frac{d}{2}+1\right)}\delta^d\exp\left((\delta+a/2)a-\frac{(R_{\min}-\delta-a)^2}{2}\right)(R_{\max}+\delta+a)^2 \\
\leq& 2\pi_k\kappa\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}\left(1-\frac{\delta}{R_{\min}}\right)^2\right)+d\log(\delta/2)\right)(R_{\max}+\delta+a)^2
\end{aligned}
\tag{6.14}
$$

Adding up I_k's and I_0, we have

$$
\begin{aligned}
&\mathbb{E}_X\left[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_i\|\|X-\boldsymbol{\mu}_j\|\right] \\
\leq& (R_{\max}+a+d)^2\exp\left(-\frac{\delta}{2}\sqrt{d}\right) \\
&\quad + 2\kappa\exp\left(R_{\min}\left(a-\frac{R_{\min}}{2}\left(1-\frac{\delta}{R_{\min}}\right)^2\right)+d\log(\delta/2)\right)(R_{\max}+\delta+a)^2
\end{aligned}
$$

Take $\delta = \frac{1}{4}\left(\frac{R_{\min}}{2}-a\right)$, we have $R_{\min}\left(a-\frac{R_{\min}}{2}\left(1-\frac{\delta}{R_{\min}}\right)^2\right)+d\log(\delta/2) \leq -\frac{\delta}{2}\sqrt{d}$. Therefore,

$$
\begin{aligned}
&\mathbb{E}_X[w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X-\boldsymbol{\mu}_i\|\cdot\|X-\boldsymbol{\mu}_j\|] \\
\leq& (1+2\kappa)\left(\frac{3}{2}R_{\max}+d\right)^2\exp\left(-\left(\frac{R_{\min}}{2}-a\right)^2\sqrt{d}/8\right)
\end{aligned}
$$

□

Proof of Theorem 2.5. Consider the difference of the gradient corresponding

to $\boldsymbol{\mu}_i$, without loss of generality, assume $i = 1$.

$$\nabla_{\boldsymbol{\mu}_1} Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t) - \nabla q(\boldsymbol{\mu}^t) = \mathbb{E}(w_1(X;\boldsymbol{\mu}^t) - w_1(X;\boldsymbol{\mu}^*))(X - \boldsymbol{\mu}_1^t) \tag{6.15}$$

For any given X, consider the function $\boldsymbol{\mu} \to w_1(X;\boldsymbol{\mu})$, we have

$$\nabla_{\boldsymbol{\mu}} w_1(X;\boldsymbol{\mu}) = \begin{pmatrix} w_1(X;\boldsymbol{\mu})(1 - w_1(X;\boldsymbol{\mu}))(X - \boldsymbol{\mu}_1)^T \\ -w_1(X;\boldsymbol{\mu})w_2(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_2)^T \\ \vdots \\ -w_1(X;\boldsymbol{\mu})w_r(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_r)^T \end{pmatrix} \tag{6.16}$$

Let $\boldsymbol{\mu}^u = \boldsymbol{\mu}^* + u(\boldsymbol{\mu}^t - \boldsymbol{\mu}^*), \forall u \in [0,1]$, obviously $\boldsymbol{\mu}^u \in \otimes_{i=1}^r \mathbb{B}(\boldsymbol{\mu}_i^*, \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|) \subset \otimes_{i=1}^r \mathbb{B}(\boldsymbol{\mu}_i^*, a)$. By Taylor's theorem,

$$\begin{aligned}
&\|\mathbb{E}(w_1(X;\boldsymbol{\mu}_1^t) - w_1(X;\boldsymbol{\mu}_1^*))(X - \boldsymbol{\mu}_1^t)\| = \left\| \mathbb{E}\left[\int_{u=0}^1 \nabla_u w_1(X;\boldsymbol{\mu}^u) du (X - \boldsymbol{\mu}_1^t) \right] \right\| \\
=& \left\| \int_{u=0}^1 \mathbb{E} w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^u)^T(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*)(X - \boldsymbol{\mu}_1^t) du \right. \\
&\qquad \left. - \sum_{i \neq 1} \int_{u=0}^1 \mathbb{E} w_1(X;\boldsymbol{\mu}^u) w_i(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_2^u)^T(\boldsymbol{\mu}_2^t - \boldsymbol{\mu}_2^*)(X - \boldsymbol{\mu}_1^t) du \right\| \\
\leq& U_1 \|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_1^*\|_2 + \sum_{i \neq 1} U_i \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2
\end{aligned} \tag{6.17}$$

where

$$\begin{aligned}
U_1 &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X;\boldsymbol{\mu}^u)(1 - w_1(X;\boldsymbol{\mu}^u))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^u)^T\|_{op} \\
U_i &= \sup_{u \in [0,1]} \|\mathbb{E} w_1(X;\boldsymbol{\mu}^u) w_i(X;\boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_2^u)^T\|_{op}
\end{aligned}$$

For U_1 by triangle inequality we have,

$$\begin{aligned} U_1 &\leq \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1-w_1(X;\boldsymbol{\mu}^u))(X-\boldsymbol{\mu}_1^u)(X-\boldsymbol{\mu}_1^u)^T\|_{op} \\ &\quad + \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1-w_1(X;\boldsymbol{\mu}^u))(\boldsymbol{\mu}_1^u-\boldsymbol{\mu}_1^t)(X-\boldsymbol{\mu}_1^u)^T\|_{op} \\ &\leq \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1-w_1(X;\boldsymbol{\mu}^u))(X-\boldsymbol{\mu}_1^u)(X-\boldsymbol{\mu}_1^u)^T\|_{op} \\ &\quad + a \sup_{u\in[0,1]} \|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1-w_1(X;\boldsymbol{\mu}^u))(X-\boldsymbol{\mu}_1^u)\| \end{aligned} \tag{6.18}$$

We now develop an uniform bound for the operator norm. For any $u \in [0,1]$, there exists a rotation matrix O, such that all $R\boldsymbol{\mu}_i^u, i \in [r]$ have non-zero entries in the leading $\tilde{d}$ coordinates, and zeros for the remaining $[d-r]_+$ coordinates. Denote $\tilde{X} := OX$, then $\tilde{X}|Z = i \sim \mathcal{N}(O\boldsymbol{\mu}_i^*, I_d)$. Let

$$O\boldsymbol{\mu}_i^u = [\tilde{\boldsymbol{\mu}}_i^u, 0_{[d-r]_+}] \text{ and } O\boldsymbol{\mu}_i^* = [v_i^{\tilde{d}}, v_i^{[d-r]_+}], \quad \tilde{\boldsymbol{\mu}}_i^u \in \mathbb{R}^{\tilde{d}}$$

For ease of notation, we assume $d \geq r$ for now, the other case can be derived without much modification. We can rewrite

$$(X-\boldsymbol{\mu}_1^u)(X-\boldsymbol{\mu}_1^u)^T = O^T \begin{bmatrix} (\tilde{X}^r-\tilde{\boldsymbol{\mu}_1^u})(\tilde{X}^r-\tilde{\boldsymbol{\mu}_1^u})^T & (\tilde{X}^r-\tilde{\boldsymbol{\mu}_1^u})(\tilde{X}^{d-r})^T \\ (\tilde{X}^{d-r})(\tilde{X}^r-\tilde{\boldsymbol{\mu}_1^u})^T & (\tilde{X}^{d-r})(\tilde{X}^{d-r})^T \end{bmatrix} O$$

Note by the rotation, $w_i(X;\boldsymbol{\mu})$ only depend on the first r coordinates. And by isotropicity, $\tilde{X}^r$ and $\tilde{X}^{d-r}$ are independent. By $\mathbb{E}\tilde{X}^{d-r} = 0$ (since we assume that the centroid of the means is at zero, and a rotation does not change that) and $\mathbb{E}\tilde{X}^{d-r}(\tilde{X}^{d-r})^T = I_{d-r} + \sum_i \pi_i (v_i^{d-r})(v_i^{d-r})^T$, we have,

$$\begin{aligned} &\|\mathbb{E}w_1(X;\boldsymbol{\mu}^u)(1-w_1(X;\boldsymbol{\mu}^u))(X-\boldsymbol{\mu}_1^u)(X-\boldsymbol{\mu}_1^u)^T\|_{op} = \left\| \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \right\|_{op} \\ &\leq \max\{\|D_1\|_{op}, \|D_2\|_{op}\} \end{aligned}$$

D_1 and D_2 are defined below. Applying Lemma 6.9 with dimension $\tilde{d}$, when $R_{\min} = \Omega(\sqrt{\tilde{d}})$,

$$\begin{aligned}\|D_1\|_{op} = &\|\mathbb{E} w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u))(\tilde{X}^{\tilde{d}} - \tilde{\boldsymbol{\mu}}_1^u)(\tilde{X}^{\tilde{d}} - \tilde{\boldsymbol{\mu}}_1^u)^T\|_{op} \\ \leq &2r\left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)\end{aligned}$$

For D_2, by independence and Lemma 6.9, when $R_{\min} = \Omega(\sqrt{\tilde{d}})$,

$$\begin{aligned}\|D_2\|_{op} = &\left\|\mathbb{E} w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}; \tilde{\boldsymbol{\mu}}^u))\left(I_{[d-r]_+} + \sum_i \pi_i (v_i^{[d-r]_+})(v_i^{[d-r]_+})^T\right)\right\|_{op} \\ = &\left\|\left(\mathbb{E}_{\tilde{X}_{\tilde{d}}} w_1(\tilde{X}_{\tilde{d}}; \tilde{\boldsymbol{\mu}}^u)(1 - w_1(\tilde{X}_{\tilde{d}}; \tilde{\boldsymbol{\mu}}^u))\right) \cdot \mathbb{E}_{X_{[d-r]_+}}\left(I_{[d-r]_+} + \sum_i \pi_i (v_i^{[d-r]_+})(v_i^{[d-r]_+})^T\right)\right\|_{op} \\ \leq &(R_{\max}^2 + 1) 2r \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)\end{aligned}$$

Combining the two and plugging in Eq. (6.18),

$$\begin{aligned}U_1 \leq &2r \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right) \cdot \\ &\left(\max\left\{\left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2, (R_{\max}^2 + 1)\right\} + a\left(\frac{3}{2}R_{\max} + \tilde{d}\right)\right) \\ \leq &2r\left(2R_{\max} + \tilde{d}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)\end{aligned}$$

The max will always be achieved at the first term as $\tilde{d} \geq 1$. Similarly, with the same rotation, for $U_i, i \neq 1$,

$$U_i \leq \sup_u \|\mathbb{E} w_1(X; \boldsymbol{\mu}^u) w_i(X; \boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_1^u)(X - \boldsymbol{\mu}_i^u)^T\|_{op} + a\|\mathbb{E} w_1(X; \boldsymbol{\mu}^u) w_i(X; \boldsymbol{\mu}^u)(X - \boldsymbol{\mu}_i^u)\|$$

By Lemma 6.10, when $R_{\min} = \Omega(\sqrt{\tilde{d}})$, we have

$$\begin{aligned}
U_i \leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right) \cdot \\
& \left(\max\left\{(1+2\kappa)\left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2, 2r(R_{\max}^2+1)\right\} + 2ra\left(\frac{3}{2}R_{\max} + \tilde{d}\right)\right) \\
\leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)\left(\frac{3}{2}R_{\max} + \tilde{d}\right) \\
& \cdot \left(\max\{(1+2\kappa), 2r\}\left(\frac{3}{2}R_{\max} + \tilde{d}\right) + 2ra\right) \\
\leq & \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)\left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2 \cdot \max\{3r, r+2\kappa+1\} \\
\leq & r(2\kappa+4)\left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right)
\end{aligned}$$

The second inequality is because $R_{\max}^2 + 1 \leq \left(\frac{3}{2}R_{\max} + \tilde{d}\right)^2$ and the third inequality is because $2a \leq \frac{3}{2}R_{\max} + \tilde{d}$. Taking back to Eq. (6.17), and summing over $i \in [r]$, we have

$$\begin{aligned}
& \|\nabla_{\boldsymbol{\mu}_i} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) - \nabla_{\boldsymbol{\mu}_i} q(\boldsymbol{\mu})\| \\
\leq & r(2\kappa+4)\left(2R_{\max} + \tilde{d}\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \sqrt{\tilde{d}}/8\right) \sum_{i=1}^{r} \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|
\end{aligned}$$

This completes the proof. □

6.2.2 Proof of Theorem 2.1

Proof of Theorem 2.1. By Theorem 2.5 and Theorem 2.4, it suffices to check $\gamma \leq \pi_{\min}$. Solving the inequality we have

$$a \leq \frac{R_{\min}}{2} - \frac{2\sqrt{2}}{\sqrt[4]{\tilde{d}}}\sqrt{\log\left(\frac{r^2(2\kappa+4)(2R_{\max}+\tilde{d})^2}{\pi_{\min}}\right)}$$

Combined with the condition in Theorem 2.5, we have

$$\begin{aligned} a \leq & \frac{R_{\min}}{2} - \max\left\{\frac{2\sqrt{2}}{\sqrt[4]{\tilde{d}}}\sqrt{\log\left(\frac{r^2(2\kappa+4)(2R_{\max}+\tilde{d})^2}{\pi_{\min}}\right)}, \right. \\ & \left. \sqrt{\tilde{d}}\max(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3})\right\} \\ = & \frac{R_{\min}}{2} - \sqrt{\tilde{d}}o(R_{\min}) \end{aligned}$$

because

$$\begin{aligned} & \max\left\{c\sqrt{\log(c_1\frac{r^2\kappa}{\pi_{\min}} + 2\log\left(2R_{\max}+\tilde{d}\right)}, \sqrt{\tilde{d}}\max\{c_2\sqrt{\log(R_{\min}/4)_+}, 8\sqrt{3}\}\right\} \\ \leq & \max\left\{c\sqrt{\log(c_1\frac{r^2\kappa}{\pi_{\min}} + c_2R_{\max} + c_3\tilde{d})}, c'\sqrt{\tilde{d}}\sqrt{\log(R_{\max}+e)}\right\} \\ \leq & \sqrt{\tilde{d}}O\left(\sqrt{\log\left(\max\left\{\frac{r^2\kappa}{\pi_{\min}}, R_{\max}, \tilde{d}\right\}\right)}\right) \end{aligned}$$

The condition in Theorem 2.5 can be rewritten as

$$a \leq \frac{R_{\min}}{2} - \sqrt{\tilde{d}}O\left(\sqrt{\log\left(\max\left\{\frac{r^2\kappa}{\pi_{\min}}, R_{\max}, \tilde{d}\right\}\right)}\right)$$

□

6.3 Proofs for sample-based gradient EM

In this section we develop the error bound for sample-based gradient EM. Our proof is based on the Rademacher complexity theory and some new tools for contraction result. In [76], Maurer has the following contraction result for the complexity defined over countable sets.

Lemma 6.11 (Theorem 3 [76])**.** *Let X be nontrivial, symmetric and sub-gaussian. Then there exists a constant $C < \infty$, depending only on the distribution of X, such that for any countable set $\mathcal{S}$ and function $h_i : \mathcal{S} \to \mathbb{R}$, $f_i : \mathcal{S} \to \mathbb{R}^k$, $i \in [n]$ satisfying $\forall s, s' \in \mathcal{S}, |h_i(s) - h_i(s')| \leq L\|f(s) - f(s')\|$. If ϵ_{ik} is an independent doubly indexed Rademacher sequence, we have,*

$$\mathbb{E} \sup_{s \in \mathcal{S}} \sum_i \epsilon_i h_i(s) \leq \mathbb{E}\sqrt{2}L \sup_{s \in \mathcal{S}} \sum_{i,k} \epsilon_{ik} f_i(s)_k,$$

where $f_i(s)_k$ is the k-th component of $f_i(s)$.

We prove Lemma 2.2 by generalizing this result to any subset of separable Banach space.

Proof of Lemma 2.2. First note that a subset of a separable subspace is separable, and has a dense countable subset; lets call this $\mathcal{S}_0$. Now note that if the Lipschitz condition holds for $s, s' \in \mathcal{S}$, then it also holds for $s, s' \in \mathcal{S}_0$. Now applying Lemma 6.11, we see that

$$\mathbb{E} \sup_{s \in \mathcal{S}_0} \sum_i \epsilon_i h_i(s) \leq \mathbb{E}\sqrt{2}L \sup_{s \in \mathcal{S}_0} \sum_{i,k} \epsilon_{ik} f_i(s)_k,$$

All we need to prove is that the two supremas over $\mathcal{S}_0$ on the LHS and RHS of the above equation can be replaced by supremum over $\mathcal{S}$. We will only show this for the LHS. The argument for the RHS is identical. In order to show this, we need to also make sure that $g(s) := \sum_i \epsilon_i h_i(s)$ over $\mathcal{S}$ is measurable. We show this using standard tools from measure theory.

We want to show that:

$$\sup_{s \in \mathcal{S}} g(s) = \sup_{s \in \mathcal{S}_0} g(s). \tag{6.19}$$

Since $g(s)$ is continuous, its also measurable for all $s \in \mathcal{S}$. The above statement, once proven, essentially implies that the sup over $\mathcal{S}$ is the same as the sup over a countable set $\mathcal{S}_0$. Since pointwise sup over measurable functions is measurable, we are done. We now prove Eq. (6.19). It is clear that, $\sup_{s \in \mathcal{S}} g(s) \geq \sup_{s \in \mathcal{S}_0} g(s)$. So all we need is to prove that for all $\epsilon > 0$.

$$\sup_{s \in \mathcal{S}} g(s) \leq \sup_{s \in \mathcal{S}_0} g(s) + \epsilon \tag{6.20}$$

Since $g(s)$ is continuous, let $D_1(s) = \{s' \in \mathcal{S} : |g(s) - g(s')| \leq \epsilon\}$. Furthermore, since $\mathcal{S}_0$ is dense in $\mathcal{S}$, we also have $D_2(s, \epsilon) := D_1(s) \cap \mathcal{S}_0 \neq \phi$. So for each $s \in \mathcal{S}$, and $\epsilon > 0$, $\exists s' \in \mathcal{S}_0$ (to be precise, $s' \in D_2(s, \epsilon)$) such that $g(s) \leq g(s') + \epsilon$. Taking a sup over the LHS over $\mathcal{S}$ and a sup of RHS over $\mathcal{S}_0$, we get Eq. (6.20). This completes the proof. □

Proof of Theorem 2.2. For any unit vector u, the Rademacher complexity of

$\mathcal{F}$ is

$$\begin{aligned} R_n(\mathcal{F}) =& \mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle X_i - \boldsymbol{\mu}_1, u \rangle \\ \leq & \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle X_i, u \rangle}_{(D)} + \underbrace{\mathbb{E}_X \mathbb{E}_\epsilon \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n} \sum_{i=1}^n \epsilon_i w_1(X_i; \boldsymbol{\mu}) \langle \boldsymbol{\mu}_1, u \rangle}_{(E)} \end{aligned} \tag{6.21}$$

We bound the two terms separately. Define $\eta_j(\boldsymbol{\mu}) : \mathbb{R}^{rd} \to \mathbb{R}^r$ to be a vector valued function with the k-th coordinate

$$[\eta_j(\boldsymbol{\mu})]_k = \frac{\|\boldsymbol{\mu}_1\|^2}{2} - \frac{\|\boldsymbol{\mu}_k\|^2}{2} + \langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1 \rangle + \log\left(\frac{\pi_k}{\pi_1}\right)$$

We claim

$$|w_1(X_j; \boldsymbol{\mu}) - w_1(X_j; \boldsymbol{\mu}')| \leq \frac{\sqrt{r}}{4} \|\eta_j(\boldsymbol{\mu}) - \eta_j(\boldsymbol{\mu}')\| \tag{6.22}$$

This vectorized Lipschitz condition simply follows from the fact that

$$\begin{aligned} w_1(X_j, \boldsymbol{\mu}) =& \frac{1}{1 + \sum_{k=2}^r \exp([\eta_j(\boldsymbol{\mu})]_k)} \\ \frac{\partial w_1(X_j, \boldsymbol{\mu})}{\partial [\eta_j(\boldsymbol{\mu})]_k} =& \frac{\exp([\eta_j(\boldsymbol{\mu})]_k)}{(1 + \sum_{k=2}^r \exp([\eta_j(\boldsymbol{\mu})]_k))^2} \leq \frac{1}{4} \end{aligned}$$

so $w_1(X_j, \boldsymbol{\mu})$ is $\frac{1}{4}$-Lipschitz continuous w.r.t. $[\eta_j(\boldsymbol{\mu})]_k$. By Lemma 6.8, $w_1(X_j, \boldsymbol{\mu})$ is $\frac{\sqrt{r}}{4}$ Lipschitz w.r.t $\eta_j(\boldsymbol{\mu})$. Now let $\psi_j(\boldsymbol{\mu}) = w_1(X_j; \boldsymbol{\mu}) \langle X_j, u \rangle$.

With Lipschitz property (6.22) and by Lemma 6.11, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\epsilon_j w_1(X_j;\boldsymbol{\mu})\langle X_j,u\rangle\right] \leq \mathbb{E}\left[\frac{1}{n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}[\eta_j(\boldsymbol{\mu})]_k\frac{\sqrt{2r}}{4}\langle X_j,u\rangle\right] \\
=&\mathbb{E}\left[\frac{\sqrt{2}r^{\frac{1}{2}}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=2}^{r}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle+\log(\frac{\pi_k}{\pi_1})\right)\langle X_j,u\rangle\right] \\
\leq&\underbrace{\mathbb{E}\left[\frac{\sqrt{2r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log(\frac{\pi_k}{\pi_1})\right)\langle X_j,u\rangle\right]}_{(D.1)} \\
&+\underbrace{\mathbb{E}\left[\frac{\sqrt{2r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\langle X_j,\boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\langle X_j,u\rangle\right]}_{(D.2)}
\end{aligned} \tag{6.23}$$

To bound (D.1), note that the sum over $k=1,\cdots,r$ can be considered as an inner product of two vectors in $\mathbb{R}^r$. The supremum of $\|\boldsymbol{\mu}\|$ can be bounded as $\max_{\boldsymbol{\mu}\in\mathbb{A}}\|\boldsymbol{\mu}_i\| \leq \|\boldsymbol{\mu}^*_{\max}\|+a \leq \frac{3}{2}R_{\max}$.

$$\begin{aligned}
(D.1) =&\mathbb{E}\left[\frac{\sqrt{2r}}{4}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\begin{pmatrix}\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_1\|^2}{2}+\log(\frac{\pi_1}{\pi_1})\\ \vdots \\ \frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_r\|^2}{2}+\log(\frac{\pi_r}{\pi_1})\end{pmatrix}^T\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\langle X_j,u\rangle\\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jr}\langle X_j,u\rangle\end{pmatrix}\right] \\
\leq& cr(9R^2_{\max}/4+\log(\kappa))\mathbb{E}\left\|\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\langle X_j,u\rangle\\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jr}\langle X_j,u\rangle\end{pmatrix}\right\|
\end{aligned} \tag{6.24}$$

By Lemma 6.6, and $\|u\|=1$, we know $\langle X_j,u\rangle$ is sub-Gaussian with parameter upper bounded by $1+R_{\max}$. So each element of the vector in Equation 6.24 is the average of n independent mean 0 sub-Gaussian random variables with sub-gaussian norm upper bounded by $1+R_{\max}$ (since w.l.o.g we have assumed

that $\sigma = 1$ and $\max_i \|\mu\| \leq R_{\max}$, by Lemma 6.7). Consequently, $\forall k \in [r]$, $\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u_1\rangle\right| \leq c(1+R_{\max})/\sqrt{n}$ for some global constant c [104], and

$$\begin{aligned}(D.1) \leq & cr^{3/2}(9R_{\max}^2/4 + \log(\kappa))(1+R_{\max})\frac{1}{\sqrt{n}} \\ \leq & cr^{3/2}(1+R_{\max})^3 \max\{1, \log(\kappa)\}\frac{1}{\sqrt{n}}\end{aligned}$$

On the other hand, for $(D.2)$, we have

$$\begin{aligned}(D.2) =& \mathbb{E}\left[\frac{\sqrt{2r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^n\sum_{k=1}^r \epsilon_{jk}\langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1\rangle\langle X_j, u\rangle\right] \\ =& \mathbb{E}\left[\frac{\sqrt{2r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{k=1}^r (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T\left(\sum_{j=1}^n \epsilon_{jk}X_jX_j^T\right)u\right] \\ \leq & \sum_{k=1}^r \mathbb{E}\left[\frac{\sqrt{2r}}{4}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_1\|\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_jX_j^T\right\|_{op}\right] \\ \leq & \sum_{k=1}^r \frac{\sqrt{2r}}{2}\|\boldsymbol{\mu}_{\max}\|\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_jX_j^T\right\|_{op}\right]\end{aligned} \tag{6.25}$$

For each $k \in [r]$, the operator norm $\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_jX_j^T\|_{op}$ can be bounded by the same discretization technique with the 1/2-covering of the unit sphere. To be specific, since for any matrix A, $\|A\|_{op} = \sup_{u\in\mathbb{S}^{d-1}}\|Au\|$,

$$\forall u, \exists u_j \ s.t. \ \|Au\| \leq \|Au_j\| + \|A\|_{op}\|u - u_j\| \leq \max_j \|Au_j\| + \frac{1}{2}\|A\|_{op}$$

Taking $\sup_{u\in\mathbb{S}^{d-1}}$ on the left side, we get $\|A\|_{op} \leq 2\max_j\|Au_j\|$. Therefore $\|\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}X_jX_j^T\|_{op} \leq 2\max_\ell \frac{1}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u_\ell\rangle^2$. The square of sub-gaussian random variable $\langle X_j, u_\ell\rangle$ is sub-exponential, from Lemma 5.14 in [104] we know

$$\mathbb{E}\left[\exp\left(\frac{1}{n}\sum_{j=1}^n \epsilon_{jk}\langle X_j, u\rangle^2 t\right)\right] \leq \exp\left(\frac{c_4 t^2(1+R_{\max})^4}{n}\right)$$

With the 1/2-covering number of $\mathcal{S}^{d-1}$ bounded by $\exp(2d)$, we have

$$\mathbb{E}\left[\exp\left(t\cdot\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\|_{op}\right)\right]\leq\exp\left(2d+\frac{c_5t^2(1+R_{\max})^4}{n}\right)$$

Hence, $\forall t>0$,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right]=&\frac{1}{t}\log\left(\exp\left(t\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right]\right)\right)\\ \leq&\frac{1}{t}\log\left(\mathbb{E}\left[\exp\left(t\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right)\right]\right)\\ \leq&\frac{2d}{t}+\frac{ct(1+R_{\max})^4}{n}\end{aligned}$$

Taking $t=c\frac{\sqrt{nd}}{(1+R_{\max})^2}$,

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_jX_j^T\right\|_{op}\right]\leq c\sqrt{\frac{d}{n}}(1+R_{\max})^2$$

Plugging back to Eq. (6.25), and use $\sup_{\boldsymbol{\mu}\in\mathbb{A}}\|\boldsymbol{\mu}\|\leq\sup_k\|\boldsymbol{\mu}_k^*\|+a\leq\frac{3}{2}R_{\max}$, we have

$$(D.2)\leq\frac{cr(1+R_{\max})^3\sqrt{d}}{\sqrt{n}}$$

Plugging the bound back to Eq. (6.23), we have

$$(D)\leq\frac{cr^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}}{\sqrt{n}}$$

Apply Lemma 6.11 on the (E) term in Eq. (6.21), we have

$$
\begin{aligned}
(E) =& \mathbb{E}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}} \frac{1}{n}\sum_{j=1}^{n} \epsilon_j w_i(X_j;\boldsymbol{\mu})\langle \boldsymbol{\mu}_i, u\rangle\right] \\
\leq& \mathbb{E}\left[\frac{\sqrt{2r}}{4n}\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\langle X_j, \boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle+\log(\frac{\pi_k}{\pi_1})\right)\langle \boldsymbol{\mu}_i, u\rangle\right] \\
\leq& \underbrace{\frac{\sqrt{2r}}{4}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log\frac{\pi_k}{\pi_1}\right)\langle \boldsymbol{\mu}_i, u\rangle\right]}_{E.1} \\
&+ \underbrace{\frac{\sqrt{2r}}{4}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\langle X_j, \boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle\langle \boldsymbol{\mu}_i, u\rangle\right]}_{E.2}
\end{aligned}
$$

We will now bound $(E.1)$ and $(E.2)$.

$$
\begin{aligned}
(E.1) &\leq \frac{\sqrt{2r}}{4}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\left(\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_k\|^2}{2}+\log\frac{\pi_k}{\pi_1}\right)\sup_{\boldsymbol{\mu}\in\mathbb{A}}\langle \boldsymbol{\mu}_i, u\rangle\right] \\
&\leq \frac{\sqrt{2r}}{4}R_{\max}\mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\begin{pmatrix}\frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_1\|^2}{2}+\log(\frac{\pi_1}{\pi_1})\\ \vdots \\ \frac{\|\boldsymbol{\mu}_1\|^2}{2}-\frac{\|\boldsymbol{\mu}_r\|^2}{2}+\log(\frac{\pi_r}{\pi_1})\end{pmatrix}^T\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jr}\end{pmatrix}\right] \\
&\leq crR_{\max}(9R_{\max}^2/4+\log\kappa)E_\epsilon\left\|\begin{pmatrix}\frac{1}{n}\sum_{j=1}^{n}\epsilon_{j1}\\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n}\epsilon_{jr}\end{pmatrix}\right\| \qquad (6.26)
\end{aligned}
$$

Note that each element of the vector in Equation 6.26 is the average of n i.i.d mean 0 Radamacher random variables, which are essentially sub-gaussian radnom variables with subgaussian norm upper bounded by 1. Consequently, $\forall k \in [r]$, $\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}\right| \leq c'/\sqrt{n}$ for some global constant c [104], and

$$(E.1) \leq c' r^{3/2} R_{\max}(9R_{\max}^2/4+\log\kappa)/\sqrt{n}$$

As for (E.2), we have

$$
\begin{aligned}
(E.2) &\leq \frac{\sqrt{2r}}{4}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{r}\epsilon_{jk}\langle X_j, \boldsymbol{\mu}_k-\boldsymbol{\mu}_1\rangle \sup_{\boldsymbol{\mu}\in\mathbb{A}}\langle \boldsymbol{\mu}_i, u\rangle\right] \\
&\leq \frac{3\sqrt{2r}}{8}R_{\max}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}\sum_{k=1}^{r}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^T\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right)\right] \\
&\leq \frac{3\sqrt{2r}}{8}R_{\max}\sum_{k=1}^{r}\mathbb{E}_{X,\epsilon}\left[\sup_{\boldsymbol{\mu}\in\mathbb{A}}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1)^T\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right)\right] \\
&\leq \frac{9\sqrt{2r}}{8}R_{\max}^2\sum_{k=1}^{r}\mathbb{E}_{X,\epsilon}\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right\| \qquad (6.27)
\end{aligned}
$$

In Eq (6.27), the vector $\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j$ is the average of n independent mean zero isotropic subgaussian random vectors. Another using of the discretizing technique along with the moment generating function with $t \geq 0$ gives:

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right\| &\leq 2\max_{\ell}\langle\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j, u_\ell\rangle \\
E\left[\exp\left(t\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right\|\right)\right] &\leq \sum_{\ell}E\left[\exp\left(2\frac{t}{n}\sum_{j=1}^{n}\epsilon_{jk}\langle X_j, u_\ell\rangle\right)\right] \\
&\leq \exp\left(2d+\frac{c'(1+R_{\max})^2t^2}{n}\right) \\
E\left\|\frac{1}{n}\sum_{j=1}^{n}\epsilon_{jk}X_j\right\| &\leq \frac{c''+2d+\frac{c'(1+R_{\max})^2t^2}{n}}{t} \qquad \text{Using Jensen's inequality}
\end{aligned}
$$

Taking $t = \Theta\left(\sqrt{nd}/(1+R_{\max})\right)$,

$$(E.2) \leq cr^{3/2}R_{\max}^2(1+R_{\max})\sqrt{d}/\sqrt{n}$$

Thus, combing (E.1) and (E.2) we get:

$$(E) \leq \frac{cr^{3/2}(1+R_{\max})^3\max\{1,\log(\kappa)\}\sqrt{d}}{\sqrt{n}}$$

The final bound follows by combining (D) and (E):

$$R_n(\mathcal{F}) \leq \frac{cr^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1,\log(\kappa)\}}{\sqrt{n}}$$

□

For proving Theorem 2.6 we first recall the following symmetrization lemma in learning theory.

Lemma 6.12 (See e.g. [80])**.** *Let $\mathcal{F}$ be a function class with domain X. Let $\{X_1, X_2, \cdots, X_n\}$ be a set of sample generated by a distribution $\mathbb{P}$ on X. Assume σ_i are i.i.d. Rademacher variables, then*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}(\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(X_i))\right) \leq 2R_n(\mathcal{F})$$

Here $R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right]$ is the Rademacher complexity.

Proof of Theorem 2.7. We will use the notation $X_i^j = (X_i, \cdots, X_j)$ for all sequences, and sequence concatenation is denoted multiplicatively: $x_i^j x_{j+1}^k = x_i^k$. We will also use that fact that X_1^0 is the empty set. The proof will proceed via the Azuma-Hoeffding-McDiarmid method of martingale differences. Defining $V_i = \mathbb{E}[g|X_1^i] - \mathbb{E}[g|X_1^{i-1}]$, we see that $g(X) - E[g(X)] = \sum_i V_i$. We also note that V_i is a function X_1^i. We have,

$$\mathbb{E}[g|X_1^i] = \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P(x_{i+1}^n) g(X_1^i x_{i+1}^n),$$

which along with Jensen's inequality gives:

$$
\begin{aligned}
e^{\lambda V_i} &= e^{\lambda \sum_{x_i', x_{i+1}^n} P(x_{i+1}^n) P(x_i') (g(X_1^{i-1}, x_i, x_{i+1}^n) - g(X_1^{i-1}, x_i', x_{i+1}^n))} \\
&\leq \sum_{x_i', x_{i+1}^n} P(x_{i+1}^n) P(x_i') e^{\lambda (g(X_1^{i-1}, x_i, x_{i+1}^n) - g(X_1^{i-1}, x_i', x_{i+1}^n))} \\
\mathbb{E}\left[e^{\lambda V_i} | X_1^{i-1}\right] &\leq \sum_{x_{i+1}^n} P(x_{i+1}^n) \sum_{x_i, x_i'} P(x_i) P(x_i') e^{\lambda (g(X_1^{i-1}, x_i, x_{i+1}^n) - g(X_1^{i-1}, x_i', x_{i+1}^n))}
\end{aligned}
$$

For fixed $X_1^{i-1} \in \mathcal{X}_1^{i-1}$ and $x_{i+1}^n \in \mathcal{X}_{i+1}^n$, define $F_i : \mathcal{X}_i \to \mathbb{R}$ by $F_i(y) = g(X_1^{i-1} y x_{i+1}^n)$. Let X' denote $X_1^{i-1} x_i' X_{i+1}^n$, which only differ with X on the i-th position. Using the definition of $g(X)$ and denoting by $\tilde{\boldsymbol{\mu}}$ the $\boldsymbol{\mu}$ that achieves the supremum in $g(X_1^{i-1} y x_{i+1}^n)$, we get:

$$
\begin{aligned}
& F_i(y) - F_i(y') \\
=& \sup_{\boldsymbol{\mu} \in \mathbb{A}} \left(\frac{1}{n} \sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}) \langle X_i - \boldsymbol{\mu}_1, u \rangle - \mathbb{E}_X w_1(X; \boldsymbol{\mu}) \langle X - \boldsymbol{\mu}_1, u \rangle \right) \\
& \quad - \sup_{\boldsymbol{\mu} \in \mathbb{A}} \left(\frac{1}{n} \sum_{i=1}^n w_1(X_i'; \boldsymbol{\mu}) \langle X_i' - \boldsymbol{\mu}_1, u \rangle - \mathbb{E}_{X'} w_1(X'; \boldsymbol{\mu}) \langle X' - \boldsymbol{\mu}_1, u \rangle \right) \\
\leq& \left(\frac{1}{n} \sum_{i=1}^n w_1(X_i; \tilde{\boldsymbol{\mu}}) \langle X_i - \tilde{\boldsymbol{\mu}}_1, u \rangle - \mathbb{E}_X w_1(X; \tilde{\boldsymbol{\mu}}) \langle X - \tilde{\boldsymbol{\mu}}_1, u \rangle \right) \\
& \quad - \left(\frac{1}{n} \sum_{i=1}^n w_1(X_i'; \tilde{\boldsymbol{\mu}}) \langle X_i' - \tilde{\boldsymbol{\mu}}_1, u \rangle - \mathbb{E}_X w_1(X'; \tilde{\boldsymbol{\mu}}) \langle X' - \tilde{\boldsymbol{\mu}}_1, u \rangle \right) \\
=& \frac{1}{n} \left(w_1(y; \tilde{\boldsymbol{\mu}}) \langle y - \tilde{\boldsymbol{\mu}}_1, u \rangle - w_1(y'; \tilde{\boldsymbol{\mu}}) \langle y' - \tilde{\boldsymbol{\mu}}_1, u \rangle \right)
\end{aligned}
$$

Take $\bar{\boldsymbol{\mu}}$ as the maximizer for the supremum of X' we get the other side of the

inequality. Hence

$$\begin{aligned}|F_i(y) - F_i(y')| \leq & \frac{1}{n} \sup_{\boldsymbol{\mu} \in \mathbb{A}} \left(|\langle y - \boldsymbol{\mu}, u \rangle| + |\langle y' - \boldsymbol{\mu}, u \rangle| \right) \\ \leq & \frac{1}{n} \left(|\langle y - \boldsymbol{\mu}^*_{Z_y}, u \rangle| + |\langle y' - \boldsymbol{\mu}^*_{Z'_y}, u \rangle| + 4R_{\max} \right) := \rho(y, y')\end{aligned}$$

Note for all t such that $|t| < s$, we have $e^t + e^{-t} \leq e^s + e^{-s}$, we have

$$e^{\lambda(F(y)-F(y'))} + e^{-\lambda(F(y)-F(y'))} \leq e^{\lambda\rho(y,y')} + e^{-\lambda\rho(y,y')}$$

By symmetry, we have:

$$\begin{aligned}\sum_{y,y'} P(y)P(y')e^{\lambda(F(y)-F(y'))} \leq & \frac{1}{2} \left(\mathbb{E}_{y,y'} e^{\lambda\rho(y,y')} + \mathbb{E}_{y,y'} e^{-\lambda\rho(y,y')} \right) \\ = & \mathbb{E}_\epsilon \mathbb{E}_{y,y'} e^{\lambda\epsilon\rho(y,y')} \overset{(i)}{\leq} e^{\lambda^2/n^2} E[e^{4\lambda\epsilon R_{\max}/n}] \\ \overset{(ii)}{\leq} & e^{\lambda^2/n^2 + 8\lambda^2 R^2_{\max}/n^2} \leq e^{\lambda^2(1+3R_{\max})^2/n^2}\end{aligned}$$

where ϵ is a Rademacher random variable independent of y, y'. Note that $\epsilon|\langle y - \boldsymbol{\mu}^*_{Z_y}, u \rangle|$ is identically distributed as a Gaussian random variable with mean zero and variance 1. Also since by construction y and y' are independent, inequality (i) follows using the moment generating function of a Gaussian. Inequality (ii) follows from Hoeffding's Lemma (Eq (3.16) in [48]) since $\epsilon \in [-1, 1]$. Therefore,

$$\mathbb{E}\left[e^{\lambda V_i} | X_1^{i-1}\right] \leq e^{\lambda^2/n^2 + 8\lambda^2 R^2_{\max}/n^2} \leq e^{\lambda^2(1+3R_{\max})^2/n^2} \tag{6.28}$$

Applying standard Markov inequality, we have

$$\begin{aligned}
&P(g(X) - \mathbb{E}g(X) > t) = P\left(\sum_{i=1}^{n} V_i > t\right) \\
\leq& e^{-\lambda t}\mathbb{E}\left[\Pi_{i=1}^{n} e^{\lambda V_i}\right] \\
\overset{(i)}{=}& e^{-\lambda t}\mathbb{E}\left[\mathbb{E}\left[\Pi_{i=1}^{n} e^{\lambda V_i}|X_1^{n-1}\right]\right] \\
=& e^{-\lambda t}\mathbb{E}\left[\Pi_{i=1}^{n-1} e^{\lambda V_i}\mathbb{E}\left[e^{\lambda V_n}|X_1^{n-1}\right]\right] \\
\overset{(ii)}{=}& e^{-\lambda t}\mathbb{E}\left[\Pi_{i=1}^{n}\mathbb{E}\left[e^{\lambda V_i}|X_1^{i-1}\right]\right] \\
\overset{(iii)}{\leq}& \exp\left(-\lambda t + \frac{\lambda^2(1+3R_{\max})^2}{n}\right)
\end{aligned}$$

where step (ii) follows by applying step (i) repeatedly and step (iii) follows by applying Eq (6.28). Optimizing over λ we have $P(g(X) - \mathbb{E}g(X) > t) \leq \exp\left(-\frac{nt^2}{4(1+R_{\max})^2}\right)$. Taking $t = 2(1+3R_{\max})\sqrt{\frac{d\log n}{n}}$, we have

$$P\left(g(X) - \mathbb{E}g(X) > 2(1+3R_{\max})\sqrt{\frac{d\log n}{n}}\right) \leq n^{-d}$$

□

6.4 Initialization

This section provides the number of initializations needed for the condition in Theorem 2.1.

Proposition 6.1. *Let $\pi_i = \frac{1}{M}, \forall i \in [M]$, $R_{\min} = \Omega(\sqrt{d})$, and let a satisfy the conditions in Theorem 2.1. Then with $\frac{\log(1/\delta)}{\sqrt{2\pi M}}\left(\frac{e}{1-e^{-a\sqrt{d}/2}}\right)^M$ initializations, the probability of having at least one good initialization is greater than $1-\delta$.*

The proof follows directly from some combinatorial arguments and Lemma 6.5.

Proof of Proposition 6.1. Define event $\mathcal{E}_{init}(a) = \{\mu_i^0 \in \mathbb{B}_{\mu_i^*}(a),\ \forall i \in [M]\}$. By equal weights assumption, the probability of randomly sampled M points having exactly one from each cluster is $\frac{M!}{M^M}$. By Sterling's formula, we have $M! \geq \sqrt{2\pi M}e^{-M}$. For each center, by Lemma 6.5 we have the probability of it lying in $\mathbb{B}_{\mu_i^*}(a)$ is no less than $1 - e^{-a\sqrt{d}/2}$. Hence

$$P(\mathcal{E}_{init}(a)) \geq \sqrt{2\pi M}\left(\frac{1 - e^{-a\sqrt{d}/2}}{e}\right)^M =: p$$

Now assume the number of initializations is T, in order to satisfy the required property, we need $(1 - P(\mathcal{E}_{init}(a)))^T \leq \delta$. A sufficient condition is

$$T \geq \frac{\log(1/\delta)}{\log(1-p)}$$

Note that $\log(1-x) \geq -x, \forall 0 \leq x \leq 0.5$. Since $p < .5$ for $M \geq 2$, we see that as long as $T \geq \frac{\log(1/\delta)}{\sqrt{2\pi M}}\left(\frac{e}{1-e^{-a\sqrt{d}/2}}\right)^M$, with probability $1-\delta$ we will have a good initialization. □

Remark 6.1. Perhaps not so surprisingly, the above theorem requires a stronger separation condition, i.e. $R_{\min} = \Omega(\sqrt{d})$, whereas all our analysis requires $R_{\min} = \Omega(\sqrt{d_0})$ where $d_0 := \min(d, M)$ can be thought of as effective dimension. This difficulty can be alleviated by using projections schemes similar to those in [8, 63]. We leave this for future work.

Chapter 7

Appendix in SDP-based Kernel Clustering

7.1 Sub-gaussian random vector

In our analysis, we make use of some useful properties of sub-gaussian random variables, which are defined by the following equivalent properties. More discussions on this topic can be found in [104].

Lemma 7.1 ([104]). *The sub-gaussian norm of X is denoted by $\|X\|_{\psi_2}$,*

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}.$$

Every sub-gaussian random variable X satisfies:

(1) $P(|X| > t) \leq \exp(1 - ct^2/\|X\|_{\psi_2}^2)$ for all $t \geq 0$;

(2) $(\mathbb{E}|X|^p)^{1/p} \leq \|X\|_{\psi_2}\sqrt{p}$ for all $p \geq 1$. In particular, $Var(X) \leq 2\|X\|_{\psi_2}^2$.

(3) Consider a finite number of independent centered sub-gaussian random variables X_i. Then $\sum_i X_i$ is also a centered sub-gaussian random variable. Moreover,

$$\|\sum_i X_i\|_{\psi_2}^2 \leq C \sum_i \|X_i\|_{\psi_2}^2$$

We say that a random vector $X \in \mathbb{R}^n$ is sub-gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$.

We will also see the square of sub-gaussian random variables, the following lemma shows it will be sub-exponential. A random variable is sub-exponential if the following equivalent properties hold with parameters $K_i > 0$

differing from each other by at most an absolute constant factor.

$$P(|X| > t) \leq \exp(1 - t/K_1) \quad \text{for all } t \geq 0; \tag{7.1}$$

$$(\mathbb{E}|X|)^{1/p} \leq K_2 p \quad \text{for all } p \geq 1; \tag{7.2}$$

$$\mathbb{E}\exp(X/K_3) \leq e. \tag{7.3}$$

Lemma 7.2 ([104]). *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$$

We have a Bernstein-type inequality for independent sum of sub-exponential random variables.

Lemma 7.3 ([104]). *Let $X_1, \cdots, X_N$ be independent centered sub-exponential random variable, and $M = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

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

where $c > 0$ is an absolute constant.

7.2 Proof of Theorem 2.8

To prove Theorem 2.8, we work with the elementwise expansion, for ease of notation, we slightly abuse K and $\tilde{K}$ to represent $K^{\mathcal{I}\times\mathcal{I}}$ and $\tilde{K}^{\mathcal{I}\times\mathcal{I}}$ in

this proof. We use c to represent any constant that does not depend on the parameters, and its value can change from line to line. For $i \in C_k, j \in C_\ell$, recall that W_i is sub-gaussian random vector with mean 0, covariance $\sigma_k^2 I$ and sub-gaussian norm bounded by b. We have

$$\|Y_i - Y_j\|_2^2 = \|\mu_k - \mu_\ell\|_2^2 + 2\frac{(W_i - W_j)^T}{\sqrt{d}}(\mu_k - \mu_\ell) + \frac{\|W_i - W_j\|_2^2}{d} \tag{7.4}$$

As W_i and W_j are independent, $W_i - W_j$ has mean 0 and covariance $(\sigma_k^2 + \sigma_\ell^2)I$.

Define

$$\beta_{ij} = \|W_i - W_j\|_2^2/d - (\sigma_k^2 + \sigma_\ell^2),$$
$$\alpha_{ij} = (W_i - W_j)'(\mu_k - \mu_\ell)/\sqrt{d}.$$

Hence $\mathbb{E}\beta_{ij} = 0$. By the Lipschitz continuity of f,

$$|K_{ij} - \tilde{K}_{ij}| \leq 2C_0|\beta_{ij} + 2\alpha_{ij}| \tag{7.5}$$

By Lemma 7.1-(3), α_{ij} is also sub-gaussian, with sub-gaussian norm upper bounded by $2bd_{k\ell}^2 C/d$, for some $C > 0$. Then by Lemma 7.1-(1), $\exists C_1 > 0$ s.t.

$$P\left(|\alpha_{ij}| \geq c\sqrt{\frac{\log d}{d}}\right) \leq d^{-C_1 c^2} \tag{7.6}$$

To bound β_{ij}, note each summand in Eq. (7.7) is a squared sub-gaussian random variable, thus is a sub-exponential random variable by Lemma 7.2.

$$\beta_{ij} = \sum_{d=1}^{d}(W_i^{(d)} - W_j^{(d)})^2/d - (\sigma_k^2 + \sigma_\ell^2). \tag{7.7}$$

By Lemma 7.3 with $t = c\sqrt{\frac{\log d}{d}}$, we see that with $a = (1, \ldots, 1)/p$, $\min\left(c^2\frac{t^2}{M^2\|a\|_2^2}, c\frac{t}{M\|a\|_\infty}\right) = \min\left(\frac{c^2\log d}{M^2}, \frac{c\sqrt{d\log d}}{M}\right) \geq c'\log p$ for large enough p. Thus $\exists C_2 > 0$ such that for large enough p,

$$P\left(|\beta_{ij}| \leq c\sqrt{\frac{\log d}{d}}\right) \geq 1 - d^{-C_2c^2} \tag{7.8}$$

By union bound, for some $\rho > 0$, with probability at least $1 - n^2d^{-\rho c^2}$,

$$\sup_{i,j\in\mathcal{I}} |K_{ij} - \tilde{K}_{ij}| \leq c\sqrt{\frac{\log d}{d}}.$$

7.3 Proof of Lemma 2.3

Define a diagonal matrix D where $D_{ii} = f(\sigma_k^2)$, if $i \in C_k$ and 0 if $i \in \mathcal{O}$. Write $\tilde{K}_0 = \tilde{K} - I + D^2$, which is basically replacing the diagonal of $\tilde{K}$ to make it blockwise constant. By the fact $f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2) = f(d_{k\ell}^2)f(\sigma_k^2)f(\sigma_\ell^2)$, $\tilde{K}_0$ has the decomposition $\tilde{K}_0 = DZBZ^TD$ where $B \in \mathbb{R}_{r\times r}$ and $B_{k\ell} = f(d_{k\ell}^2)$. In fact, B is exactly the Gaussian kernel matrix generated by $\{\mu_i\}_{i=1}^r$ centers, and is strictly positive semi-definite when the scale parameter $\eta \neq 0$ and centers are all different. Hence $\tilde{K}_0$ is rank r.

$$\lambda_r(DZBZ^TD) = \lambda_r(B^{1/2}Z^TD^2ZB^{1/2}) = \lambda_r(BZ^TD^2Z)$$

The first equality uses the fact that XX^T and X^TX has the same set of eigenvalues. The second step uses the fact that B is full rank, since all clusters have distinct means. Now B and Z^TD^2Z are both $r \times r$ positive

definite matrices. So the rth eigenvalue is the smallest eigenvalue. Now we use, $\lambda_{\min}(BZ^TD^2Z) \geq \lambda_{\min}(B)\lambda_{\min}(Z^TD^2Z)$ and have

$$\lambda_r(\tilde{K}_0) \geq \lambda_r(Z^TD^2Z)\lambda_r(B) \geq \frac{n}{r}\lambda_{\min}(B) \cdot \min_k \left(f(\sigma_k^2)\right)^2.$$

Then $\lambda_r(\tilde{K}_0) = \Omega(\frac{n}{r})$. On the other hand, $\|I - D^2\|_2 \leq \max_k(1 - f(2\sigma_k^2))$. Let $\lambda_r(\tilde{K}), \lambda_{r+1}(\tilde{K})$ be the r^{th} and $r+1^{th}$ eigenvalue of $\tilde{K}$, by Weyl's inequality,

$$\lambda_r(\tilde{K}) \geq \lambda_r(\tilde{K}_0) - \max_k(1 - f(2\sigma_k^2)) = \Omega(\frac{n}{r}\lambda_{\min}(B))$$

$$\lambda_{r+1}(\tilde{K}) \leq \max_k(1 - f(2\sigma_k^2)) = O(1) \tag{7.9}$$

Putting pieces together,

$$\lambda_r(\tilde{K}) - \lambda_{r+1}(\tilde{K}) \geq \frac{n}{r}\lambda_{\min}(B) \cdot \min_k \left(f(\sigma_k^2)\right)^2 - 2\max_k(1 - f(2\sigma_k^2)) = \Omega\left(\frac{n}{r}\lambda_{\min}(B)\right).$$

7.4 Proof of Lemma 2.4

Proof. First note that $\hat{X}$ is the optimal solution of (SDP-1), so $\langle K, \hat{X}\rangle \geq \langle K, X_0\rangle$. Hence $\langle K - \tilde{K}, \hat{X} - X_0\rangle \geq \langle \tilde{K}, X_0 - \hat{X}\rangle$.

Let $a := \min_k f(2\sigma_k^2)$, $b := \max_{k\neq\ell} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2)$ and $\gamma_{min} := a - b$, we have

$$
\begin{aligned}
&\langle \tilde{K}, X_0 - \hat{X} \rangle \\
&= \sum_k \sum_{i \in \widetilde{C}_k} \left(\sum_{j \in \widetilde{C}_k} f(2\sigma_k^2)(1 - \hat{X}_{ij}) - \sum_{\ell \neq k} \sum_{j \in \widetilde{C}_\ell} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2) \hat{X}_{ij} \right) \\
&\geq \sum_k \sum_{i \in \widetilde{C}_k} \left(a \sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij}) - b \sum_{\ell \neq k} \sum_{j \in \widetilde{C}_\ell} \hat{X}_{ij} \right) \\
&\geq \sum_k \sum_{i \in \widetilde{C}_k} \left(a \sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij}) - b \left(\frac{n}{r} - \sum_{j \in \widetilde{C}_k} \hat{X}_{ij} \right) \right) \\
&\geq \gamma_{\min} \sum_k \sum_{i \in \widetilde{C}_k} \sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij})
\end{aligned}
\tag{7.10}
$$

On the other hand, by the fact that $\hat{X}_{ij} \geq 0$ and row sum is n/r,

$$
\begin{aligned}
\|X_0 - \hat{X}\|_1 &= \sum_k \sum_{i \in \widetilde{C}_k} \left(\sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij}) + \sum_{\ell \neq k} \sum_{j \in \widetilde{C}_\ell} \hat{X}_{ij} \right) \\
&= \sum_k \sum_{i \in \widetilde{C}_k} \left(\sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij}) + \left(n/r - \sum_{j \in \widetilde{C}_k} \hat{X}_{ij} \right) \right) \\
&\leq 2 \sum_k \sum_{i \in \widetilde{C}_k} \sum_{j \in \widetilde{C}_k} (1 - \hat{X}_{ij})
\end{aligned}
\tag{7.11}
$$

Equations (7.10) and (7.11) gives us:

$$
\|X_0 - \hat{X}\|_1 \leq \frac{2}{\gamma_{\min}} \langle \tilde{K}, X_0 - \hat{X} \rangle \leq \frac{2 \langle K - \tilde{K}, \hat{X} - X_0 \rangle}{\gamma_{\min}}
$$

□

7.5 Proof of Theorem 2.9

By Lemma 2.4,

$$\|X_0 - \hat{X}\|_1 \leq \frac{2\langle \tilde{K}, X_0 - \hat{X}\rangle}{\gamma_{\min}} \leq \frac{2\langle K - \tilde{K}, \hat{X} - X_0\rangle}{\gamma_{\min}}$$

Divide the inner product into inlier part and outlier part, and note that $0 < |K_{ij} - \tilde{K}_{ij}| < 1, \forall i, j$. By Theorem 2.8, w.p. at least $1 - n^2 d^{-\rho c^2}$, we have

$$\begin{aligned}
&\langle K - \tilde{K}, \hat{X} - X_0\rangle \\
=&\langle K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}, \hat{X} - X_0\rangle + \langle K^{\mathcal{R}} - \tilde{K}^{\mathcal{R}}, \hat{X} - X_0\rangle \\
\leq&\|\hat{X} - X_0\|_1 \cdot \|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_\infty + \sum_{(i,j)\in\mathcal{R}} (\hat{X}_{ij} - (X_0)_{ij})(K_{ij} - \tilde{K}_{ij}) \\
\leq&\|\hat{X} - X_0\|_1 \cdot \|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_\infty + \sum_{(i,j)\in\mathcal{R}} \hat{X}_{ij}(K_{ij} - \tilde{K}_{ij}) - \sum_{(i,j)\in\mathcal{R}} (X_0)_{ij}(K_{ij} - \tilde{K}_{ij}) \\
\leq&\|\hat{X} - X_0\|_1 \cdot \|K^{\mathcal{I}\times\mathcal{I}} - \tilde{K}^{\mathcal{I}\times\mathcal{I}}\|_\infty + \sum_{(i,j)\in\mathcal{R}} \hat{X}_{ij} + \sum_{(i,j)\in\mathcal{R}} (X_0)_{ij} \\
\leq&C\sqrt{\frac{\log d}{d}}\|X_0 - \hat{X}\|_1 + \frac{4mn}{r}
\end{aligned}$$

Thus,

$$\left(\gamma_{\min} - 2C\sqrt{\frac{\log d}{d}}\right)\|\hat{X} - X_0\|_1 \leq \frac{4mn}{r}$$

When $\sqrt{\frac{\log d}{d}} = o(\gamma_{\min})$, rearranging terms gives

$$\|X_0 - \hat{X}\|_1 \leq \frac{\frac{4mn}{r}}{\gamma_{\min} - C\sqrt{\frac{\log d}{d}}} \tag{7.12}$$

$$\leq \frac{4mn}{r\gamma_{\min}}\left(1 + \frac{C}{\gamma_{\min}}\sqrt{\frac{\log d}{d}}\right) = O\left(\frac{mn}{r\gamma_{\min}}\right) \tag{7.13}$$

7.6 Davis-Kahan Theorem

Theorem 7.1 ([119]). *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_{s-1} - \lambda_s) > 0$, where $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. Let $d_0 := s - r + 1$, and let $V = (v_r, v_{r+1}, \cdots, v_s) \in \mathbb{R}^{d\times d_0}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \cdots, \hat{v}_s) \in \mathbb{R}^{d\times d_0}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$, for $j = r, r+1, \cdots, s$. Then there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d_0\times d_0}$ such that*

$$\|\hat{V}\hat{O} - V\|_F \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_F}{\min(\lambda_{r-1} - \lambda_r, \lambda_{s-1} - \lambda_s)}.$$

7.7 Proof of Theorem 2.10

Proof. Let R be a $n \times n$ matrix with $R(\mathcal{O}, \mathcal{O}) = I$ and zero otherwise, $\hat{V}_r = RV$, $\hat{V_{\mathcal{O}}} = (I - R)V$. $\hat{V_{\mathcal{I}}}^T \hat{V_{\mathcal{I}}} = \hat{V}^T(I - R)\hat{V}$. For any input matrix W, define $\text{loss}_k(W) := \min_{M \text{ has exactly } k \text{ unique rows}} \|W - M\|_F^2$ as the k-means loss of clustering W corresponding to cluster number k. Furthermore, define two feasible sets: $\mathcal{C}_1 = \{M \in \mathbb{R}^{n\times r} : M_{\mathcal{I}} \text{ has exactly } r \text{ unique rows}\}$ and $\mathcal{C}_2 = \{M \in \mathbb{R}^{n\times r} : M_{\mathcal{I}} \text{ has no more than } r - 1 \text{ unique rows}\}$. We want to obtain a condition such that

$$\min_{M\in\mathcal{C}_1} \|\hat{V} - M\|_F^2 < \min_{M\in\mathcal{C}_2} \|\hat{V} - M\|_F^2 \tag{7.14}$$

Intuitively, this condition indicates the k-means loss of inlier nodes assigned to no more than $r - 1$ clusters is strictly larger than the k-means

loss for assigning inliers to exactly r clusters. By optimality, $\min_{M\in\mathcal{C}_1}\|\hat{V}-M\|_F^2 \leq \|\hat{V}-VO\|_F^2$, therefore a sufficient condition of Eq. (7.14) would be $\|\hat{V}-VO\|_F^2 < \min_{M\in\mathcal{C}_2}\|\hat{V}-M\|_F^2$. Now, we will obtain a lower bound on the k-means loss on $\mathcal{C}_2$. In order to do so, we will use [87] to write the k-means loss for any number of clusters k and input matrix W as the following 0-1 SDP problem for any input matrix W.

$$\begin{aligned}\text{loss}_k(W) = \min_X \quad & \text{trace}(WW^T(I-X)),\\ s.t \quad & X\mathbf{1}=\mathbf{1}, X=X^T,\ X\geq 0,\ \text{trace}(X)=k,\ X^2=X.\end{aligned}$$

Note that by relaxing the constraints, we can see that:

$$\text{loss}_k(W) \geq \min_X \text{trace}(WW^T(I-X)), \quad \text{s.t}\ \ X=X^T, X^2=X, \text{trace}(X)=k$$

The right hand side is essentially finding the trailing k eigenvectors of WW^T [87]. Let the singular values of W be $\sigma_1,\ldots,\sigma_r$.

$$\text{loss}_k(W) \geq \sum_{i=k+1}^{r}\sigma_i^2 \tag{7.15}$$

Let $M^* = \arg\min_{M\in\mathcal{C}_2}\|\hat{V}-M\|_F^2$, then

$$\begin{aligned}\min_{M\in\mathcal{C}_2}\|\hat{V}-M\|_F^2 =& \|\hat{V}-M^*\|_F^2\\ =& \|\hat{V}_{\mathcal{I}}-M^*_{\mathcal{I}}\|_F^2 + \|\hat{V}_{\mathcal{O}}-M^*_{\mathcal{O}}\|_F^2\\ \geq& \min_{s\leq r-1}\text{loss}_s(\hat{V}_{\mathcal{I}}) + 0\end{aligned}$$

The last inequality comes from the fact that $M_{\mathcal{J}}^*$ has no more than $r-1$ unique rows since $M^* \in \mathcal{C}_2$. Note that loss_s is non-increasing as s increases. To see this, consider the following procedure. Suppose the solution for $(k-1)$ centroids are $\{c_i\}_{i=1}^{k-1}$, now generate a feasible k centroid solution by keeping $\{c_i\}_{i=1}^{k-1}$ and picking the k^{th} centroid as the point that has largest distance with its corresponding centroid (there will always exist such a point that does not overlap with the existing centroids as long as loss is greater than 0). This consists an upper bound for the k-means loss with k clusters, which is smaller than the k-means loss with $k-1$ clusters.

Therefore without loss of generality, we assume the inliers are assigned $r-1$ clusters and one cluster contains only outliers. By Eq. (7.15) we have

$$\text{loss}_{r-1}(\hat{V}_{\mathcal{J}}) \geq \sigma_r(\hat{V}_{\mathcal{J}})^2 = \lambda_r(\hat{V}_{\mathcal{J}}^T \hat{V}_{\mathcal{J}}) \geq \lambda_r(\hat{V}^T\hat{V}) - \|\hat{V}^T R\hat{V}\| \geq 1 - \|\hat{V}_{\mathcal{O}}\|_F^2$$

Now, $\|\hat{V}_{\mathcal{O}}\|_F \leq \|V_{\mathcal{O}}O\|_F + \|\hat{V}_{\mathcal{O}} - V_{\mathcal{O}}O\|_F$ However, recall that $V = Z\nu$, and since $V^TV = I_r$, $\nu^T\nu = r/nI$. Thus every row of V is of norm $\sqrt{\frac{r}{n}}$. Using $(a+b)^2 \leq 2(a^2+b^2)$, we have:

$$\|\hat{V}_{\mathcal{O}}\|_F^2 \leq 2(\|V_{\mathcal{O}}O\|_F^2 + \|\hat{V}_{\mathcal{O}} - V_{\mathcal{O}}O\|_F^2) \leq 2\left(\frac{mr}{n} + \|\hat{V} - VO\|_F^2\right)$$

Let $u_{\hat{V}}^2$ denote an upper bound on $\|\hat{V} - VO\|_F^2$, then we have:

$$\text{loss}_{r-1}(\hat{V}_{\mathcal{J}}) \geq 1 - 2\left(\frac{mr}{n} + u_{\hat{V}}^2\right)$$

On the other hand, $\text{loss}_r(\hat{V}) \leq \|\hat{V} - VO\|_F^2 \leq u_{\hat{V}}^2$ by optimality. Hence, we use the condition,

$$1 - 2\left(\frac{mr}{n} + u_{\hat{V}}^2\right) \geq u_{\hat{V}}^2 \qquad \Rightarrow \qquad 3u_{\hat{V}}^2 + 2\frac{mr}{n} < 1 \tag{7.16}$$

□

Proof of Corollary 2.1

Proof. By Eq. (2.18), we have for eigenvectors of $\hat{X}$,

$$\|\hat{U} - UO\|_F \leq O_P\left(\sqrt{\frac{mr}{n\gamma_{\min}}}\right)$$

Plug it to Theorem 2.10 we have $u_{\hat{V}} = C\sqrt{\frac{mr}{n\gamma_{\min}}}$, therefore

$$m < \frac{n\gamma_{\min}}{r(C + 2\gamma_{\min})} = \frac{C'n\gamma_{\min}}{r}$$

For K-SVD, by Eq. (2.19), $u_{\hat{V}} = \max\left\{O_P\left(\frac{\sqrt{mn}}{\lambda_r - \lambda_{r+1}}\right), O_P\left(\frac{n\sqrt{\log d/d}}{\lambda_r - \lambda_{r+1}}\right)\right\}$.

We first consider the scenario where $m = O\left(\frac{n\log d}{d}\right)$, now $u_{\hat{V}} = \frac{C_1 n\sqrt{\log p/p}}{\lambda_r - \lambda_{r+1}}$. Plugging this into inequality (7.16), we have

$$m < \frac{n}{2r}\left(1 - \frac{Cn^2 \log d}{d(\lambda_r - \lambda_{r+1})^2}\right)$$

When $\frac{d}{\log d} > 2r + \frac{Cn^2}{(\lambda_r(\tilde{K}) - (\lambda_{r+1}(\tilde{K}))^2}$, we have $\frac{n}{2r}\left(1 - \frac{Cn^2\log d}{d(\lambda_r - \lambda_{r+1})^2}\right) > \frac{n\log d}{d}$, therefore $m = O\left(\frac{n\log d}{d}\right) = O\left(\frac{n}{2r + \frac{Cn^2}{(\lambda_r - \lambda_{r+1})^2}}\right)$.

In the second scenario where $m = \Omega\left(\frac{n\log p}{p}\right)$, we have $u_{\hat{V}} = \frac{C_2\sqrt{mn}}{\lambda_r - \lambda_{r+1}}$. Now (7.16) solves

$$m < \frac{Cn}{\frac{n^2}{(\lambda_r - \lambda_{r+1})^2} + C'r} \tag{7.17}$$

which shares the same formulation as the first condition.

In particular, when all clusters share the same variance, by Lemma 2.3, $\lambda_r - \lambda_{r+1} = \Theta\left(\frac{n\gamma_{\min}}{r}\right)$. Substituting into Eq. (7.17), we have $m < \frac{Cn\gamma_{\min}^2}{r^2}$. $\square$

7.8 Proof of Lemma 2.5

We prove the result for k-means on $\hat{X}$. Let $\hat{U}$ be the top r eigenvectors of $\hat{X}$, $U \in \mathbb{R}^{n\times r}$ be the top r eigenvector of X_0, then by construction, it can be written as $U = \begin{bmatrix} U^{\mathcal{I}} \\ U^{\mathcal{O}} \end{bmatrix}$. Let $\nu \in \mathbb{R}^{r\times r}$ be the population value of the eigenvector corresponding to each cluster, $U = Z\nu$. U is a unit basis so we know $I = U^T U = \nu^T Z^T Z \nu = \frac{n}{r}\nu^T\nu$. So $\nu^T\nu = \frac{r}{n} I_r$.

Define $\mathcal{C} = \{M \in \mathbb{R}^{n\times r} : M \text{ has no more than} r \text{ unique rows}\}$. Then minimizing the k-means objective for $\hat{U}$ is equivalent to

$$\min_{\{m_1,\cdots,m_r\}\subset\mathbb{R}^r} \sum_i \min_g \|\hat{u}_i - m_g\|_2^2 = \min_{M\in\mathcal{C}} \|\hat{U} - M\|_F^2$$

So $C = [c_1, \cdots, c_n] = \arg\min_{M\in\mathcal{C}} \|\hat{U} - M\|_F^2$ and $\|C - \hat{U}\| \le \|Z\nu O - \hat{U}\|$. c_i is the center assigned to point i by running k-means on $\hat{U}$.

When $i, j \in \mathcal{I}, Z_i \neq Z_j$,

$$\|Z_i\nu - Z_j\nu\| = \|(Z_i - Z_j)\nu\| \ge \sqrt{2} \min_{x:\|x\|^2=1} \sqrt{x^T\nu^T\nu x} = \sqrt{\frac{2r}{n}}$$

So

$$\|c_i - Z_j\nu O\| \ge \|Z_i\nu - Z_j\nu\| - \|c_i - Z_i\nu O\| \ge \sqrt{\frac{2r}{n}} - \sqrt{\frac{r}{2n}} = \sqrt{\frac{r}{2n}} \quad (7.18)$$

Therefore when $i, j \in \mathcal{I}$ and $Z_i \neq Z_j$, $\|c_i - Z_i \nu O\| < \sqrt{\frac{r}{2n}} \Rightarrow \|c_i - Z_i \nu O\|_2 < \|c_i - Z_j \nu O\|_2$, which means node i is correctly clustered.

Now we bound the cardinality of $\mathcal{M}$.

$$\begin{aligned}
|\mathcal{M}| &\leq \frac{2n}{r} \sum_{i \in \mathcal{I}} \|c_i - Z_i \nu O\|_F^2 \\
&= \frac{2n}{r} \|C^{\mathcal{I}} - U^{\mathcal{I}} O\|_F^2 \\
&\leq \frac{2n}{r} (\|C^{\mathcal{I}} - \hat{U}^{\mathcal{I}}\|_F + \|\hat{U}^{\mathcal{I}} - U^{\mathcal{I}} O\|_F)^2 \\
\|C^{\mathcal{I}} - \hat{U}^{\mathcal{I}}\|_F^2 &= \|\hat{U} - C\|_F^2 - \|C^{\mathcal{O}} - \hat{U}^{\mathcal{O}}\|_F^2 \\
&\leq \|\hat{U} - C\|_F^2 \leq \|\hat{U} - UO\|_F^2
\end{aligned}$$

Therefore,

$$|\mathcal{M}| \leq \frac{2n}{r} (\|\hat{U} - UO\|_F + \|\hat{U}^{\mathcal{I}} - U^{\mathcal{I}} O\|_F)^2 \leq \frac{8n}{r} \|\hat{U} - UO\|_F^2$$

For k-means procedure on K, note that $\tilde{K}$ is blockwise constant except for the diagonals. It can be shown that the top r eigenvectors of $\tilde{K}$ are also piecewise constant. The rest of the analysis is similar to that of $\hat{X}$.

7.9 Proof of Corollary 2.2

Proof. Denote by d_0 the distance between clusters, $\alpha = f(2\sigma^2)$, $\beta = f(d_0^2 + 2\sigma^2)$, hence $\gamma_{\min} = \alpha - \beta$. Then $\tilde{K}$ has the form $(\alpha - \beta) X_0 + \beta E + (1 - \alpha) I$, and $\lambda_r(\tilde{K}) \geq \gamma_{\min} n / r$, since $\beta E + (1 - \alpha) I$ is positive semidefinite.

On the other hand, from Lemma 2.3 and Eq. (7.9), $\lambda_{r+1}(\tilde{K}) \leq 1 - f(2\sigma^2) \leq 1$. Hence $\lambda_r - \lambda_{r+1} \geq \frac{n}{r} \gamma_{\min} - 1$. By Lemma 2.5 the misclassification

rate of K-SVD becomes:

$$
\begin{aligned}
|\mathcal{M}_{ksvd}| &\leq C\frac{n}{r}\left(\frac{2^{3/2}\|\tilde{K}-K\|_F}{\lambda_r(\tilde{K})-\lambda_{r+1}(\tilde{K})}\right)^2 \\
&\leq C\frac{n}{r}\left(\frac{\max\left\{n\sqrt{\frac{\log d}{d}},\sqrt{mn}\right\}}{\frac{n}{r}\gamma_{\min}}\right)^2 \\
&\leq \max\left(O_P\left(\frac{mr}{\gamma_{\min}^2}\right), O_P\left(\frac{nr\log d/d}{\gamma_{\min}^2}\right)\right)
\end{aligned}
$$

□

Chapter 8

Appendix for Semi-definite Relaxation for Dense and Sparse Stochastic Block Models

In this chapter, we present the detailed proofs for guarantees of semi-definite relaxation for both dense and sparse networks generated by stochastic block model. The proof for dense graphs is in Section 8.1. And the proof for sparse graph can be found in Section 8.3.

8.1 Proofs for dense networks

8.1.1 Proof of Theorem 3.1 and 3.2

Proof of Theorem 3.2. The construction (3.11)-(3.14) together with X_0 is a primal-dual certificate, if (3.6)-(3.9) are satisfied. In view of the fact that both Λ and X are positive semi-definite, $\langle \Lambda, X \rangle = 0$ is equivalent to $\lambda X = 0$. We need to check the following:

(a) $\Lambda X = 0$;

(b) $\Lambda \succeq 0$;

(c) $\Gamma_{uv} \geq 0, \forall u, v$.

Note that span(X)=span($\mathbf{1}_{S_k}$), therefore we only need to show $\Lambda \mathbf{1}_{S_k} = 0, \forall k \in [r]$. Or equivalently $\Lambda_{S_k} \mathbf{1}_{m_k} = 0$ and $\Lambda_{S_k S_\ell} \mathbf{1}_{m_\ell} = 0$. The latter holds by (3.11). For the former, recall that $\alpha_{S_k}^T \mathbf{1}_{m_k} = \frac{1}{m_k} \left(\mathbf{1}_{m_k}^T A_{S_k} \mathbf{1}_{m_k} \right) + \phi_k$.

$$
\begin{aligned}
0 =& \Lambda_{S_k} \mathbf{1}_{m_k} = -A_{S_k} \mathbf{1}_{m_k} + (\mathbf{1}_{m_k} \alpha_{S_k}^T \mathbf{1}_{m_k} + \alpha_{S_k} \mathbf{1}_{m_k}^T \mathbf{1}_{m_k}) + \beta \mathbf{1}_{m_k} \\
=& - A_{S_k} \mathbf{1}_{m_k} + \left(\frac{\mathbf{1}_{m_k}^T A_{S_k} \mathbf{1}_{m_k}}{m_k} + \phi_k \right) \mathbf{1}_{m_k} + A_{S_k} \mathbf{1}_{m_k} + \phi_k \mathbf{1}_{m_k} + \beta \mathbf{1}_{m_k} \\
=& \left(\frac{\mathbf{1}_{m_k}^T A_{S_k} \mathbf{1}_{m_k}}{m_k} \right) \mathbf{1}_{m_k} + 2\phi_k \mathbf{1}_{m_k} + \beta \mathbf{1}_{m_k}
\end{aligned}
$$

The equation holds by taking

$$\phi_k = -\frac{1}{2}\left(\beta + \frac{\mathbf{1}_{m_k}^T A_{S_k} \mathbf{1}_{m_k}}{m_k}\right). \tag{8.1}$$

Positive Semidefiniteness of Λ For (b), since $\text{span}(\mathbf{1}_{S_k}) \subset \ker(\Lambda)$, it suffices to show that for any $u \in \text{span}(\mathbf{1}_{S_k})^\perp$, $u^T \Lambda u \geq \epsilon \|u\|^2$. Consider the decomposition $u = \sum_k u_{S_k}$, where $u_{S_k} := u \circ \mathbf{1}_{S_k}$, and $u_{S_k} \perp \mathbf{1}_{m_k}$.

$$\begin{aligned}
u^T \Lambda u =& \sum_k u_{S_k}^T \Lambda_{S_k} u_{S_k} + \sum_{k \neq \ell} u_{S_k}^T \Lambda_{S_k S_\ell} u_{S_\ell} \\
=& -\sum_k u_{S_k}^T A_{S_k} u_{S_k} + \beta \sum_k u_{S_k}^T u_{S_k} - \sum_{k \neq \ell} u_{S_k}^T A_{S_k S_\ell} u_{S_\ell} \\
=& -\sum_k u_{S_k}^T (A - P)_{S_k} u_{S_k} - \sum_{k \neq \ell} u_{S_k}^T (A - P)_{S_k S_\ell} u_{S_\ell} + \beta \|u\|_2^2 \\
=& -u^T A u + \beta \|u\|_2^2 \geq \epsilon \|u\|^2
\end{aligned}$$

In order to have $\beta \geq \|A - P\|_2$, using Lemma 3.1, we propose the following sufficient condition:

$$\beta = \Omega(\sqrt{n p_{\max}}) \geq \|A - P\|_2 \tag{8.2}$$

Positiveness of Γ For (c), denote $d_i(S_k) = \sum_{j \in S_k} A_{i,j}$, which is the number of edges from node i to cluster k, and $\bar{d}_i(S_k) = \frac{d_i(S_k)}{m_k}$. Define the average degree between two clusters as $\bar{d}(S_k S_\ell) = \frac{\sum_{i \in S_\ell} d_i(S_k)}{m_\ell}$. For $k \neq \ell$, we plug

(3.13) into (3.11) and get

$$\begin{aligned}\Gamma_{S_kS_\ell} =& - A_{S_kS_\ell} + (I - \frac{1}{m_k}E_{m_k})A_{S_kS_\ell}(I - \frac{1}{m_\ell}E_{m_\ell}) + \frac{1}{m_k}\left(A_{S_k}\mathbf{1}_{m_k} + \phi_k\mathbf{1}_{m_k}\right)\mathbf{1}_{m_\ell}^T \\ &+ \mathbf{1}_{m_k}\frac{1}{m_\ell}\left(\mathbf{1}_{m_\ell}^T A_{S_\ell} + \phi_\ell\mathbf{1}_{m_\ell}^T\right) \\ =& - \frac{1}{m_k}E_{m_k}A_{S_kS_\ell} - \frac{1}{m_\ell}A_{S_kS_\ell}E_{m_\ell} + \frac{1}{m_km_\ell}E_{m_k}A_{S_kS_\ell}E_{m_\ell} + \\ &\left(\frac{E_{S_kS_\ell}A_{S_\ell}}{m_\ell} + \frac{A_{S_k}E_{S_kS_\ell}}{m_k}\right) + \left(\frac{\phi_k}{m_k} + \frac{\phi_\ell}{m_\ell}\right)E_{m_k,m_\ell}\end{aligned} \tag{8.3}$$

Therefore for $u \in C_k, v \in C_\ell$, we have

$$\Gamma_{uv} = -\bar{d}_v(S_k) - \bar{d}_u(S_\ell) + \bar{d}(S_kS_\ell) + \bar{d}_v(S_\ell) + \bar{d}_u(S_k) + \frac{\phi_k}{m_k} + \frac{\phi_\ell}{m_\ell} \tag{8.4}$$

Plugging in Eq (8.1), we have $\Gamma_{uv} \geq 0$ equivalent to

$$\begin{aligned}\bar{d}_u(S_k) - \bar{d}_u(S_\ell) + \frac{1}{2}\left(\bar{d}(S_kS_\ell) - \bar{d}(S_kS_k)\right) + \bar{d}_v(S_\ell) - \bar{d}_v(S_k) + \frac{1}{2}\left(\bar{d}(S_kS_\ell) - \bar{d}(S_\ell S_\ell)\right) \\ - \frac{\beta}{2m_\ell} - \frac{\beta}{2m_k} \geq 0\end{aligned} \tag{8.5}$$

By Chernoff bound, we have

$$\begin{aligned}P\left(\bar{d}_u(S_k) \leq B_{kk} - \sqrt{\frac{6B_{kk}\log n}{m_k}}\right) &\leq n^{-3} \\ P\left(\bar{d}_u(S_\ell) \geq B_{k\ell} + \sqrt{\frac{18B_{k\ell}\log n}{m_\ell}}\right) &\leq n^{-3} \\ P\left(\bar{d}(S_kS_k) \geq B_{kk} + \sqrt{\frac{18B_{kk}\log n}{m_k(m_k-1)}}\right) &\leq n^{-3} \\ P\left(\bar{d}(S_kS_\ell) \leq B_{k\ell} - \sqrt{\frac{6B_{k\ell}\log n}{m_km_\ell}}\right) &\leq n^{-3}\end{aligned}$$

Apply union bound we have,

$$P\left(\bar{d}_u(S_k) - \bar{d}_u(S_\ell) + \frac{1}{2}\left(\bar{d}(S_kS_\ell) - \bar{d}(S_kS_k)\right) + \bar{d}_v(S_\ell) - \bar{d}_v(S_k) + \frac{1}{2}\left(\bar{d}(S_kS_\ell) - \bar{d}(S_\ell S_\ell)\right) \leq \right.$$
$$\frac{1}{2}(B_{kk} - B_{k\ell}) + \frac{1}{2}(B_{\ell\ell} - B_{k\ell}) - \sqrt{6\log n}\left(\sqrt{\frac{B_{kk}}{m_k}} + \sqrt{\frac{p_\ell}{m_\ell}}\right)$$
$$\left. -\sqrt{18B_{k\ell}\log n\left(\frac{1}{m_k} + \frac{1}{m_\ell}\right)}\right) \leq 4n^{-3}$$

We then apply union bound over all pairs of nodes and clusters, and combined with Eq. (3.15), $\Gamma_{uv} \geq 0$ for all pairs of (u, v) if

$$\frac{1}{2}(B_{kk} - B_{k\ell}) + \frac{1}{2}(B_{\ell\ell} - B_{k\ell}) - \sqrt{6\log n}\left(\sqrt{\frac{B_{kk}}{m_k}} + \sqrt{\frac{B_{\ell\ell}}{m_\ell}}\right)$$
$$- \sqrt{18B_{k\ell}\log n\left(\frac{1}{m_k} + \frac{1}{m_\ell}\right)} - c\frac{\sqrt{np_{\max}}}{m_{\min}} \geq 0$$

The proof follows by relaxing $B_{kk} - B_{k\ell}$ with the minimum over all clusters. □

For the proof of Theorem 3.1, we use the same dual certificate construction Eq. (3.11)-(3.14), with $\beta = \lambda$. The existence of the primal-dual certificate is guaranteed by the proof of Theorem 3.2.

8.1.2 Proof of Proposition 3.1

Proof. When $\lambda \geq \|A\|_{op}$, $\tilde{A} = A - \lambda I \preceq 0$. From the constraint we know that $X \succeq 0$, and has at least one eigenvalue 1 with eigenvector $\mathbf{1}/\sqrt{n}$. Consider an eigen-decomposition $X = \frac{1}{n}\mathbf{1}\mathbf{1}^T + \sum_{i=2}^n s_i u_i u_i^T$ where $s_i \geq 0$. Then the objective is

$$\langle \tilde{A}, X\rangle = \mathbf{1}^T\tilde{A}\mathbf{1}/n + \sum_i s_i u_i^T \tilde{A} u_i$$

Note that $s_i \geq 0$ and $\tilde{A} \preceq 0$, so the above objective is maximized when $s_i = 0, \forall i \geq 2$. Therefore $X^* = \mathbf{1}\mathbf{1}^T/n$. □

8.2 Proof of Lemma 3.2

We start with the following lemma, whose proof can be found in [79].

Lemma 8.1. *For any X that satisfies $X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}$, we have $\|X\|_F^2 \leq trace(X)$.*

Proof of Lemma 3.2. Note that both X_0 and X_M are in the feasible set $\mathcal{F}$, by optimality, we have $\langle M, X_M \rangle \geq \langle M, X_0 \rangle$. We construct Q as stated in the lemma to obtain: $\langle Q, X_M - X_0 \rangle$, $\langle M - Q, X_M - X_0 \rangle \geq \langle Q, X_0 - X_M \rangle$. Note that Q is constant on diagonal blocks and upper bounded by q_k on off-diagonal blocks, with respect to the clustering of nodes. Using the fact that $|C_k| = m_k$, we have:

$$\begin{aligned}
\langle M, X_0 - X_M \rangle &= \sum_k \sum_{i \in C_k} \left(\beta_k^{(in)} \sum_{j \in C_k} \left(\frac{1}{m_k} - \hat{X}_{ij} \right) + \sum_{\ell \neq k} \sum_{j \in C_\ell} Q_{ij}(0 - (X_M)_{ij}) \right) \\
&\geq \sum_k \sum_{i \in C_k} \left(\beta_k^{(in)} \sum_{j \in C_k} \left(\frac{1}{m_k} - (X_M)_{ij} \right) - \beta_k^{(out)} \sum_{\ell \neq k} \sum_{j \in C_\ell} (X_M)_{ij} \right) \\
&= \sum_k \sum_{i \in C_k} \left(\beta_k^{(in)} \left(1 - \sum_{j \in C_k} (X_M)_{ij} \right) - \beta_k^{(out)} \left(1 - \sum_{j \in C_k} (X_M)_{ij} \right) \right) \\
&= \sum_k \sum_{i \in C_k} (\beta_k^{(in)} - \beta_k^{(out)}) \left(1 - \sum_{j \in C_k} (X_M)_{ij} \right) \geq \min_k (\beta_k^{(in)} - \beta_k^{(out)}) \sum_k \sum_{i \in C_k} \left(1 - \sum_{j \in C_k} (X_M)_{ij} \right)
\end{aligned}$$

The third line and last inequality uses the constraint that $\sum_j \hat{X}_{ij} = 1$,

and $1-\sum_{j\in C_k}\hat{X}_{ij}\geq 1-\sum_j \hat{X}_{ij}=0$. On the other hand,

$$\|X_M-X_0\|_F^2=\|X_M\|_F^2-\|X_0\|_F^2+2\langle X_0-X_M,X_0\rangle$$

By Lemma 8.1, and the fact that $\|X_0\|_F^2=r$, we have $\|X_M\|_F^2-\|X_0\|_F^2\leq \text{trace}(X_M)-r=0$. Since $\min_k(\beta_k^{(in)}-\beta_k^{(out)})\geq 0$,

$$\|X_M-X_0\|_F^2\leq 2\langle X_0-X_M,X_0\rangle=2\sum_k\sum_{i\in C_k}\sum_{j\in C_k}\frac{1}{m_k}\left(\frac{1}{m_k}-(X_M)_{ij}\right)$$

$$=2\sum_k\sum_{i\in C_k}\frac{1}{m_k}\left(1-\sum_{j\in C_k}(X_M)_{ij}\right)\leq\frac{2}{m_{\min}}\sum_k\sum_{i\in C_k}\left(1-\sum_{j\in C_k}(X_M)_{ij}\right)$$

$$\leq\frac{2}{m_{\min}\min_k(\beta_k^{(in)}-\beta_k^{(out)})}\langle Q,X_0-X_M\rangle\leq\frac{2}{m_{\min}\min_k(\beta_k^{(in)}-\beta_k^{(out)})}\langle M-Q,X_M-X_0\rangle$$

□

8.3 Analysis of sparse graph

We first introduce the following result on sparse graph with Grothendieck's inequality by [43].

Lemma 8.2 ([43]). *Let $\mathcal{M}_G^+=\{X:X\succeq 0,diag(X)\preceq I_n\}$, $A=(a_{ij})\in\mathbb{R}^{n\times n}$ be a symmetric matrix whose diagonal entries equal 0, and entries above the diagonal are independent random variables satisfying $0\leq a_{ij}\leq 1$. Let $P=E[A|Z]$. Assume that $\bar{p}:=\frac{2}{n(n-1)}\sum_{i<j}Var(a_{ij})\geq\frac{9}{n}$. Then, with probability at least $1-e^3 5^{-n}$, we have $\max_{X\in\mathcal{M}_G^+}|\langle A-P,X\rangle|\leq K_G\|A-P\|_{\ell_\infty\to\ell_1}\leq 3K_G\bar{p}^{1/2}n^{3/2}$, where K_G is the Grothendieck's constant, and its best know upper bound is 1.783.*

Proof of Proposition 3.2. Notice that A and $P := E[A|Z]$ has zero diagonals. Therefore,

$$\begin{aligned}\langle P - Q, X_A - X_0\rangle =& \sum_k \sum_{i\in C_k} p_k \left(\frac{1}{m_k} - (X_A)_{ii}\right) \\ \leq& \sum_k p_k - p_{\min}\text{trace}(X_A) \leq r(p_{\max} - p_{\min})\end{aligned}, \tag{8.6}$$

where $p_{\max} = \max_k p_k$ and $p_{\min} = \min_k p_k$. Thus by Lemma 3.2 and Eq (8.6),

$$\|X_A - X_0\|_F^2 \leq \frac{2}{m_{\min}\min_k(p_k - q_k)}\left(\langle A - P, X_A - X_0\rangle + r(p_{\max} - p_{\min})\right)$$

In sparse regime, both $m_{\min}X_0$ and $m_{\min}X_A$ belong to the set $\mathcal{M}_G^+$. Let $g = n\bar{p} \geq 9$, applying Lemma 8.2 we get with probability at least $1 - e^3 5^{-n}$,

$$\|X_A - X_0\|_F^2 \leq \frac{22\sqrt{n^3\bar{p}}}{m_{\min}^2\min_k(p_k - q_k)} + \frac{2r(p_{\max} - p_{\min})}{m_{\min}\min_k(p_k - q_k)}$$

Substituting $p_k = a_k/n, q_k = b_k/n$, and using the fact that

$$\frac{2r(p_{\max} - p_{\min})}{m_{\min}\min_k(p_k - q_k)} = \frac{2rm_{\min}(p_{\max} - p_{\min})}{m_{\min}^2\min_k(p_k - q_k)} \leq \frac{2\max_k a_k}{m_{\min}^2\min_k(p_k - q_k)} = o(\sqrt{n^3\bar{p}}),$$

Recall that $\alpha := m_{\max}/m_{\min}$, we get with probability tending to 1,

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq \frac{23n^2\sqrt{g}}{rm_{\min}^2\min_k(a_k - b_k)} \leq \frac{23\alpha^2 r\sqrt{g}}{\min_k(a_k - b_k)}.$$

□

Chapter 9

Appendix for Covariate Regularized Community Detection

9.1 Proof of Lemma 8.1

Proof of Lemma 8.1. We first show that for all such X, the eigenvalues of X are in $[0, 1]$. Let v_i be the eigenvector of X corresponding to the i^{th} largest eigenvalue λ_i. Since X is positive semi-definite, $\lambda_i \geq 0, \forall i$. Without loss of generality, let $i^* = \arg\max_i |v_1(i)|$, i.e. be the index of the entry with the largest absolute value of v_1. Since $Xv_1 = \lambda_1 v_1$, and $\sum_j X_{ij} = 1, X_{ij} \geq 0$, we have:

$$|\lambda_1 v_1(i^*)| = |\sum_j X_{i^*j} v_1(j)| \leq \sum_j X_{i^*j} |v_1(j)| \leq |v_1(i^*)|.$$

Therefore $|\lambda_1| \leq 1$.

$$\|X\|_F^2 = \sum_i \lambda_i^2 \leq \sum_i \lambda_i = \text{trace}(X)$$

□

9.2 Proof of Proposition 4.1

Proof of Proposition 4.1. Recall that by definition, for $i \in C_k$, $Y_i - \mu_k$ is sub-gaussian random vector with sub-gaussian norm ψ_k. Using the following concentration inequality from [52] for sub-gaussian random vectors, we have:

$$\text{For } i \in C_k,\ P(\|Y_i - \mu_k\|_2^2 > \psi_k^2(d + 2\sqrt{td} + 2t)) \leq e^{-t}$$

We take $t = c_k^2 d$ for $c_k \geq 1$. Since $1 + 2c_k + 2c_k^2 \leq 5c_k^2$ for $c_k \geq 1$, we get $P(\|X - \mathbb{E}X\|^2 \leq 5c_k^2 \psi_k^2 d) \geq 1 - \exp(-c_k^2 d)$. Let $\Delta_k = \sqrt{5} c_k \psi_k \sqrt{d}$, we can

divide the nodes into "good nodes" (those close to their population mean) $\mathcal{S}_k$ and the rest as follows:

$$\mathcal{S}_k = \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}, \qquad \mathcal{S} = \cup_{k=1}^{r} \mathcal{S}_k \tag{9.1}$$

Let $m_c^{(k)} = m_k - |\mathcal{S}_k|$. We want to bound $m_c^{(k)}$ with high probability. Note that $m_c^{(k)} = \sum_{i \in C_k} \mathbf{1}(\|Y_i - \mu_k\| \geq \Delta_k)$ is a sum of i.i.d random variables. Therefore, using the Hoeffding bound we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq m_k \delta\right) \leq \exp(-2m_k\delta^2)$$

Using $\delta = \sqrt{\log m_k / 2m_k}$, we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq \sqrt{m_k \log m_k / 2}\right) \leq \frac{1}{m_k}$$

Since $P(i \notin \mathcal{S}_k) \leq \exp(-c_k^2 d)$, we have:

$$P\left(m_c^{(k)} \geq m_k \exp(-c_k^2 d) + \sqrt{m_k \log m_k / 2})\right) \leq \frac{1}{m_k}$$

Finally, using union bound over all clusters we get:

$$P\left(m_c \geq \sum_k m_k e^{-c_k^2 d} + \sum_k \sqrt{m_k \log m_k / 2}\right) \leq \sum_k \frac{1}{m_k} \tag{9.2}$$

Now define

$$(K_I)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i,j \in C_k \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \neq \ell \end{cases} \tag{9.3}$$

By Lemma 3.2, all diagonal blocks are blockwise constant and the off-diagonal blocks are upper bounded by $f(d_{k\ell} - \Delta_k - \Delta_\ell)$. Let $\nu_k = f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$, and $\gamma = \min_k \nu_k$. If $\nu_k \geq 0$, we have

$$\|X_K - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma} \langle K - K_I, X_K - X_0 \rangle$$

Apply Grothendieck's inequality,

$$\|X_K - X_0\|_F^2 \le \frac{2K_G}{m_{\min}^2\gamma}\|K - K_I\|_{\ell_\infty\to\ell_1} \tag{9.4}$$

Now it remains to bound the $\ell_\infty \to \ell_1$ norm of $K - K_I$. Note that if $i \in S_k, j \in S_\ell, k \neq \ell$, then by a simple use of triangle inequality we have $K_{ij} \le f(d_{k\ell} - \Delta_k - \Delta_\ell)$, so $K_{ij} = (K_I)_{ij}$; and if $i, j \in S_k$, then $K_{ij} \ge f(2\Delta_k)$.

$$\begin{aligned}
\|K - K_I\|_{\ell_\infty\to\ell_1} &= \max_{x,y\in\{\pm\}^n} \sum_{i,j} x_i y_j \left(K_{ij} - (K_I)_{ij}\right) \\
&\le \max_{x,y\in\{\pm\}^n} \sum_{i,j\in\mathcal{S}} x_i y_j \left(K_{ij} - (K_I)_{ij}\right) + \max_{x,y\in\{\pm\}^n} \sum_{i\notin\mathcal{S}\cup j\notin\mathcal{S}} x_i y_j \left(K_{ij} - (K_I)_{ij}\right) \\
&\overset{(i)}{\le} \max_{x,y\in\{\pm\}^n} \sum_{i,j\in\mathcal{S}} x_i y_j \left(K_{ij} - (K_I)_{ij}\right) + 2m_c n \\
&\overset{(ii)}{=} \max_{x,y\in\{\pm\}^n} \sum_k \sum_{i,j\in\mathcal{S}_k} x_i y_j \left(K_{ij} - f(2\Delta_k)\right) + 2m_c n \\
&\le \sum_k m_k^2(1 - f(2\Delta_k)) + 2m_c n
\end{aligned} \tag{9.5}$$

where (i) is due to $|K_{ij} - (K_I)_{ij}| \le 1$, and (ii) comes from the definition of K_I. Now Eq 9.4 follows as

$$\begin{aligned}
\|X_K - X_0\|_F^2 \le& \frac{4K_G\left(\sum_k m_k^2(1 - f(2\Delta_k)) + 2m_c n\right)}{m_{\min}^2\gamma} \\
=& \frac{4K_G}{m_{\min}^2}\sum_k \left(m_k^2\frac{1 - f(2\Delta_k)}{\gamma} + 2m_k n e^{-c_k^2 d}/\gamma\right) + \frac{\sqrt{2}K_G n}{m_{\min}^2\gamma}\sum_k \sqrt{m_k \log m_k}
\end{aligned} \tag{9.6}$$

Recall that $f(x) = \exp(-\eta x^2)$, and $\gamma = \min_k \{f(2\Delta_k) - \max_{\ell\neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)\}$. For simplicity, we assume $c_k = c_0$. We take $c_0 = \sqrt{\log\left(\frac{d_{\min}^2}{\psi_{\max}^2 d}\right)/d}$ and the scale

parameter $\eta = \frac{\phi}{20c_0^2\psi_{\max}^2 d}$, for some $\phi > 0$, which will be chosen later. Furthermore, we also define

$$\xi = \frac{d_{\min}}{2\sqrt{5}c_0\psi_{\max}\sqrt{d}} - 1. \tag{9.7}$$

If $\xi > 1$, then $d_{min} > 4\sqrt{5}c_0\psi_{max}\sqrt{d}$, and hence $\gamma > 0$. Also, since $\eta(d_{\min} - 2\sqrt{5}c_0\psi_{\max}\sqrt{d})^2 = \phi\xi^2, \forall k, \ell \in [r]$, if $d_{\min} := \min_{k\ell} d_{k\ell} > 4\sqrt{5}c_0\psi_{\max}\sqrt{d}$, then

$$\gamma \geq f(2\sqrt{5}c_0\psi_{\max}\sqrt{d}) - f(d_{\min} - 2\sqrt{5}c_0\psi_{\max}\sqrt{d}) = \exp(-\phi) - \exp(-\phi\xi^2).$$

and

$$1 - f(2\Delta_k) \leq 1 - f(2\sqrt{5}c_0\psi_{\max}\sqrt{d}) = 1 - \exp(\phi).$$

Recall $\alpha = \frac{m_{\max}}{m_{\min}}$,

$$\begin{aligned}
&\|X_K - X_0\|_F^2 \qquad (9.8)\\
\leq& 4K_G r\alpha^2 \cdot \frac{1 - f(2\sqrt{5}c_0\psi_{\max}\sqrt{d}) + 2r\exp(-c_0^2 d)}{\gamma} + \frac{2\sqrt{2}K_G m_{\max} r^2\sqrt{m_{\max}\log m_{\max}}}{\gamma m_{\min}^2}\\
\leq& \frac{4K_G r\alpha^2}{\gamma}\left(1 - \exp(-\phi) + \frac{2r\psi_{\max}^2\sqrt{d}}{d_{\min}^2} + r\sqrt{\log m_{\max}/2m_{\max}}\right)\\
\leq& 4K_G r\alpha^2\left(\underbrace{\frac{(1 - \exp(-\phi) + 2r\psi_{\max}^2 d/d_{\min}^2}{\exp(-\phi) - \exp(-\phi\xi^2)}}_{A} + \underbrace{\frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)}}_{B}\right) \qquad (9.9)
\end{aligned}$$

We will first bound part (A).

$$(A) = \frac{\exp(\phi) - 1 + \exp(\phi)\frac{2r\psi_{\max}^2 d}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \overset{(i)}{\leq} \frac{\phi + \frac{\phi^2}{2}\exp(\phi) + \exp(\phi)\frac{2r\psi_{\max}^2 d}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \tag{9.10}$$

where (i) uses the Mean value theorem: for $e^x - 1 \leq x + e^y x^2/2$ for $y \in [0, x]$. If $\frac{d_{\min}}{\psi_{\max}\sqrt{d}} > \max\left\{1, \frac{180}{d}\right\}$, using the fact that $\log x \leq \sqrt{x}$, we have:

$$\frac{d_{\min}^2}{\psi_{\max}^2 d} > \frac{180}{d^2}\frac{d_{\min}}{\psi_{\max}} > \frac{180}{d}\log\left(\frac{d_{\min}^2}{\psi_{\max}^2 d}\right) = 180c_0^2.$$

Using Eq 9.7, we see that $\xi > \frac{\sqrt{180}}{2\sqrt{5}} - 1 = 2$, and hence $\gamma > 0$. Now we pick $\phi = \frac{\log \xi}{\xi^2}$.

Now we will use this to obtain a lower bound on $1 - \exp(\phi - \phi\xi^2)$. Since $\xi \geq 2$, we have $\xi^2/4 \geq 1$. Hence

$$\begin{aligned}
1 - \exp(\phi - \phi\xi^2) &\geq 1 - \exp(\phi\xi^2/4 - \phi\xi^2) \\
&= 1 - \exp(-\phi 3\xi^2/4) = 1 - \exp(-3\log\xi/4) = 1 - \xi^{-3/4} \\
&\geq 1 - 2^{-3/4} = .4
\end{aligned}$$

Using the fact that the function $\frac{\log x}{x^2}$ is monotonically decreasing when $x > 2$, we see that $\phi < \log 2/2^2$ and $\exp(\phi) \leq 1.2$. Furthermore,

$$\gamma \geq \exp(-\phi)(1 - \exp(\phi(1 - \xi^2))) \geq .3 \tag{9.11}$$

Now Eq. (9.10) yields:

$$\begin{aligned}
(A) &\leq \frac{\phi + 1.2\left(\frac{\phi^2}{2} + \frac{2r\psi_{\max}^2 d}{d_{\min}^2}\right)}{.4} \leq \frac{c\log\xi}{\xi^2} + \frac{3r\psi_{\max}^2 d}{d_{\min}^2} \\
&\overset{(ii)}{\leq} \frac{c'\log(\xi+1)}{(\xi+1)^2} + \frac{3r\psi_{\max}^2 d}{d_{\min}^2} \leq c''\frac{\psi_{\max}^2 d}{d_{\min}^2}\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right) + \frac{3r\psi_{\max}^2 d}{d_{\min}^2},
\end{aligned}$$

for some constant c. To get (ii), note that

$$\frac{\log\xi}{\xi^2} \leq \frac{\log(\xi+1)}{\xi^2} \leq \frac{2.25\log(\xi+1)}{(\xi+1)^2}, \forall \xi > 2$$

Finally, we bound (B) in Eq 9.9 using Eq 9.11.

$$(B) = \frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)} \leq c_1 r\sqrt{\frac{\log m_{\max}}{m_{\max}}}$$

for some constant $c_1 > 0$. Putting pieces together, we have

$$\frac{\|X_K - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 \max\left(\frac{\psi_{\max}^2 d}{d_{\min}^2}\max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}\sqrt{d}}\right), r\right\}, r\sqrt{\frac{\log m_{\max}}{m_{\max}}}\right)$$

□

9.2.1 Analysis for $X_{A+\lambda K}$

Proof of Theorem 4.1. Let K_I be defined as in Eq (9.3). Let $\gamma = \min_k(p_k - q_k + \lambda(f(2\Delta_k) - \max_{\ell\neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)))$. When $\gamma \geq 0$, Lemma 3.2 with $Q = ZBZ^T + \lambda K_I$, we have

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma}\left(\langle A - P, X_{A+\lambda K} - X_0\rangle + r(p_{\max} - p_{\min}) + \lambda\langle K - K_I, X_{A+\lambda K} - X_0\rangle\right)$$

Now by Grothendieck's inequality on both $\langle A - P, X_{A+\lambda K} - X_0\rangle$ and $\langle K - K_I, X_{A+\lambda K} - X_0\rangle$, one gets,

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq \frac{2K_G}{m_{\min}^2\gamma}\left(2\|A - P\|_{\ell_\infty\to\ell_1} + r(p_{\max} - p_{\min}) + 2\lambda\|K - K_I\|_{\ell_\infty\to\ell_1}\right)$$

By Lemma 8.2 and Eq (9.5),

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq \frac{4K_G}{m_{\min}^2\gamma}\left(6\sqrt{n^3\bar{p}} + \lambda\left(2m_c n + \sum_k m_k^2(1 - f(2\Delta_k))\right)\right)$$

Recall that for the sparse graph, $p_k = a_k/n, q_k = b_k/n, g = \bar{p}/n$. Using $\lambda = \ell/n$, $m_k = n\pi_k$, $m_{\min} = n\pi_{\min}$, and $\pi_0 := \sum_k(m_k\exp(-\Delta_k^2/(5\psi_k^2)) +$

$\sqrt{m_k \log m_k/2})/n$ in conjunction with Eq (9.2), we get with probability tending to 1,

$$\|X_{A+\lambda K} - X_0\|_F^2 \leq 4K_G \frac{6\sqrt{g} + \ell\left(2\pi_0 + \sum_k \pi_k^2(1 - f(2\Delta_k))\right)}{\pi_{\min}^2 \min_k(a_k - b_k + \ell\nu_k)}$$

□

9.3 Analysis of covariate clustering when $d \gg r$

Before proving Lemma 4.1, we clearly state our assumptions and other useful lemmas.

Assumption 9.1. *We assume that M is of rank $r-1$, i.e. the means are not collinear, or linearly dependent, other than the fact that they are centered.*

Lemma 9.1. *Let $M = \sum_k \pi_k \mu_k \mu_k^T$ and S be the covariance matrix of n data points from a sub-gaussian mixture, then $S = M + \sum_i \pi_i \sigma_i^2 I_d$. Let $\hat{S}$ be the sample covariance matrix $\hat{S} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^T}{n}$. We have $\|\hat{S} - S\| \leq C\sqrt{\frac{d \log n}{n}}$ for some constant C with probability bigger than $1 - O(n^{-d})$.*

This is a direct consequence of Corollary 5.50 from [104]. The main ingredient of the proof is provided below.

Lemma 9.2. *Let U_{r-1} be the top $r-1$ eigenvectors of $\hat{S}$ estimated using P_1, and λ be the smallest positive eigenvalue of M. For any vector v in the span of $\{\mu_i\}_{i=1}^r$, as long as $\lambda > 5\left(\psi_{\max}^2 + C\sqrt{\frac{d \log^2 n}{n}}\right)$ we have $\|U_{r-1}^T v\| \geq \|v\|/2$ with probability at least $1 - \tilde{O}(n^{-d})$.*

Proof. Take $n_1 = \frac{n}{\log n}$ and v to be a vector in the span of $\{\mu_i\}_{i=1}^r$. By definition, we have $\|Mv\| \geq \lambda\|v\|$. Let $R = \hat{S} - S$. Denote $\bar{\sigma}^2 = \sum_i \pi_i \sigma_i^2$, by Lemma 9.1, $S = M + \bar{\sigma}^2 I_d$. We also know that $\bar{\sigma}^2 \leq \sigma_{\max}^2 \leq \psi_{\max}^2$ by the property of sub-gaussian distributions. Since S is estimated from P_1 with n_1 points, applying Lemma 9.1 with $n = n_1$ we get $\|R\| \leq \epsilon = C\sqrt{\frac{d \log n_1}{n_1}}$. By Weyl's inequality, $\|\hat{S}v\| = \|(M + R + \sum_i \sigma_i^2 I_d)v\| \geq (\lambda - \sigma_{\max}^2 - \epsilon)\|v\|$. Let $U_{r:d}$ be the eigenspace orthogonal to U_{r-1}.

Assume the contradiction that $\|U_{r-1}^T v\| < \|v\|/2$. Then there has to be a unit d dimensional vector $u \in \text{span}(U_{r:d})$, such that $|u^T v| > \|v\|/2$. On one hand, if we write $u = c\frac{v}{\|v\|} + \sqrt{1-c^2}v^\perp$, for $|c| > 1/2$ and some unit vector $v^\perp$ orthogonal to v, we have $\|\hat{S}u\| \geq \frac{\lambda - \sigma_{\max}^2 - \epsilon}{2} - \sqrt{1-c^2}\|\hat{S}v^\perp\|$. Note $\|\hat{S}v^\perp\| = \|(M + R + \bar{\sigma}^2 I_d)v^\perp\|$. Since $v^\perp$ is orthogonal to the span of M, $\|\hat{S}v^\perp\| \leq (\sigma_{\max}^2 + \epsilon)$. Hence

$$\|\hat{S}u\| \geq \frac{\lambda - 3(\sigma_{\max}^2 + \epsilon)}{2}. \tag{9.12}$$

On the other hand, since $u \in \text{span}(U_{r:d})$, by Weyl's inequality, $\|\hat{S}u\| \leq |\lambda_k(\hat{S})| \leq \sigma_{\max}^2 + \epsilon$. This contradicts with Eq. (9.12) since we assume $\lambda > 5(\psi_{\max}^2 + \epsilon) \geq 5(\sigma_{\max}^2 + \epsilon)$. The result is proven by contradiction. □

Remark 9.1. Note that the result can be generalized to non-spherical case as long as the largest eigenvalue of covariance matrix for each cluster is bounded.

We are now ready to prove Lemma 4.1.

Proof of Lemma 4.1. Recall that $Y_i' = U_{r-1}^T Y_i$ where U_{r-1} and Y_i are from two different partitions and hence independent. Let $Z_i \in [r]$ denote that latent variable associated with i. Thus, $E[Y_i'|Z_i = a, P_2] = U_{r-1}^T E[Y_i|Z_i = a] = U_{r-1}^T \mu_a$. Thus the means of the new mixture are $\mu_a' := U_{r-1}^T \mu_a$ and the covariance matrix is isotropic, i.e. $E[(Y_i' - \mu_a')(Y_i' - \mu_a')^T | P_2, Z_i = a] = \sigma_a^2 I_{r-1}$. Furthermore, using Lemma 9.2 we have $\min_{k \neq \ell} \|\mu_k' - \mu_\ell'\| = \min_{k \neq \ell} \|U_{r-1}^T(\mu_k - \mu_\ell)\| \geq \|d_{\min}\|/2$. Since this requires an application of Lemma 9.2 to each of the vectors $\mu_k - \mu_\ell$, $k, \ell \in [r]$, the success probability is at least $1 - \tilde{O}(r^2 n^{-d})$ by union bound. □

Chapter 10

Bibliography

[1] Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, pages 676–684, 2015.

[2] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[3] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[5] Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.*, 41(4):2097–2122, 2013.

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

[7] Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems*, pages 2098–2106, 2015.

[8] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[9] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 02 2017.

[10] Peter J Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):253–273, 2016.

[11] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

[12] Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

[13] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[14] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[15] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[16] Mark Braverman, Konstantin Makarychev, Yury Makarychev, and Assaf Naor. The grothendieck constant is strictly smaller than krivine's bound. In *Forum of Mathematics, Pi*, volume 1, page e4. Cambridge Univ Press, 2013.

[17] T Tony Cai, Xiaodong Li, et al. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

[18] Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[19] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[20] Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, (just-accepted), 2016.

[21] Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. *arXiv preprint arXiv:1512.08425*, 2015.

[22] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

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

[24] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1):882–938, 2016.

[25] Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, pages 799–819, 2007.

[26] Richard Combes. An extension of mcdiarmid's inequality. *arXiv preprint arXiv:1511.05240*, 2015.

[27] Denis Conniffe. Expected maximum log likelihood estimation. *The Statistician*, pages 317–329, 1987.

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

[29] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

[30] Sanjoy Dasgupta and Leonard J Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 152–159. Morgan Kaufmann Publishers Inc., 2000.

[31] J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

[32] Kris De Brabanter, Kristiaan Pelckmans, Jos De Brabanter, Michiel Debruyne, Johan AK Suykens, Mia Hubert, and Bart De Moor. Robustness

of kernel based regression: a comparison of iterative weighting schemes. In *Artificial Neural Networks–ICANN 2009*, pages 100–110. Springer, 2009.

[33] Michiel Debruyne, Mia Hubert, and Johan AK Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9(10), 2008.

[34] Michiel Debruyne, Mia Hubert, and Johan Van Horebeek. Detecting influential observations in kernel pca. *Computational Statistics & Data Analysis*, 54(12):3007–3019, 2010.

[35] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[36] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on KDD*, pages 551–556. ACM, 2004.

[37] Lian Duan, Lida Xu, Ying Liu, and Jun Lee. Cluster-based outlier detection. *Annals of Operations Research*, 168(1):151–168, 2009.

[38] Noureddine El Karoui et al. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

[39] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.

[40] Zachary Friggstad, Mohsen Rezapour, and Mohammad R Salavatipour. Local search yields a ptas for k-means in doubling metrics. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 365–374. IEEE, 2016.

[41] Jorge Gil-Mendieta and Samuel Schmidt. The political network in mexico. *Social Networks*, 18(4):355–381, 1996.

[42] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

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

[44] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

[45] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster

recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

[46] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016.

[47] Greg Hamerly and Charles Elkan. Learning the k in k-means. In *In Neural Information Processing Systems*, page 2003. MIT Press, 2003.

[48] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[49] Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2008.

[50] Jake M Hofman and Chris H Wiggins. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701, 2008.

[51] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[52] Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

[53] Ute Jacob, Aaron Thierry, Ulrich Brose, Wolf E Arntz, Sofia Berg, Thomas Brey, Ingo Fetzer, Tomas Jonsson, Katja Mintenbeck, Christian Mollmann, et al. The role of body size in complex food webs: A cold case. *Advances In Ecological Research*, 45:181–223, 2011.

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

[55] Jiashun Jin, Zheng Tracy Ke, Wanjie Wang, et al. Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics*, 45(5):2151–2189, 2017.

[56] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, 2012.

[57] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[58] Dae-Won Kim, Ki Young Lee, Doheon Lee, and Kwang H Lee. Evaluation of the performance of clustering algorithms in kernel-induced feature space. *Pattern Recognition*, 38(4):607–611, 2005.

[59] JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565, 2012.

[60] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 28–36, 2014.

[61] Brian Kulis, Arun C Surendran, and John C Platt. Fast low-rank semidefinite programming for embedding and clustering. In *AISTATS*, pages 235–242, 2007.

[62] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems 24*, pages 1413–1421. 2011.

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

[64] Kenneth Lange. A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 425–437, 1995.

[65] Pierre Latouche, Etienne Birmele, and Christophe Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

[66] Can M Le and Elizaveta Levina. Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*, 2015.

[67] Can M Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*, 2015.

[68] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[69] Jing Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

[70] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[71] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.

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

[73] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley sym-*

posium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[74] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, pages 715–742, 2010.

[75] Jirı Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[76] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[77] Julian J McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *NIPS*, volume 2012, pages 548–56, 2012.

[78] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[79] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

[80] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[81] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. *arXiv preprint arXiv:1504.05910*, 2015.

[82] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, pages 814–827, New York, NY, USA, 2016. ACM.

[83] Elchanan Mossel and Jiaming Xu. Local algorithms for block models with side information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 71–80. ACM, 2016.

[84] Iftekhar Naim and Daniel Gildea. Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. *arXiv preprint arXiv:1206.6427*, 2012.

[85] MEJ Newman and Aaron Clauset. Structure and inference in annotated networks. *arXiv preprint arXiv:1507.04001*, 2015.

[86] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[87] Michael L. Overton and Robert S Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357, 1993.

[88] Art B Owen and Patrick O Perry. Bi-cross-validation of the svd and the nonnegative matrix factorization. *The annals of applied statistics*, pages 564–594, 2009.

[89] Rajendra Pamula, Jatindra Kumar Deka, and Sukumar Nandi. An outlier detection method based on clustering. In *2011 Second International Conference on EAIT*, pages 253–256. IEEE, 2011.

[90] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLOS genetics.*, 2(12):2074–2093, 2006.

[91] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, 2000. Morgan Kaufmann.

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

[93] William Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. *arXiv preprint arXiv:1507.05605*, 2015.

[94] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[95] Maria A Riolo, George T Cantwell, Gesine Reinert, and MEJ Newman. Efficient method for estimating the number of communities in a network. *arXiv preprint arXiv:1706.02324*, 2017.

[96] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[97] D Franco Saldana, Yi Yu, and Yang Feng. How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017.

[98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[99] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[100] Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pages 3960–3984, 2009.

[101] Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804, 1956.

[102] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

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

[104] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[105] Nicolas Verzelen, Ery Arias-Castro, et al. Detection and feature selection in sparse mixture models. *The Annals of Statistics*, 45(5):1920–1950, 2017.

[106] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

[107] YX Wang and Peter J Bickel. Likelihood-based model selection for stochastic block models. *arXiv preprint arXiv:1502.02069*, 2015.

[108] Haolei Weng and Yang Feng. Community detection with nodal information. *arXiv preprint arXiv:1610.09735*, 2016.

[109] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[110] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc., 2016.

[111] Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[112] Linli Xu, Koby Crammer, and Dale Schuurmans. Robust support vector machine training via convex outlier ablation. In *AAAI*, volume 6, pages 536–542, 2006.

[113] Bowei Yan and Purnamrita Sarkar. Convex relaxation for community detection with covariates. *arXiv preprint arXiv:1607.02675*, 2016.

[114] Bowei Yan and Purnamrita Sarkar. On robustness of kernel clustering. In *Advances in Neural Information Processing Systems*, pages 3090–3098, 2016.

[115] Bowei Yan, Purnamrita Sarkar, and Xiuyuan Cheng. Exact recovery of number of blocks in blockmodels. *arXiv preprint arXiv:1705.08580*, 2017.

[116] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient em on multi-component mixture of gaussians. In *Advances in Neural Information Processing Systems*, pages 6959–6969, 2017.

[117] Liuqin Yang, Defeng Sun, and Kim-Chuan Toh. Sdpnal+: a majorized semismooth newton-cg augmented lagrangian method for semidefinite programming with nonnegative constraints. *Mathematical Programming Computation*, 7(3):331–366, 2015.

[118] Miin-Shen Yang and Kuo-Lung Wu. A similarity-based robust clustering method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(4):434–448, 2004.

[119] Y Yu, T Wang, and RJ Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[120] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[121] Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*, 2017.

[122] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[123] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*, 2015.

[124] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.