Spectral clustering in the Gaussian mixture block model

Shuangping Li* Tselil Schramm†

Abstract

Gaussian mixture block models are distributions over graphs that strive to model modern networks: to generate a graph from such a model, we associate each vertex i with a latent "feature" vector $u_i \in \mathbb{R}^d$ sampled from a mixture of Gaussians, and we add edge (i, j) if and only if the feature vectors are sufficiently similar, in that $\langle u_i, u_j \rangle \geqslant \tau$ for a pre-specified threshold τ. The different components of the Gaussian mixture represent the fact that there may be different types of nodes with different distributions over features—for example, in a social network each component represents the different attributes of a distinct community. Natural algorithmic tasks associated with these networks are embedding (recovering the latent feature vectors) and clustering (grouping nodes by their mixture component).

In this paper we initiate the study of clustering and embedding graphs sampled from high-dimensional Gaussian mixture block models, where the dimension of the latent feature vectors $d \to \infty$ as the size of the network $n \to \infty$. This high-dimensional setting is most appropriate in the context of modern networks, in which we think of the latent feature space as being high-dimensional. We analyze the performance of canonical spectral clustering and embedding algorithms for such graphs in the case of 2-component spherical Gaussian mixtures, and begin to sketch out the information-computation landscape for clustering and embedding in these models.

*Stanford University. `fifalsp@stanford.edu`.
†Stanford University. `tselil@stanford.edu`. Supported by NSF CAREER award # 2143246.

Contents

1 Introduction

For algorithmic problems arising in data science, it is useful to study "model organisms:" families of synthetic datasets which aspire to faithfully represent the input data, while being simple enough to admit provable guarantees for the algorithms in question.

Consider, for example, the task of clustering in social networks: we observe an unlabeled graph in which each node belongs to one of several communities, and we would like to partition the graph into communities. This problem is hard to study in the traditional worst-case analysis framework for several reasons. Firstly, many of the objectives associated with finding community structure (e.g. sparsest cut, correlation clustering) are computationally intractable [MS90, BBC04] (even to approximate [CKK+06]) in the worst case. Secondly, it is not even clear that sparsest cut and other proxy clustering objectives faithfully capture community structure in real-life networks; in fact, they are sometimes poorly correlated with network structure even in non-adversarial synthetic models (see Section 3.1 of [Abb17]). Model organisms (such as the stochastic block model) allow researchers to "theoretically benchmark" the performance of their clustering algorithms. Even if the algorithm has no provable guarantees for worst-case inputs, at least we can rest assured that it works well in a simple generative data model.

Perhaps the most popular model for this task is the *Stochastic Block Model* (SBM) [HLL83], a generalization of an Erdős-Rényi graph in which each node belongs to an unknown community, and each edge is present independently with probability p between nodes in the same community and probability q between nodes in different communities (see [Abb17] for a definition of the model in its full generality). When $p > q$, this model reflects the empirical observation that social networks tend to have denser connections within communities. The SBM has been a successful model organism in that it has allowed for a nuanced study of spectral methods, motif (small subgraph) counting, and other algorithms for community detection (e.g. [MNS15, Mas14, BMR21, DdNS22]), and a rich mathematical theory has developed to describe its behavior (e.g. [MPW16, MNS16, AS15], see also the survey [Abb17]). However, this simple model leaves much to be desired because it fails to capture much of the structure of real-life social networks (see e.g. the discussion in [GMPS18]).

We study the following *Gaussian mixture block model* (GMBM) as a model organism for community recovery. The model is meant to reflect the conception of networks wherein each node is associated with a latent *feature vector* which describes its characteristics, and pairs of nodes with similar feature vectors are more likely to be connected. To generate a sample from the spherical 2-community Gaussian mixture block model, $G \sim \boldsymbol{G}_{n,d}(p, \mu)$ with $n, d \in \mathbb{Z}_+, \mu \in \mathbb{R}_+, p \in [0, 1]$, (1) independently sample n latent vectors from a mixture of Gaussians in $\mathbb{R}^d$,

$$u_1, \dots, u_n \sim \frac{1}{2}\mathcal{N}(-\mu \cdot e_1, \tfrac{1}{d} I_d) + \frac{1}{2}\mathcal{N}(\mu \cdot e_1, \tfrac{1}{d} I_d),$$

then (2) for each $i \neq j \in [n]$, add edge (i, j) to G if and only if the corresponding vectors are τ-correlated, $\langle u_i, u_j \rangle \geqslant \tau$, where τ is chosen in advance as a function of n, d, μ, and p so that the edge probability $\mathbf{Pr}[(i, j) \in E(G)]$ is p. Note that we ultimately observe only G and not the latent embedding $u_1, \dots, u_n$.

Each Gaussian component of the mixture represents the characteristics of a community as a distribution over feature space, and the distance between the means, 2μ, is a measure of the communities' separation. The edge criterion $\langle u_i, u_j \rangle \geqslant \tau$ reflects the intuition that nodes with similar feature vectors are more likely to be connected; the larger we set τ, the more stringent the connection criterion is, and therefore the sparser the resulting network becomes.

Variants of this model have been studied in the past, albeit with minor variations (for example, u_i sampled from the uniform distribution over the sphere rather than a Gaussian mixture) [GMPS18, EMP22]. But to date, the focus has been on the (more mathematically tractable) low-dimensional regime, where d remains fixed as $n \to \infty$.

In this work, we will study the performance of spectral algorithms in the GMBM in the *high-dimensional regime*, where $d \to \infty$ as $n \to \infty$. This high-dimensional setting is more compatible with our conception of modern networks, in which we think of the feature space as being large, on a scale comparable to the networks' size. We will show that so long as p is chosen to ensure that the network is not too sparse and so long as the dimension of the feature space d is not too large relative to the number of nodes n, the canonical spectral embedding algorithm provides a good estimate of the latent embedding $u_1, \dots, u_n$ (up to rotation). Further, if the separation between the communities μ is large enough, the spectral embedding allows us to test for the presence of and/or recover the communities.

Spectral embedding methods are used widely throughout network science, and our analysis is the first that provides provable guarantees for their performance in the relatively realistic high-dimensional Gaussian mixture block model. However, our work is merely an initial step; we will formulate several open questions for future research, both towards the goal of increasing the realism of the model, and better understanding the information-computation landscape of this simplest model.

1.1 Our results

We begin by defining our model and formulating our algorithmic objectives.

Definition 1.1 (Gaussian mixture block model)**.** For $n, d \in \mathbb{Z}_+$ and $\mu \geqslant 0, p \in [0,1]$, the *n-vertex, d-dimensional, μ-separated, 2-community Gaussian mixture block model with edge probability p* is the distribution over n-vertex graphs $G \sim \boldsymbol{G}_{n,d}(p,\mu)$ defined by the following sampling procedure:

1. Independently sample n d-dimensional vectors $u_1, \dots, u_n \sim \frac{1}{2}\mathcal{N}(-\mu \cdot e_1, \frac{1}{d}I_d) + \frac{1}{2}\mathcal{N}(\mu \cdot e_1, \frac{1}{d}I_d)$.
2. Let $V(G) = [n]$ and add (i,j) to $E(G)$ if and only if $\langle u_i, u_j \rangle \geqslant \tau$,

where $\tau = \tau(n,d,\mu,p)$ is a threshold chosen in advance so that $\mathbf{Pr}[\langle u_i, u_j\rangle \geqslant \tau] = p$.

We say that vertex $i \in [n]$ comes from community $+1$ if u_i comes from the component in the mixture with mean $\mu \cdot e_i$; otherwise we say that vertex i comes from community -1.

Remark 1.2. Because of the rotational invariance of $\mathcal{N}(0, \frac{1}{d}I_d)$, the distribution over graphs produced by the mixture with means $\pm\mu \cdot e_1$ is equivalent to that produced by the means $\pm\theta$ for any $\theta \in \mathbb{R}^d$ with $\|\theta\| = \mu$. We suspect that our techniques will generalize in a straightforward manner to the case in which the points are sampled from a mixture with any pair of means $\frac{1}{2}\mathcal{N}(\theta_1, \frac{1}{d}I_d) + \frac{1}{2}\mathcal{N}(\theta_2, \frac{1}{d}I_d)$, or to the case when edges are based on the distance criterion $\|u_i - u_j\| < \sigma$ (so that $\theta_1 = -\theta_2$ is equivalent to the general case) rather than correlation criterion. We opted to analyze only this special case in order to keep the proofs simpler.

Problem 1.3. There are three algorithmic problems we associate with the GMBM:

1. **Latent vector recovery:** Given $G \sim \boldsymbol{G}_{n,d}(p,\mu)$, can we estimate the latent vectors $u_1, \dots, u_n$ up to rotation?
2. **Hypothesis testing:** Can we tell if the data that generated G comes from two distinct clusters? Formally, we would like to hypothesis test between the "one community" null hypothesis that $G \sim \boldsymbol{G}_{n,d}(p,0)$ (so that $\mu = 0$ and the two components of the mixture are merged) and the "two community" alternative hypothesis $G \sim \boldsymbol{G}_{n,d}(p,\mu)$ with $\mu > 0$.
3. **Clustering:** When $\mu > 0$, can we partition the vertices $V(G)$ into sets $S, V \setminus S$ so that for all (or most) $i \in S$, i comes from community $+1$ and for all (or most) $j \in V \setminus S$, j is in community -1?

Spectral algorithm. Spectral methods are broadly employed for clustering and embedding tasks in network science. Though many variations on the basic concept exist, in this work we are mostly concerned with providing rigorous guarantees for this model organism, and we analyze a very canonical and "vanilla" spectral algorithm. Roughly, the algorithm is as follows: given input graph G on n vertices,

(a) Assemble the $n \times n$ adjacency matrix A with $A_{i,j} = \mathbb{1}[(i,j) \in E(G)]$, then

(b) Compute the top $d+1$ eigenvalues and unit eigenvectors $\{(\eta_i, w_i)\}_{i=0}^{d}$ of A, where $\eta_0 \geqslant \cdots \geqslant \eta_{n-1}$.

If d is unknown, in step (b) we simply look for a spectral gap: we find the smallest $i > 0$ so that $\eta_i - \eta_{i+1}$ is sufficiently large, and we let $i = d$. We then use this spectral information to solve vector recovery, hypothesis testing, or clustering, employing a slightly different final step in each case:

- For the task of recovery, we define the vectors $\hat{u}_1, \dots, \hat{u}_n$ with $\hat{u}_j(i) \propto w_i(j)$, and let $\hat{u}_j$ be our estimate for u_j (up to rotation).[1]
- For the task of hypothesis testing, we check if $\eta_1 > \theta$ for a threshold $\theta \in \mathbb{R}$ chosen as a function of n, d, μ, τ (larger η_1 corresponds to the case when $\mu > 0$).
- Finally, for the task of clustering, we embed the points on the line according to the vector w_1 and we let community membership be defined by a threshold cut along the line.

For each task, we are able to prove that the above algorithm works, provided the number of vertices n is large enough as a function of the dimension d, p is large enough so that the graph is not too sparse, and the separation μ between the cluster centers is not too small (except for some gaps of logarithmic width). In what follows, we use $\overset{\log}{\ll}$ to denote $\ll$ with a polylog factor. The following diagram summarizes our results; we give the theorem statements below.

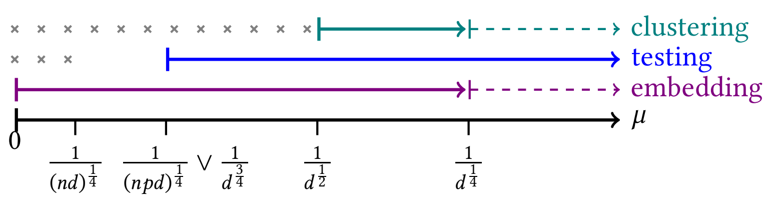

Figure 1: Diagram illustrating the range of μ for which we show that the spectral algorithm completes each task successfully (up to logarithmic factors), all under the condition that $1 \overset{\log}{\ll} d \overset{\log}{\ll} pn$. The solid lines correspond to our theorems. The dashed teal line indicates that beyond $d^{-1/4}$, each community corresponds to a distinct connected component in the graph and thus spectral clustering trivially succeeds. Similarly, the dashed violet line indicates that beyond $d^{-1/4}$, the community labels suffice to recover an approximate embedding. The gray x's mark a range in which clustering/testing is impossible even when the latent embedding is known (lower bounds for clustering in [Nda22], for testing in Appendix A).

Theorem 1.4 (Latent vector recovery/embedding). *Suppose that $n, d \in \mathbb{Z}_+$ and $\mu \in \mathbb{R}_+$, and $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, satisfy the conditions $\log^{16} n \ll d < n$, $\mu^2 \leqslant 1/(\sqrt{d}\log n)$, and $pn \gg 1$. Then given $G \sim \boldsymbol{G}_{n,d}(p,\mu)$ generated by latent vectors $u_1, \dots, u_n \in \mathbb{R}^d$, the spectral algorithm described above produces vectors $\hat{u}_1, \dots, \hat{u}_n$ which satisfy*

$$\mathop{\mathbf{E}}_{i,j\sim[n]} |\langle \hat{u}_i, \hat{u}_j\rangle - \langle u_i, u_j\rangle| \ll \max\left\{ \sqrt{\frac{\log 1/p}{d}},\ \mu^2 \cdot \sqrt{\frac{d}{\log \frac{1}{p}}},\ \sqrt{\frac{d}{np\log\frac{1}{p}}} \right\} \log^9 n \mathop{\mathbf{E}}_{i,j\sim[n]} |\langle u_i, u_j\rangle|,$$

with high probability as n goes to infinity.

As long as $1 \overset{\log}{\ll} d \overset{\log}{\ll} pn$ and $\mu \overset{\log}{\ll} d^{-1/4}$, the relative error in Theorem 1.4 is $o(1)$ and the $\hat{u}_i$ recover the u_i approximately, up to rotation. See also Theorem 4.21 which states an approximation result in terms of the spectral distance between the matrices whose columns are given by $\hat{u}_i$ and u_i respectively.

[1]Note vector entries $u_j(\cdot)$ and $w_i(\cdot)$ are indexed starting at 1, whereas the eigenvectors $w_{(\cdot)}$ are indexed starting at 0.

The condition $d \ll pn$ asks that the dimension not exceed the average vertex degree. It is not difficult to see that if $d \to \infty$ too fast relative to n, the geometry disappears (because the quantities $\{\langle u_i, u_j \rangle\}_{i,j\in[n]}$ become increasingly independent). Similarly, when p is small more information is "lost." So it is unsurprising that we see an upper bound on d as a function of n and p. However, it is not clear to us whether $d \ll np$ is sharp (even up to logarithmic factors); we discuss more below in the "lower bounds" paragraph.

The condition $\mu \overset{\log}{\ll} d^{-1/4}$ ensures that the separation between the communities does not "drown out" other geometric information present in the graph. Indeed, whenever $\mu \gg (1/\sqrt{d\tau}) = \Theta((d \log \frac{1}{p})^{-1/4})$, the number of edges between communities is of a smaller order than the number of edges within each community, and whenever $\mu \gg d^{-1/4} \log^{1/4} n$, the components of the graph corresponding to the communities become disconnected with high probability.[2] In this $\mu \overset{\log}{\gg} d^{-1/4}$ range, embedding reduces to clustering, as $\hat{u}_i = x_i \mu e_1$ for x_i the ± 1 label of node i is a good approximate embedding for u_i.

Theorem 1.5 (Hypothesis Testing)**.** *Define the one-community model to be the null hypothesis $H_0 = \boldsymbol{G}_{n,d}(p, 0)$ and the two-community mixture model to be the alternative hypothesis $H_1 = \boldsymbol{G}_{n,d}(p, \mu)$. If d, n, μ, p satisfy*

$$\mu^2 \geqslant \max\left\{\sqrt{\frac{\log 1/p}{d^3}}, \sqrt{\frac{1}{npd\log\frac{1}{p}}}\right\} \log^9 n, \qquad \log^{16} n \ll d < n, \qquad pn \gg 1, \qquad p \in [0, 1/2 - \varepsilon],$$

for any constant $\varepsilon > 0$, then if we run the spectral algorithm described above on input graph G we have that

$$\min\{\mathbf{Pr}(\mathit{accept}\ H_0 \mid G \sim H_0),\ \mathbf{Pr}(\mathit{reject}\ H_0 \mid G \sim H_1)\} \geqslant 1 - o_n(1),$$

with high probability as n goes to infinity. In other words, both type 1 and type 2 error go to zero as $n \to \infty$.

Theorem 1.5 requires that $\mu \overset{\log}{\gg} \max(d^{-3/4}, (npd)^{-1/4})$ for the errors to vanish. This is much smaller than $1/\sqrt{d}$ whenever $d \overset{\log}{\ll} pn$. In other words, whenever $d \overset{\log}{\ll} pn$, we can tell apart the two models even if the mean separation is $o(1/\sqrt{d})$ and almost-exact clustering is impossible. A second moment computation shows that $\mu = \Omega((nd)^{-1/4})$ is necessary for testing (Appendix A).

Theorem 1.6 (Spectral clustering)**.** *Suppose $d, n \in \mathbb{Z}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, and $\mu > 0$ satisfy the conditions $d^{-1/2} \ll \mu \leqslant d^{-1/4} \log^{-1/2} n$, $\log^{16} n \ll d < n$ and $pn \gg 1$. If $G \sim \boldsymbol{G}_{n,d}(p, \mu)$, then with high probability the spectral algorithm described above correctly labels (up to a global sign flip) a*

$$1 - O\left(\frac{1}{\mu\sqrt{d}} + \sqrt{\max\left\{\mu^2 \cdot \sqrt{\frac{\log 1/p}{d}}, \frac{1}{\mu^2\sqrt{npd\log\frac{1}{p}}}\right\} \log^9 n}\right) \text{-fraction of the vertices.}$$

Theorem 1.6 implies that when $1 \overset{\log}{\ll} d \overset{\log}{\ll} pn$ and $d^{-1/2} \overset{\log}{\ll} \mu \overset{\log}{\ll} d^{-1/4}$, the spectral clustering algorithm clusters a $(1-o(1))$-fraction of vertices correctly. As discussed briefly above, we expect that when μ exceeds the range covered by Theorem 1.6, the spectral algorithm also works because the cut between the $+1$ and -1 labeled vertices becomes sparse (and eventually has no crossing edges with high probability when $\mu \gg (\frac{1}{d} \log n)^{1/4}$); however this regime requires a different analysis than our theorem (and can probably avoid the trace method), and we defer it to future work. The bound $d^{-1/2} \overset{\log}{\ll} \mu$ matches the lower bound in the Gaussian mixture model (when the u_i are not latent but observed) up to logarithmic factors [Nda22].

Lower bounds. It is interesting to know whether our results are tight. A priori, in some regimes we can deduce information-theoretic barriers to solving the embedding and/or hypothesis testing problem by appealing to a known barrier for another model:

[2] For reference, so long as $d = \omega(1)$ and μ is small enough that it does not dominate the edge probability, $\tau = \Theta(\sqrt{\frac{1}{d}} \log 1/p)$.

- When μ is too small (precisely, when $\mu \lesssim \sqrt{\frac{1}{d} + \frac{1}{\sqrt{nd}}}$), clustering in the underlying Gaussian mixture model is information-theoretically impossible [Nda22]. Given only $G \sim \boldsymbol{G}_{n,d}(\mu, \tau)$, we have access to strictly less information than if we were handed the latent embedding, so clustering in G in this regime is impossible.
- Similarly, we show in Appendix A that when $\mu \lesssim \left(\frac{1}{nd}\right)^{1/4}$, the underlying Gaussian mixture model cannot be hypothesis tested against $\mathcal{N}(0, \frac{1}{d}\mathbb{1})$ from n samples. This implies that testing is impossible in this regime.
- When $d = \Omega(\log n)$ and $\mu = 0$, the graph we observe is a random geometric graph with points sampled from $\mathcal{N}(0, \frac{1}{d}I_d)$, which is not too different from a random geometric graph over $\mathcal{S}^{d-1}$. A random geometric graph over $\mathcal{S}^{d-1}$ with average degree np is known to be indistinguishable from Erdős-Rényi $\boldsymbol{G}(n, p)$ when $d \gtrsim n^3 p^2$, and is conjectured indistinguishable when $d \gtrsim (nH(p))^3$ for $H(\cdot)$ the binary entropy function [LMSY22b]. In this regime, the embedding that generated the GMBM graph G is likely not identifiable, and at the very least it is not meaningful because the geometry has effectively "disappeared" in the observed graph.

These lower bounds are inherited by $\boldsymbol{G}_{n,d}(p, \mu)$ from these simpler models, and they are consistent with our results which do need μ not too small to cluster and d not too large relative to np. However, these lower bounds do not give the full story—the first two bounds do not account for the information lost when the embedding is thrown out, and the latter lower bound is very far from our result,[3] and does not explain what happens when μ is large.

The condition $d = O(nH(p))$ for $H(\cdot)$ the binary entropy function seems plausibly tight for spectral algorithms; see Section 1.3 for more discussion. We also mention that since the span of the embedding vectors $u_1, \dots, u_n$ has rank at most n, when the embedding dimension $d > n$, there is always some set of vectors in n dimensions, $v_1, \dots, v_n \in \mathbb{R}^n$, so that $\langle u_i, u_j\rangle = \langle v_i, v_j\rangle$ for all i, j; hence the embedding which produced the graph is certainly no longer unique up to rotation. However, it is possible that the embedding may still be identifiable (up to rotation) in the $d > n$ regime, because the u_i also behave like Gaussian vectors; it is not clear whether the same is true for alternate embeddings such as the one given by $\{v_i\}_{i\in[n]}$.

We leave proving lower bounds as an interesting open question. In the case of embedding, one might imagine that in the large-d regime, it might be possible to demonstrate (via direct calculation or otherwise) that the posterior distribution over embeddings has high entropy. An alternate approach in the sparse regime might be to try to adapt the arguments for lower bounds in the stochastic block model (using information flow on trees); see [EMP22] for a discussion of this possibility.

1.2 Related work

Gaussian mixture block model and variations. Most of the prior work on geometric block models in the literature focuses on the low-dimensional regime, where the dimension d is held fixed as $n \to \infty$. For example, in [ABARS20], the authors study the performance of a (somewhat different) spectral algorithm for the approximate clustering problem in the special case of $d = 2$. Another previously studied slight variation on our model was studied in [GMPS18, GMPS19]. There, the authors introduce a variant of the model in which the ith node's latent vector z_i is sampled uniformly at random from the unit sphere $\mathcal{S}^{d-1}$, and each node also has a latent community label $x_i \in [k]$. Edge (i, j) is present if and only if $\langle z_i, z_j\rangle \geqslant \tau_{x_i,x_j}$, so the connectivity threshold depends on the communities that the nodes belong to. Though not exactly the same, this is not too dissimilar from our model.[4] The authors study the relatively low-dimensional case

[3] We require $d \lesssim np$, rather than $d \lesssim (nH(p))^3$, though we are not sure if either of these bounds is sharp.

[4] In our case $u_i = Z_i + x_i\mu e_1$ for $Z_i \sim \mathcal{N}(0, \frac{1}{d}I_d)$ (almost like vectors on the sphere when d is large) and $x_i \in \{\pm 1\}$ the community label of node i, so that $\langle u_i, u_j\rangle = \langle Z_i, Z_j\rangle + \mu\langle x_j Z_i + x_i Z_j, e_1\rangle + x_i x_j \mu^2$. Now the condition $\langle u_i, u_j\rangle \geqslant \tau$ is equivalent to the condition $\langle Z_i, Z_j\rangle \geqslant \tau - x_i x_j \mu^2$ up to the random fluctuation $\mu\langle x_j Z_i + x_i Z_j, e_1\rangle$.

$d = O(\log n)$ for this model, and give a clustering algorithm based on counting motifs (small subgraphs) that works in some parameter regimes. See also [ABD21] for a result on the success of spectral clustering when the feature vectors are drawn from the uniform measure over the torus $\mathbb{T}^d$ with $d = O(1)$.

Other works have considered variants of the classical stochastic block model which incorporate higher-dimensional geometry. In [SB18, ABS21] the authors consider a version of the SBM in which one observes a known embedding of the nodes in $\mathbb{R}^d$, but the community labels of the nodes are latent and the edges are a function of both the embedding and the labels. Their setting differs significantly from ours because the node embedding is not latent. Another instance of a geometric variant of the block model is the mixed-membership k-community stochastic block model [ABFX08]. There each node i has a latent k-dimensional "community membership" distribution $u_i \in \Delta_k$ where Δ_k denotes the k-dimensional simplex; u_i is supposed to represent node i's fractional belonging to each of the k different communities. The u_i are sampled independently from a Dirichlet distribution, and then the presence of the edge (i, j) is a randomized function of u_i and u_j.

In this context (more sophisticated) spectral algorithms are also known to recover the latent embedding, even in the sparse and high-dimensional regime where k grows with n [HS17]. The underlying geometry in the mixed-membership block model is quite different from the Gaussian mixture block model (as the embedding vectors u_i are supposed to represent something else), so the result of [HS17] does not imply anything for our setting. Still, the fact that the techniques in these cases are similar further points to a potential universality of methods for recovering embeddings in random geometric graphs. We also think it likely that the spectral algorithms from [HS17] could give sharper (up to logarithmic factors) algorithmic thresholds, for instance removing polylogs from the inequality $d \lesssim np$ and potentially allowing us to handle sparse graphs with $np = \Theta(1)$; however in this work our goal was to analyze the more "canonical" and efficient basic spectral method.

Recovering embeddings of random geometric graphs. A *random geometric graph* is any graph which is generated by sampling points according to a measure on a metric space, associating each point to a vertex, and connecting vertices according to a probability which depends on their distance in the metric. The GMBM is a geometric random graph where the metric space is Euclidean space and the measure is a mixture of two spherical Gaussians. Recovering the embedding is a natural algorithmic task on random geometric graphs; we mention a couple of relevant works here.

Motivated by social networks, the work [EMP22] considers the task of recovering the embedding of a random geometric graph over a single unit sphere $\mathcal{S}^{d-1}$, where the edge indicator for (i, j) is distributed as a Bernoulli$(\frac{1}{n}\phi(\langle u_i, u_j \rangle))$ for u_i, u_j the latent embedding vectors (the fact that the edge probabilities are a function of distance is supposed to mimic community structure). They prove that so long as ϕ's spectrum satisfies certain conditions (which restrict their result to the low-dimensional setting), the spectral embedding approximates the latent embedding well.

The work of [EMP22] makes use of techniques used in the study of kernel random matrices, which is a class of random matrices that includes the adjacency matrices of random geometric graphs. These are matrices which are sampled by first sampling n latent vectors $u_1, \dots, u_n$ uniformly from $\mathcal{S}^{d-1}$, and then taking the (i, j)-th entry to be a (deterministic) function ϕ of $\langle u_i, u_j \rangle$. The spectrum (eigenvalues and eigenvectors) of a kernel random matrix can be thought of as a random approximation to the spectrum of the associated integral operator. A sufficiently strong quantitative bound on the strength of this approximation in the context of the random geometric graph kernel function, $\phi(x) = \mathbb{1}[x \geqslant \tau]$, would imply that spectral algorithms recover the latent embedding. For example, in the low-dimensional setting where $d = O(1)$, the classical results [KG00] imply that a spectral algorithm recovers the embedding of a kernel random matrix up to error which goes to zero as $n \to \infty$. See also [AVY19].

In the high dimensional setting less is known. A series of works [EK10, CS13, DV13, Bor13, FM19, LY22] characterizes the empirical spectral densities of well-behaved kernel functions, under restrictions on the relationship between d and n (state of the art requires $d = n^{1/k}$ for integer k). Recently, in a push to more finely characterize the non-asymptotic performance of Kernel methods in machine learning, a series of papers has made progress on understanding the spectral edge and eigenvectors of certain kernel random matrices as well. See [FM19, GMMM21, MMM21, MMM22], which study a fairly flexible class of kernel functions (though as far as we understand, their results do not accommodate random geometric graphs). Our result shows that part of the spectrum of the random geometric graph approximates the associated operator's spectrum well. We mention as well the work [LMSY22a], in which the authors obtain sharp bounds on the spectral gap of random geometric graphs on the high-dimensional sphere.

Clustering Gaussian mixtures. The well-studied problem of clustering mixture distributions, and specifically Gaussian mixtures, is the easier version of our problem in which we are already given access to the latent embedding of our vertices in $\mathbb{R}^d$. The performance of spectral clustering algorithms in this context has been well studied, with a focus on the many-cluster case (see e.g. [VW04], and [Nda22] for the case of 2 Gaussians).

Our problem is harder because we do not observe this latent embedding. Our clustering algorithms work by trying to find an approximate embedding, and then applying spectral clustering. It is interesting to ask whether we could use a more sophisticated clustering algorithm on our approximate embedding. Though the robust Gaussian clustering problem is well-studied, the works that we are aware of consider contamination models in which a small fraction of the data points have been corrupted arbitrarily (e.g. [DKK$^+$18, HL18, KSS18]). Here, we are instead interested in the case where all of the points may be corrupted, but the corruption overall is bounded in operator norm.

Comparison to the Stochastic Block Model. Of interest is how our results compare to the classic two-community stochastic block model (SBM). The distribution $\mathrm{SSBM}(n, 2, A, B)$ is defined as follows: every vertex $i \in [n]$ is assigned a community label independently from $\mathrm{Unif}(\{\pm 1\})$, and then each edge (i, j) is added independently with probability A if nodes i, j belong to the same community, and with probability B otherwise. The signal-to-noise ratio in this model can be expressed as $\lambda(n, A, B) = \sqrt{\frac{n(A-B)^2}{2(A+B)}}$.

The clustering and testing problems have been well-studied in the SBM (embedding has no analogue). Here we focus on the $\Omega(\log n)$-average-degree results, as they are most directly comparable to our setting. For clustering, almost exact recovery (a $1 - o_n(1)$ fraction of the vertices are labelled correctly) can be achieved if and only if $\lambda(n, A, B) = \omega(1)$ [YP14, MNS14, AS15]. Some works have studied the performance of spectral algorithms for clustering in SBMs, including [Bop87, McS01, CWA12, Vu18, RCY11]. The setting in each of these works differs slightly, but the bottom line is that exact recovery (where all vertices are labeled correctly) by spectral algorithm is possible if $\lambda(n, A, B) = \Omega(\sqrt{\log n})$.

For the task of testing, the null hypothesis is the Erdős-Rényi graph $G(n, \frac{A+B}{2})$ with n vertices and edge probability $\frac{A+B}{2}$. It is known that if $\lambda(n, A, B) > 1$, then consistent testing is possible, in both the bounded degree case [MNS15] (nA and nB are constants as $n \to \infty$) and in the growing degree case [Jan95, Ban18] ($nA, nB \to \infty$ as $n \to \infty$). Furthermore, there are tests with polynomial time complexity in the both cases [MNS15, BM17, MS16]. When $\lambda(n, A, B) < 1$, the two distributions are asymptotically mutually contiguous, thus there is no consistent test.

In order to compare with the stochastic block model, we first match the parameters in our setting to the SBM case. In the GMBM setting, when $\mu \leqslant d^{-1/4} \log^{-1/2}(n)$, the within-community edge probability is $A_{\mathrm{GMBM}} = p + \Theta(p\tau\mu^2 d)$, and the across-community edge probability is $B_{\mathrm{GMBM}} = p - \Theta(p\tau\mu^2 d)$. A calculation shows that if we plug these edge probabilities in the the SBM signal-to-noise ratio function, $\lambda(n, A_{\mathrm{GMBM}}, B_{\mathrm{GMBM}}) = \Theta(\sqrt{npd\log(1/p)}\mu^2)$. Requiring $\lambda(n, A_{\mathrm{GMBM}}, B_{\mathrm{GMBM}}) \gg 1$ is equivalent (up to poly-log factors) to one of our requirements for testing, that $\mu \overset{\log}{\gg} (npd\log(1/p))^{-1/4}$. Our second requirement in the GMBM, that $\mu \overset{\log}{\gg} d^{-3/4}$, is a result of the geometric structure of the model, and is the dominant term in the maximum only when $d^2 \ll np$; it is plausible that the signal-to-noise ratio for testing in the GMBM is indeed impacted by the geometry, and differs in this lower-dimensional setting.

For clustering, the requirement that $\mu \overset{\log}{\gg} d^{-1/2}$ is more strict than $\lambda(n, A_{\mathrm{GMBM}}, B_{\mathrm{GMBM}}) = \omega(1)$ when $np \gg d$. However, this requirement derives from the fact that even the Gaussian mixture itself is not clusterable when $\mu \ll d^{-1/2}$, so even given perfect access to the Gaussian mixture embedding this clustering task is impossible. For $d \gg np$, we anticipate that the threshold will be the same (or similar) for both the GMBM and the SBM. Our results primarily focus on the moderate range where $d \ll np$.

1.3 Directions for future research

Our work makes an initial study of algorithmic tasks in a basic high-dimensional geometric block model. It is our hope that this is merely a small early step, and that GMBMs will become a standard model organism for network science. Here, we highlight a couple of intriguing directions for future research.

1. *Characterize the information-computation landscape of the Gaussian Mixture block model.*

   Here, we have given polynomial-time algorithms for this basic geometric block model that work so long as certain conditions are met: $1 \overset{\log}{\ll} d \overset{\log}{\ll} pn$ in all cases, and $d^{-1/2} \overset{\log}{\ll} \mu$ for clustering (save for a logarithmic-scale window around $\mu = d^{-1/4}$, also present for embedding, where we suspect a more careful analysis might succeed), and $d^{-3/4} + (npd)^{-1/4} \overset{\log}{\ll} \mu$ for hypothesis testing. But our understanding of the information-theoretic landscape for these problems is incomplete.

   It seems that a natural requirement for the success of spectral algorithms is $d = O(nH(p))$ for $H(\cdot)$ the binary entropy function. This is because the order of random fluctuations in the spectrum of the adjacency matrix is at least $\sqrt{np}$, and the d-dimensional embedding has eigenvalues on the order of $np\tau = \Theta(np\sqrt{d^{-1}\log 1/p})$, so one should only expect spectral embedding to succeed when $np\tau > \sqrt{np}$, that is when $d = O(nH(p))$. In the sparse regime $np = O(1)$, the fluctuations of the adjacency matrix are actually of higher order than $\sqrt{np}$ (which is why our algorithm requires $np = \Omega(\operatorname{polylog} n)$), but it is possible that the precise threshold $d \sim nH(p)$ could be achieved by the non-backtracking matrix or a spectral algorithm in the style of [HS17]. It is unclear to us whether *any* algorithm, polynomial-time or otherwise, can succeed in the regime $d \gg pn$.

   In the stochastic block model, extensive study has been made of the information and computational landscape of the clustering problem, and beautiful conjectures from statistical physics have been confirmed by an elegant mathematical theory to establish the existence of *information-theoretic* and computational phase transitions. When the signal-to-noise ratio in the SBM (a function of the inter- and intra-community edge probabilities and the number of communities) is below the Kesten-Stigum threshold, the hypothesis testing problem is believed to be computationally hard; there is also an information-theoretic threshold below which it is impossible to tell a graph generated from the SBM apart from an Erdős-Rényi graph[5].

   It seems plausible that geometric block models exhibit a similarly rich computational landscape, both for clustering and separately for embedding; we feel that charting this landscape is an exciting direction for future research. The excitement is deepened by the fact that the mathematical tools used in the context of the stochastic block model seem ill-suited to this more geometric setting (see also [EMP22] for some discussion). To our knowledge, the field is wide open both on the algorithmic/mathematical side and also from the perspective of predictions in statistical physics.

[5] The information-theoretic threshold coincides with the Kesten-Stigum transition in the 2-community case [DKMZ11, Abb17], see also the more recent [MSS22].

2. *Understand the performance of spectral (and other) algorithms for GMBMs generated by a wider class of Gaussian mixtures.*

   Though the GMBM that we have studied in this paper is certainly a more realistic than, for example, the stochastic block model, it is still far from capturing most real-world settings. Recall that the underlying Gaussian mixture is supposed to model a distribution over the feature space of network nodes. Here we have only studied the case of a mixture of at most two perfectly spherical Gaussians.

   It would be interesting to understand in which scenarios the spectral embedding algorithm continues to work if the number of communities is larger than two. More problematically, it seems unlikely that spectral embedding will succeed out-of-the-box when the Gaussian covariances are far from being spherical; this is even more so true for clustering, as spectral algorithms are known to fail even when given access to the true Gaussian mixture model in the non-spherical case [AM05]. Is it possible to design other algorithms for embedding and clustering for this more general "model organism," and would such algorithms yield insights which would transfer well to practice?

2 Technical overview

Recall that we have let A denote the adjacency matrix of G and that (η_i, w_i) are the eigenvalues and corresponding unit eigenvectors of A, where $\eta_0 \geqslant \cdots \geqslant \eta_{n-1}$. Define the vectors $\hat{u}_1, \ldots, \hat{u}_n \in \mathbb{R}^d$ by setting

$$\hat{u}_j(i) := \sqrt{\frac{\max(\eta_i, 0)}{\tilde{d}\lambda_1}} w_i(j),$$

where $\tilde{d}$ and λ_1 are to be specified later. And define the $n \times d$ matrix U by putting all the vectors $\hat{u}_1, \cdots, \hat{u}_d$ as rows in the matrix. In other words, $\tilde{d}\lambda_1 \mathsf{U}\mathsf{U}^\top$ is the projection of A on to the subspace spanned by its second to $d+1$-st eigenvector. Similarly, we define the $n \times d$ matrix U by putting all the true latent vectors $u_1, \cdots, u_d$ as rows in the matrix. The key is to show that the $\hat{u}_i$ approximate the latent u_i, spectrally. For technical reasons, we treat the case when μ is small and large separately. When μ is small, we show:

Theorem 2.1. *Suppose $d, n \in \mathbb{Z}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, and $\mu > 0$ satisfy the conditions $\log^{16} n \ll d < n$, $\mu \leqslant \tau$, and $pn \gg 1$, then we have that*

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{op}} \ll \max\left\{\frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p}d\tau}\right\} \log^9(n).$$

For references' sake (and for the sake of comparison with the theorem statements in the introduction), when $\mu \overset{\log}{\ll} d^{-1/4}$ is not too large and in the high-dimensional regime $d = \Omega(\log n)$, $\tau = \Theta(\sqrt{\frac{1}{d}\log 1/p})$. This can be seen by noting that the distribution of $\langle u_i, u_j\rangle$ is close to $\mathcal{N}(0, \frac{1}{d})$, and so $\mathbf{Pr}[\langle u_i, u_j\rangle \geqslant \tau] \approx \exp(-d\tau^2/2)$.

Proving Theorem 2.1 amounts to showing that the top $d+1$-dimensional eigenspace of A is spanned by the columns of U and a non-negative vector $\tilde{\mathbb{1}}_n \in \mathbb{R}^n$ (whose entries scale as a function of the length of the corresponding u_i). Specifically, we show:

Proposition 2.2. *Suppose $d, n \in \mathbb{Z}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, and $\mu > 0$ satisfy the conditions $\log^{16} n \ll d < n$, $\mu \leqslant \tau$, and $pn \gg 1$, there exist a length-n vector $\tilde{\mathbb{1}}_n$, constants p_0, $\tilde{d}$ and λ_1 to be defined later, such that with high probability,*

$$\|A - p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{op}} \ll \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}.$$

The above is enough to imply that $\tilde{\mathbb{1}}_n$ is close to the top eigenvector of A and the columns of U are close to the span of the next d eigenvectors of A so long as

$$np \approx \sigma_{\min}(p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n) \gg \sigma_{\max}(\tilde{d}\lambda_1 \cdot UU^\top) \geqslant \sigma_{\min}(\tilde{d}\lambda_1 \cdot UU^\top) \gg \max\left\{np\tau^2, \sqrt{np}\right\} \cdot \log^9 n,$$

for $\sigma_{\min}$ and $\sigma_{\max}$ denoting the minimum and maximum singular values, respectively. Applying spectral concentration of Wishart matrices, all of the singular values of $\tilde{d}\lambda_1 \cdot UU^\top$ are of order $np\tau$ with high probability, so A's top eigenspace is well-approximated by U when $\frac{1}{\sqrt{np}} \overset{\log}{\ll} \tau \overset{\log}{\ll} 1 \iff 1 \overset{\log}{\ll} d \overset{\log}{\ll} np$ since $\tau = \Theta(\sqrt{\frac{1}{d}\log 1/p})$; this is the source of our upper bound on d.

Linear approximation of the adjacency matrix. The i,j-th entry of A is a function of the inner product of the latent u_i, u_j:

$$A_{i,j} = A_{i,j}(\langle u_i, u_j\rangle) = \mathbb{1}[\langle u_i, u_j\rangle \geqslant \tau].$$

We can understand the matrix $p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 \cdot UU^\top$ subtracted in Proposition 2.2 as a linear approximation of A in the inner products $\langle u_i, u_j\rangle$. Intuitively, if we were to express $\mathbb{1}[\langle u_i, u_j\rangle \geqslant \tau]$ as a polynomial in $\langle u_i, u_j\rangle$, we'd see that $p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top$ is roughly the zeroth order term of the polynomial and $\tilde{d}\lambda_1 UU^\top$ is the first order term. The error term $np\tau^2$ in Proposition 2.2 reflects the fact that the second order coefficient in the polynomial expansion of $A_{i,j}$ is quadratic in τ (the other error term, $\sqrt{np}$, comes from random fluctuations). So in effect, we want to show that A is well-approximated by its linear term in a polynomial basis.

Our proof will proceed by applying the trace method. We will expand A's entries in the basis of Gegenbauer polynomials, which is a basis of polynomials that enjoys nice orthogonality properties when evaluated on inner products of random vectors on the unit sphere (see Section 4.1 for details). In the high-dimensional setting $d \gg 1$, the latent vectors u_i lie roughly on the sphere $\mathcal{S}^{d-1}$, so we can write $u_i = v_i + \text{error}$, where each $v_i \sim \text{Unif}(\mathcal{S}^{d-1})$. Therefore, we can write

$$A_{i,j} = \mathbb{1}(\langle u_i, u_j\rangle \geqslant \tau) = \mathbb{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j})$$

for some $\tau^{i,j}$ close to τ. Ignoring the difference between $\tau^{i,j}$ and τ for the moment, we then expand the threshold function $\mathbb{1}(\cdot \geqslant \tau)$ in the Gegenbauer polynomial basis $q_0, q_1, \ldots$ (in this proof overview, the q_k are implicitly renormalized to ease notation), so we have that

$$A_{i,j} = \sum_{k=0}^{\infty} c_k q_k(\langle v_i, v_j\rangle),$$

and our goal now reduces to showing that when we subtract the $k=0$ and $k=1$ terms, the operator norm of the resulting matrix $A_{\geqslant 2}$ with $A_{\geqslant 2}(i,j) = \sum_{k=2}^{\infty} c_k q_k(\langle v_i, v_j\rangle)$ is bounded.

The trace method. The trace method relates the maximum eigenvalue of a matrix to the expectation of a power's trace (using Markov's inequality): for any integer ℓ,

$$\|M\| > t \implies \operatorname{tr}(M^{2\ell}) > t^{2\ell}, \quad \text{so } \mathbf{Pr}[\|M\| > e^{\varepsilon}\,\mathbf{E}[\operatorname{tr}(M^{2\ell})]^{1/2\ell}] \leqslant \mathbf{Pr}\left[\operatorname{tr}(M^{2\ell}) > e^{2\ell\varepsilon}\,\mathbf{E}[\operatorname{tr}(M^{2\ell})]\right] \leqslant e^{-2\ell\varepsilon}.$$

So choosing, say, $\varepsilon = \frac{1}{\log n}$ and $\ell = \omega(\log^2 n)$ gives us that $\|M\| \leqslant (1 + o(1))\,\mathbf{E}[\operatorname{tr}(M^{2\ell})]^{1/2\ell}$ with high probability.

The trace method thus allows us to relate the operator norm, an analytic quantity, to degree-2ℓ moments of entries of a random matrix. Specifically, we can relate the trace of a power of M to expected value of products over "walks" of length 2ℓ in K_n weighted by the expected product of the edge "weights" given by the entries of M:

$$\mathbf{E}\,\mathrm{tr}(M^{2\ell}) = \sum_{i_1,\dots,i_{2\ell}\in[n]} \mathbf{E}\left[\prod_{s=1}^{2\ell} M_{i_s,i_{s+1}}\right],$$

where the subscript $s+1$ is understood to be taken modulo 2ℓ.

We apply the trace method with $M = A_{\geqslant 2}$, and our goal becomes to upper bound $\mathbf{E}[\mathrm{tr}(A_{\geqslant 2}^{2\ell})]$ by a quantity scaling like $\tilde{O}(np\tau^2)^{2\ell}$ for $\ell = \operatorname{polylog} n$. To analyze the expected trace, we make use of the orthogonality properties of Gegenbauer polynomials evaluated on inner products of random vectors on $\mathcal{S}^{d-1}$. We have

$$\mathbf{E}\,\mathrm{tr}(A_{\geqslant 2}^{2\ell}) = \sum_{i_1,\dots,i_{2\ell}\in[n]} \mathbf{E}\left[\prod_{s=1}^{2\ell}\sum_{k=2}^{\infty} c_k q_k(\langle v_{i_s}, v_{i_{s+1}}\rangle)\right] = \sum_{\substack{i_1,\dots,i_{2\ell}\in[n]\\ k_1,\dots,k_{2\ell}\geqslant 2}} \mathbf{E}\left[\prod_{s=1}^{2\ell} c_{k_s} q_{k_s}(\langle v_{i_s}, v_{i_{s+1}}\rangle)\right]$$

The orthogonality properties of the Gegenbauer polynomials will (at a high level) allow us to eliminate summands which contain any terms of different orders, $k_s \neq k_{s'}$, unless the indices $i_s, i_{s'}$ appear with high multiplicity. This is very helpful in our accounting and allows us to show that the $c_2 \approx p\tau^2$ coefficients more-or-less dominate the summation. The all-c_2 term in which no i_s is repeated has a contribution bounded by $d^2 n^{2\ell} c_2^{2\ell}$, and since this is roughly the dominant[6] term when $\ell = \mathrm{polylog}(n)$ and $np\tau^2 \gg \sqrt{np}$,[7] we get that with high probability $\|A_{\geqslant 2}\| \leqslant \operatorname{polylog} n \cdot \left(d^2 n^{2\ell} c_2^{2\ell}\right)^{1/2\ell} = nc_2 \operatorname{polylog} n = np\tau^2 \operatorname{polylog} n$, giving us the correct order of magnitude for the error.

Accounting for large separation. Recall that (though we have momentarily ignored this detail) the coefficients in the polynomial expansion of of $A_{ij}(\langle v_i, v_j\rangle) = \mathbb{1}[\langle v_i, v_j\rangle \geqslant \tau^{i,j}]$ depend on $\tau^{i,j}$ as well, so really $A_{ij} = \sum_{k=0}^{\infty} c_k^{i,j} q_k(\langle v_i, v_j\rangle)$. The coefficients $c_k^{i,j}$ concentrate well when the separation μ is small, but when μ is large the analysis is slightly more involved when we related the gaussian mixture to the sphere. We prove a similar result with an application of the trace method:

Theorem 2.3. *Suppose $d, n \in \mathbb{Z}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, and $\mu > 0$ satisfy the conditions $\log^{16} n \ll d < n$, $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, and $pn \gg 1$, we have that with high probability,*

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{op}} \ll \max\left\{\frac{n\mu^4}{\tau}, \frac{\sqrt{n}}{\sqrt{p}d\tau}\right\}\log^9(n).$$

As in the small μ case, we prove Theorem 2.3 through the following proposition, using the same strategy that applied in the small μ case.

Proposition 2.4. *Suppose $d, n \in \mathbb{Z}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, and $\mu > 0$ satisfy the conditions $\log^{16} n \ll d < n$, $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, and $pn \gg 1$, we have that there exist a length-n vector $\tilde{\mathbb{1}}_n$ and scalars p_0, $\tilde{d}$, and λ_1 to be defined later, such that with high probability,*

$$\|A - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{op}} \ll \max\left\{npd\mu^4, \sqrt{np}\right\}\log^9(n).$$

Hypothesis testing and clustering. Once we have a good approximation to the latent embedding vectors, we can use them to hypothesis test and to cluster. For clustering, we show using standard matrix concentration techniques that the top singular vector of $UU^\top$ in $\mathbb{R}^n$ must have signs which closely match the

[6]We lose polylogarithmic factors because we do not show that it is completely dominant; we suspect that a more careful argument would be able to establish full dominance and eliminate the polylogs.

[7]When $np\tau^2 \ll \sqrt{np}$, terms where indices i_s appear with high multiplicity dominate, which gives the bound $\|A_{\geqslant 2}\| \leqslant \tilde{O}(\sqrt{np})$.

cluster labeling, and therefore by a classic eigenvector perturbation argument (the Davis-Kahan theorem) the same is true of $\mathsf{U}\mathsf{U}^\top$ because $\|UU^\top - \mathsf{U}\mathsf{U}^\top\|$ is small. For hypothesis testing, we show that if μ is large enough, U and U have a spectral gap, and thus η_1 furnishes a good hypothesis test.

3 Preliminaries and notation

We use standard big-O notation: for any A_n, B_n, we use $A_n = O(B_n)$ to denote that $\lim_{n\to\infty} \frac{A_n}{B_n} < \infty$. For any A_n, B_n, we use $A_n = o(B)$ or $A_n \ll B_n$ to denote that $\lim_{n\to\infty} \frac{A_n}{B_n} = 0$. Similarly we use $A_n = \Omega(B_n)$ to denote that $\lim_{n\to\infty} \frac{A_n}{B_n} > 0$, $A_n = \omega(B_n)$ to denote that $\lim_{n\to\infty} \frac{A_n}{B_n} = \infty$, and $A_n = \Theta(B_n)$ if $A_n = O(B_n)$ and $A_n = \Omega(B_n)$. The notation $A_n = \tilde{O}(B_n)$ to denote that there exists some constant C so that $A_n \leqslant \log^C(B_n)\cdot B_n$ for large enough n. The notation $\tilde{\Omega}(\cdot), \tilde{\Theta}(\cdot)$ applies similarly.

For any A_n, B_n, we use $A_n \overset{\text{w.h.p.}}{\ll} B_n$ to denote that for every constant $\varepsilon > 0$, we have $|A_n| \leqslant \varepsilon B_n$ for large enough n with high probability.

For any $n \times n$ matrix B, we define $\operatorname{diag}(B)$ to the $n \times n$ matrix with the same diagonal as B and all zero entries off-diagonal. The notation $\|B\|$ and $\|B\|_{\text{op}}$ denote the operator norm of B.

We use $\mathbf{1}(\cdot)$ and $\mathbb{1}(\cdot)$ interchangeably to denote the indicator function.

If we don't specify in the setting, we always assume that $n, d \in \mathbb{Z}_+$, $\mu \in \mathbb{R}_+$, $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, $\log^{16} n \ll d < n$ and $pn \gg 1$. This is the regime we focus on.

Gegenbauer polynomials

For $u, w \sim \text{Unif}(\mathcal{S}^{d-1})$, we denote by $\mathcal{D}_d$ the law of $\sqrt{d} \cdot \langle u, w\rangle$ (scaling by $\sqrt{d}$ ensures that $\mathbf{E}[\langle u, w\rangle^2] = 1$). The *Gegenbauer polynomials* are an orthonormal basis for functions in $L^2([-\sqrt{d}, \sqrt{d}], \mathcal{D}_d)$. The Gegenbauer polynomials can be obtained via application of the Gram-Schmidt process to the monomial basis, so that they naturally form a sequence of polynomials $\{q_\ell^{(d)}\}_{\ell\in\mathbb{N}}$, increasing in degree so that $\deg(q_\ell^{(d)}) = \ell$. For instance, the first three Gegenbauer polynomials are

$$q_0^{(d)}(x) = 1, \quad q_1^{(d)}(x) = x, \quad \text{and} \quad q_2^{(d)}(x) = \frac{1}{\sqrt{2}}\sqrt{\frac{d+2}{d-1}}(x^2 - 1).$$

The orthogonality of the Gegenbauer polynomials is equivalent to the property that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_d}[q_\ell^{(d)}(x) q_k^{(d)}(x)] = \mathbb{1}[k = \ell].$$

The Gegenbauer polynomials are related to the *spherical harmonics*, an orthonormal basis for functions on $\mathcal{S}^{d-1}$. We write $\{\phi_{\ell,t}(u)\}_{t\in[N_\ell^{(d)}]}$ as the spherical harmonics of degree ℓ associated to u, (and we use $N_\ell^{(d)}$ to denote the cardinality of the orthonormal degree-ℓ spherical harmonics associated with a fixed vector on $\mathcal{S}^{d-1}$). It is known that

$$N_\ell^{(d)} = \frac{2\ell + d - 2}{\ell}\binom{\ell + d - 3}{\ell - 1}.$$

If $u \sim \text{Unif}(\mathcal{S}^{d-1})$, orthonormality of the spherical harmonics implies that

$$\mathop{\mathbf{E}}_{u}\left[\phi_{\ell_1,t_1}(u)\phi_{\ell_2,t_2}(u)\right] = \mathbb{1}[\ell_1 = \ell_2, t_1 = t_2],$$

for any $\ell_1, \ell_2 \in \mathbb{Z}_{\geqslant 0}$ and $t_1, t_2 \in [N_\ell^{(d)}]$.

The spherical harmonics are related to the Gegenbauer polynomials through the addition theorem

$$q_\ell^{(d)}(\sqrt{d}\cdot\langle u,v\rangle) = \frac{1}{\sqrt{N_\ell^{(d)}}}\sum_{t\in[N_\ell^{(d)}]}\phi_{\ell,t}(u)\phi_{\ell,t}(v),$$

See [Dai13, EF14] for a proof of this statement. The addition theorem and the orthonormality of the spherical harmonics together imply the following remarkable property: if $v \sim \mathrm{Unif}(\mathcal{S}^{d-1})$, then

$$\mathop{\mathbf{E}}_{v}\left[q_\ell^{(d)}(\sqrt{d}\cdot\langle u,v\rangle)\cdot q_k^{(d)}(\sqrt{d}\cdot\langle v,w\rangle)\right] = \mathbb{1}[k=\ell]\cdot\frac{1}{\sqrt{N_\ell^{(d)}}}\cdot q_\ell^{(d)}(\sqrt{d}\cdot\langle u,w\rangle).$$

Further, from the orthonormality of the spherical harmonics we can derive that

$$q_\ell^{(d)}(\sqrt{d}) = \mathop{\mathbf{E}}_{u} q_\ell^{(d)}(\sqrt{d}\langle u,u\rangle) = \frac{1}{\sqrt{N_\ell^{(d)}}}\sum_{t\in[N_\ell^{(d)}]}\mathop{\mathbf{E}}_{u}\phi_{\ell,t}(u)^2 = \sqrt{N_\ell^{(d)}}.$$

4 Proofs of the main results

4.1 Polynomial expansion of the indicator function

We will expand the threshold function $\mathbb{1}(x \geqslant \sqrt{d}\tau)$ in the basis of d-dimensional Gegenbauer polynomials. As our vectors $u_1,\dots,u_n$ are sampled from a Gaussian mixture distribution, the Gegenbauer polynomials are no longer an orthogonal basis for $\langle u_i,u_j\rangle$. To correct for this, we begin by shifting and rescaling our vectors.

For each u_i, we define the first entry of it to be a_i and the let the remaining $d-1$ coordinates form the $(d-1)$-vector w_i. Furthermore, let $\ell_i = \|w_i\|$ and write $w_i = \ell_i v_i$. We then have that $u_i = (a_i, \ell_i v_i)$ where each v_i is a unit vector. For each $i \neq j \in [n]$, the (i,j)-th entry of the adjacency matrix is now given by

$$A_{i,j} = \mathbb{1}(\langle u_i,u_j\rangle \geqslant \tau) = \mathbb{1}(a_ia_j + \ell_i\ell_j\langle v_i,v_j\rangle \geqslant \tau) = \mathbb{1}\left(\langle v_i,v_j\rangle \geqslant \frac{\tau - a_ia_j}{\ell_i\ell_j}\right). \tag{1}$$

For notational convenience, call

$$\tau^{i,j} = \frac{\tau - a_ia_j}{\ell_i\ell_j}. \tag{2}$$

We will later show that the $\tau^{i,j}$ are well-concentrated around τ, with $\tau^{i,j} \approx (\tau\pm\mu^2)(1\pm d^{-1/2})$. The dimension of the v_i is now $d-1$, and in what follows we write $\tilde{d} = d-1$ for simplicity.

We now expand the threshold function $\mathbb{1}(x \geqslant \sqrt{\tilde{d}}\tau)$ as well as the threshold function corresponding to each (i,j) entry, $\mathbb{1}(x \geqslant \sqrt{\tilde{d}}\tau^{i,j})$, in the $\tilde{d}$-dimensional sphere $\mathcal{S}^{\tilde{d}-1}$ and define

$$\mathbb{1}(x \geqslant \sqrt{\tilde{d}}\tau) = \sum_{k=0}^{\infty} c_k q_k^{(\tilde{d})}(x), \quad \text{and} \qquad \mathbb{1}(x \geqslant \sqrt{\tilde{d}}\tau^{i,j}) = \sum_{k=0}^{\infty} c_k^{i,j} q_k^{(\tilde{d})}(x).$$

Importantly, we further define

$$\lambda_k = \frac{c_k}{\sqrt{N_k}} \quad \text{and} \quad \lambda_k^{i,j} = \frac{c_k^{i,j}}{\sqrt{N_k}}, \tag{3}$$

where $N_k = N_k^{(\tilde{d})}$ is the cardinality of the orthonormal degree-k spherical harmonics associated with any fixed vector on the sphere $\mathcal{S}^{\tilde{d}-1}$, as introduced in Section 3. As a convention, we write $p_0 := \lambda_0$ and

$p_0^{i,j} := \lambda_0^{i,j}$. In the rest of this subsection and the next, we write $q_k = q_k^{(\tilde{d})}$, omitting the superscript for simplicity.

4.2 The trace method

In order to prove Proposition 2.2, we apply the trace method to A minus a linear approximation to its top eigenspace in terms of the unit-vector inner products $\langle v_i, v_j\rangle$; this will let us better exploit orthogonality properties of Gegenbauer polynomials. The constant-order term which we subtract will not be a rank-1 matrix; we'll correct for this (accounting for the difference between Gaussian and spherical vectors) later in Section 4.4.

For simplicity, we adopt a notation and write $[a_{i,j}]_{0,n\times n}$ to denote the $n\times n$ matrix where each off-diagonal entry equals $a_{i,j}$ and the diagonal equals 0.

Proposition 4.1. *For $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$, and any $i, j \in [n]$, we have that*

$$\|A - [p_0^{i,j}]_{0,n\times n} - [\tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \ll \log^9 n \max\left\{n p \tau^2, \sqrt{np}\right\},$$

with high probability as n goes to infinity.

To simplify notation, we define the $n\times n$ matrix Q to be the left hand side in Proposition 4.1. So this implies that when $i \neq j$,

$$Q_{i,j} = \mathbb{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j}) - p_0^{i,j} - \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle.$$

And $Q_{i,i} = 0$ for any $i \in [n]$. The rest of the subsection will be devoted to the proof of Proposition 4.1. Now we briefly recall the statement of the trace method.

Lemma 4.2 (Trace Method)**.** *Let M be a symmetric matrix. Then for any even integer $\ell \geqslant 0$,*

$$\mathbf{Pr}(\|M\| \geqslant e^{\varepsilon}\,\mathbf{E}[\mathrm{tr}(M^{\ell})]^{1/\ell}) \leqslant \exp(-\varepsilon\ell).$$

We will proceed to compute $\mathbf{E}[\mathrm{tr}(Q^{\ell})]$ for an even integer ℓ. This amounts to bounding the expectation of a sum over closed walks of length ℓ in the complete graph K_n when weighted by entries of Q:

$$\mathbf{E}[\mathrm{tr}(Q^{\ell})] = \sum_{i_1,\cdots,i_\ell\in[n]} \mathbf{E}\left(\prod_{t=1}^{\ell} Q_{i_t,i_{t+1}}\right),$$

where we identify $i_{\ell+1}$ with i_1. We'll associate each sequence $\vec{i} = (i_1,\cdots,i_\ell) \in [n]^\ell$ with a (multi-)graph $\boldsymbol{H}_{\vec{i}}$ (often we will drop the subscript $\vec{i}$). We define the set of vertices in $\{i_1,\cdots,i_\ell\}$ as the vertex set and put an edge between i_t and i_{t+1} for any $1 \leqslant t \leqslant \ell$, where again we identify $i_{\ell+1}$ with i_1, allowing multi-edges.

Note that the diagonal of the matrix Q consists of all zero entries, so we only need to consider multi-graphs with no self loops. Furthermore, because $i_1,\cdots,i_\ell$ is a closed walk, all vertices in $\boldsymbol{H}_{\vec{i}}$ have even degree.

When we take the expectation over the vector v_{i_j} for any degree-2 vertex $i_j \in \boldsymbol{H}$, it can be contracted at the cost of a shrinking factor:

Lemma 4.3 (Contracting degree-2 vertices)**.** *In a path $s_1,\cdots,s_{t+1}$ of length $t \geqslant 2$ in which $s_2,\ldots,s_t$ have degree 2 in $\boldsymbol{H}$, in expectation over the randomness of $v_{s_2},\ldots,v_{s_t}$ we have,*

$$\mathop{\mathbf{E}}_{v_{s_2},\ldots,v_{s_{t-1}}}\left(\prod_{a=1}^{t} Q_{s_a,s_{a+1}}\right) = \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}}\langle v_{s_1}, v_{s_{t+1}}\rangle)\left(\prod_{a=1}^{t}\lambda_k^{s_a,s_{a+1}}\right)\sqrt{N_k}$$

Proof.

$$\mathop{\mathbf{E}}_{v_{s_2},\dots,v_{s_{t-1}}}\left(\prod_{a=1}^{t} Q_{s_a,s_{a+1}} \,\middle|\, v_{s_1}, v_{s_{t+1}}\right) = \mathop{\mathbf{E}}_{v_{s_2},\dots,v_{s_{t-1}}}\left(\prod_{a=1}^{t}\left(\sum_{k=2}^{\infty}\lambda_k^{s_a,s_{a+1}}\sqrt{N_k}q_k(\sqrt{\tilde{d}}\langle v_{s_a}, v_{s_{a+1}}\rangle)\right)\right)$$
$$= \sum_{k_1,\cdots,k_t=2}^{\infty}\prod_{a=1}^{t}\lambda_{k_a}^{s_a,s_{a+1}}\sqrt{N_{k_a}}\mathop{\mathbf{E}}_{v_{s_2},\dots,v_{s_{t-1}}}\left(\prod_{a=1}^{t} q_{k_a}(\sqrt{\tilde{d}}\langle v_{s_a}, v_{s_{a+1}}\rangle)\right).$$

The exchange of the limit and the expectation is justified by the standard dominated convergence theorem. Note that by the properties of the Gegenbauer polynomials given in Section 3, the product is only nonzero when all k_a are the same, and further

$$\mathop{\mathbf{E}}_{v_{s_{a+1}}}\left(q_{k_a}(\sqrt{\tilde{d}}\langle v_{s_a}, v_{s_{a+1}}\rangle)q_{k_a}(\sqrt{\tilde{d}}\langle v_{s_{a+1}}, v_{s_{a+2}}\rangle)\right) = \frac{1}{\sqrt{N_{k_a}}}q_{k_a}(\sqrt{\tilde{d}}\langle v_{s_a}, v_{s_{a+2}}\rangle).$$

By applying the above equation repeatedly, we have the lemma. □

If we begin with a cycle $s_1, \dots, s_{t+1} = s_1$, and contract all of the degree-2 vertices in a cycle, then this produces a self-loop. So we have the following as a corollary:

Corollary 4.4 (Contracting a cycle to a self-loop)**.** *In an induced cycle $s_1, \dots, s_{t+1} = s_1$ of length $t \geqslant 2$ in $\boldsymbol{H}$, in expectation over the randomness of $v_{s_2}, \dots, v_{s_{t-1}}$, we have*

$$\mathop{\mathbf{E}}_{v_{s_2},\dots,v_{s_{t-1}}}\left(\prod_{a=1}^{t} Q_{s_a,s_{a+1}}\right) = \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}})\left(\prod_{a=1}^{t}\lambda_k^{s_a,s_{a+1}}\right)\sqrt{N_k}.$$

Proof. This follows from Lemma 4.3 and from the fact that if $v_{s_1} = v_{s_{t+1}}$ then $\langle v_{s_1}, v_{s_{t+1}}\rangle = 1$. □

We will obtain a "contracted" graph $\tilde{\boldsymbol{H}}_{\vec{i}}$ from $\boldsymbol{H}_{\vec{i}}$ as follows: as long as there exists either a vertex of degree 2, contract it; otherwise if there exits a self-loop, remove it. When the algorithm terminates we are left with the contracted (and potentially empty) graph $\tilde{\boldsymbol{H}}_{\vec{i}}$. If $\tilde{\boldsymbol{H}}_{\vec{i}}$ is nonempty, then every vertex has degree at least 4: this is because (i) every vertex had even degree to begin with, (ii) all degrees have to be larger than two for the procedure to terminate, and (iii) our algorithm for producing $\tilde{\boldsymbol{H}}_{\vec{i}}$ maintains the invariant that all degrees are even. To see (iii) is true, note that so long as we follow the convention that a self-loop induces degree two, then vertex contractions do not change the degree of the non-contracted vertices, and further when a self loop is removed from a vertex the degree drops by two, therefore maintaining the invariant that the degree is even.

For any edge $e \in E(\tilde{\boldsymbol{H}})$ which is the result of the contraction of a path $s_1, \dots, s_{t+1} \in \boldsymbol{H}$, define

$$\tilde{Q}_e = \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}}\langle v_{s_1}, v_{s_{t+1}}\rangle)\cdot\sqrt{N_k}\cdot\prod_{a=1}^{t}\lambda_k^{s_a,s_{a+1}},$$

and for the sake of consistency if $e \in E(\tilde{\boldsymbol{H}}) \cap E(\boldsymbol{H})$ then define $\tilde{Q}_e = Q_e$ in any case. Note that this is a random variable depending only on v_{s_1} and $v_{s_{t+1}}$ conditioned on the values $\lambda_k^{i,j}$.

Further, let $C(\vec{i})$ be the set of all cycles $C = (s_1, \dots, s_{t+1} = s_1) \in \boldsymbol{H}$ that were contracted into a self-loop and removed in producing $\tilde{\boldsymbol{H}}$, and define

$$\tilde{Q}_C = \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}})\cdot\sqrt{N_k}\cdot\prod_{a=1}^{t}\lambda_k^{s_a,s_{a+1}}.$$

Note that in light of Corollary 4.4 Q_C is deterministic conditioned on the values $\lambda_k^{i,j}$.

From Lemma 4.3 and Corollary 4.4, we have

$$\mathbf{E}\left(\prod_{e\in E(\boldsymbol{H})} Q_e\right) = \prod_{C\in \mathcal{C}} \tilde{Q}_C \cdot \mathbf{E}\left(\prod_{e\in E(\tilde{\boldsymbol{H}})} \tilde{Q}_e\right). \tag{4}$$

It remains to deal with the expectation over the $\tilde{Q}_e$ in $\tilde{\boldsymbol{H}}$; here we will appeal to the fact that $|\tilde{Q}_e|$ are not too large whenever the inner products $\langle v_i, v_j\rangle$ are not too large. For any vertices $i \neq j \in [n]$, define the *good* event $\mathcal{G}_{(i,j)}$

$$\mathcal{G}_{(i,j)} := \left\{ |\langle v_i, v_j\rangle| < \frac{\log^2(n)}{\sqrt{\tilde{d}}} \right\},$$

and let $\mathcal{G} = \bigcap_{i\neq j\in[n]} \mathcal{G}_{(i,j)}$ be the event that all such inner products are good. We can bound the probability of $\overline{\mathcal{G}}$ using the following lemma.

Lemma 4.5 (Bound on inner products). *For any $i \neq j \in [n]$, we have that*

$$\mathbf{Pr}(\overline{\mathcal{G}}) \leqslant n^2 \cdot n^{-\log^3(n)/3}.$$

Proof. By direct computation with the density of $\mathcal{D}_{\tilde{d}}$,

$$\mathbf{Pr}(\overline{\mathcal{G}_{(i,j)}}) = \frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} \int_{\log^2(n)}^{\sqrt{\tilde{d}}} (1 - \tfrac{\xi^2}{\tilde{d}})^{\frac{\tilde{d}-3}{2}} d\xi \leqslant \sqrt{\tilde{d}} \exp\left(-\tfrac{\log^4(n)(\tilde{d}-3)}{2\tilde{d}}\right) \ll n^{-\frac{1}{3}\log^3(n)}.$$

By a union bound, we have the lemma. □

Now, it will be useful to divide $\tilde{\boldsymbol{H}}$ into two (multi)graphs: the contracted-edges graph induced by the edges resulting from contraction, $\boldsymbol{K}_{\tilde{i}}$, and the "uncontracted graph" $\boldsymbol{U}_{\tilde{i}}$ which remains when the edges in $\boldsymbol{K}$ are removed, with vertex set $V(\tilde{\boldsymbol{H}})$ and edge set $E(\tilde{\boldsymbol{H}}) \setminus E(\boldsymbol{K})$. This is because for a contracted edge e, $\|\tilde{Q}_e \mathbf{1}(\mathcal{G}_e)\|_\infty$ decays with p, whereas for uncontracted edges in $E(\boldsymbol{U})$ we will have to argue differently. We now have that

$$\begin{aligned}
\prod_{C\in\mathcal{C}} \tilde{Q}_C \cdot \mathbf{E}\left(\prod_{e\in E(\tilde{\boldsymbol{H}}_{\tilde{i}})} \tilde{Q}_e\right) &\leqslant \prod_{C\in\mathcal{C}} \tilde{Q}_C \cdot \mathbf{E}\left(\prod_{e\in E(\tilde{\boldsymbol{H}}_{\tilde{i}})} \tilde{Q}_e \mathbf{1}(\mathcal{G})\right) + \mathbf{Pr}(\overline{\mathcal{G}}) \cdot \|Q_e\|_\infty^\ell \\
&\leqslant \left|\prod_{C\in\mathcal{C}} \tilde{Q}_C\right| \cdot \prod_{e\in E(\tilde{\boldsymbol{H}})\setminus E(\boldsymbol{H})} \|\tilde{Q}_e \mathbf{1}(\mathcal{G}_e)\|_\infty \cdot \mathbf{E}\left(\prod_{e\in E(\tilde{\boldsymbol{H}})\cap E(\boldsymbol{H})} |Q_e \mathbf{1}(\mathcal{G}_e)|\right) + \mathbf{Pr}(\overline{\mathcal{G}}) \cdot \|Q_e\|_\infty^\ell \\
&= \left|\prod_{C\in\mathcal{C}} \tilde{Q}_C\right| \cdot \prod_{e\in E(\boldsymbol{K}_{\tilde{i}})} \|\tilde{Q}_e \mathbf{1}(\mathcal{G}_e)\|_\infty \cdot \mathbf{E}\left(\prod_{e\in E(\boldsymbol{U}_{\tilde{i}})} |Q_e \mathbf{1}(\mathcal{G}_e)|\right) + n^2 \cdot n^{-\log^3 n/3} \cdot (2 + \sqrt{\tilde{d}})^\ell,
\end{aligned} \tag{5}$$

where in the last line we used Lemma 4.5 and the following claim:

Claim 4.1. *For every edge e,*

$$\|Q_e\|_\infty \leqslant 2 + \sqrt{\tilde{d}},$$

Proof. For $e = (i, j)$, expand

$$|Q_{i,j}| = \left|1 - p_0^{i,j} + \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle\right|.$$

The bound follows from the fact that $|\langle v_i, v_j\rangle| \leqslant 1$, and $|p_0^{i,j}| \leqslant 1$, $|\sqrt{\tilde{d}}\lambda_1^{i,j}| \leqslant 1$ and $|\langle v_i, v_j\rangle| \leqslant 1$ since each $(c_k^{i,j})^2 \leqslant 1$. □

Concerning the second term in Equation (5), we will ultimately choose $\ell \ll \log^3 n$, so the second term is effectively negligible. The following lemmas provide the bounds on the edge weights needed to bound the first term in Equation (5):

Lemma 4.6 (Bounds on contracted edge weights). *There exists a constant $c > 0$ so that if $d \gg \log^{16} n$, with high probability over the random variables $\{\tau^{i,j}\}_{i,j\in[n]}$, for any $e \in E(\boldsymbol{K})$ and $C \in \mathcal{C}$,*

$$\|\tilde{Q}_e \mathbf{1}(\mathcal{G}_e)\|_\infty \leqslant \log^5(n) \cdot p \cdot (cp\tau^2)^{t(e)-1}, \qquad \textit{and} \qquad \tilde{Q}_C \leqslant \log^2(n) \cdot p \cdot (cp\tau^2)^{t(C)-2},$$

where $t(e) \geqslant 2$ is the number of edges in the path that produced edge e before contraction, and $t(C) \geqslant 2$ is the number of edges in the cycle C. Further, if $p = \Omega(1/n)$, then for any uncontracted edge $e \in E(\boldsymbol{U})$,

$$|Q_e \mathbf{1}(\mathcal{G}_e)| \leqslant \mathbf{1}[e \in E(G)] + cp\log^3 n,$$

and $\mathbf{E}[\mathbf{1}[e \in E(G)]] \leqslant cp$.

We will prove Lemma 4.6 below, in Section 4.3.

Lemma 4.7 (Bound on uncontracted edges). *There exists a constant $C > 0$ such that with high probability over the random variables $\{\tau^{i,j}\}_{i,j\in[n]}$,*

$$\mathbf{E}\left(\prod_{e\in E(\boldsymbol{U})} |Q_e \mathbf{1}(\mathcal{G}_e)|\right) \leqslant (C\log^3 n)^{|E(\boldsymbol{U})|} \cdot p^{|V(\tilde{\boldsymbol{H}})|-|E(\boldsymbol{K})|-1}.$$

Proof. Choose a spanning forest F of $\boldsymbol{U}$. Then

$$\mathbf{E}\left(\prod_{e\in E(\boldsymbol{U})} |Q_e \mathbf{1}(\mathcal{G}_e)|\right) \leqslant \prod_{e\in E(\boldsymbol{U})\setminus E(F)} \|Q_e \mathbf{1}(\mathcal{G}_e)\|_\infty \cdot \mathbf{E}\prod_{e\in E(F)} |Q_e \mathbf{1}(\mathcal{G}_e)|.$$

Now, by Lemma 4.6 there exists a constant $c > 0$ so that for any e, with high probability over the randomness of the $\tau_{i,j}$,

$$\mathbf{E}\prod_{e\in E(F)} |Q_e \mathbf{1}(\mathcal{G}_e)| \leqslant \mathbf{E}\prod_{e\in E(F)} (\mathbf{1}[e \in E(G)] + cp\log^3 n).$$

Further, for any leaf (i, j) in the spanning forest F, $\mathbf{1}[(i, j) \in E(G)]$ is independent of the remaining edge indicators, and has expectation at most cp with high probability over the $\tau^{i,j}$ (by Lemma 4.6). Applying this bound inductively, peeling off the leaves one at a time, we have

$$\mathbf{E}\prod_{e\in E(F)} |Q_e \mathbf{1}(\mathcal{G}_e)| \leqslant \mathbf{E}\prod_{e\in E(F)} (\mathbf{1}[e \in E(G)] + cp\log^3 n) \leqslant (2cp\log^3 n)^{|E(F)|}.$$

Since F is a spanning forest of $\boldsymbol{U}$, we note that $|E(F)| + |E(\boldsymbol{K})| \geqslant |V(\tilde{\boldsymbol{H}})| - 1$: the union of edges of F and $\boldsymbol{K}$ form a connected graph with $|V(\tilde{\boldsymbol{H}})|$ number of vertices. So, $|E(F)| \geqslant |V(\tilde{\boldsymbol{H}})| - |E(\boldsymbol{K})| - 1$.

Combining with Lemma 4.6 to bound $\|Q_e \mathbf{1}(\mathcal{G}_e)\|_\infty \leqslant (1 + 2c)\log^3 n$ (and taking $C = 1 + 2c$) gives our conclusion. □

Applying Lemma 4.6 and Lemma 4.7 in combination with Equation (4) and Equation (5), we conclude

$$
\begin{aligned}
&\mathbf{E}\left(\prod_{e\in E(\boldsymbol{H})} Q_e\right) - n^2 \cdot n^{-\log^3 n/3}(2+\sqrt{\tilde{d}})^\ell \\
&\leqslant \prod_{C\in\mathcal{C}} \log^2(n)\cdot p(cp\tau^2)^{t(C)-2} \cdot \prod_{e\in E(\boldsymbol{K})} \log^5(n)\cdot p(cp\tau^2)^{t(e)-1}\cdot (c\log^3 n)^{|E(\boldsymbol{U})|} p^{|V(\tilde{\boldsymbol{H}})|-|E(\boldsymbol{K})|-1} \\
&\leqslant (\log^2 n)^{|\mathcal{C}|}\cdot(\log^5 n)^{|E(\boldsymbol{K})|}\cdot(c\log^3 n)^{|E(\boldsymbol{U})|}\cdot(cp\tau^2)^{\sum_{e\in E(K)} t(e)+\sum_{C\in\mathcal{C}} t(C)-2|\mathcal{C}|-|E(\boldsymbol{K})|}\cdot p^{|\mathcal{C}|+|V(\tilde{\boldsymbol{H}})|-1} \\
&\leqslant (\log^5 n)^{|E(\tilde{\boldsymbol{H}})|+|\mathcal{C}|/2}(cp\tau^2)^{\ell-2|\mathcal{C}|-|E(\tilde{\boldsymbol{H}})|}\cdot p^{|\mathcal{C}|+|V(\tilde{\boldsymbol{H}})|-1} \qquad (6)
\end{aligned}
$$

where in the final line we use that $\ell = \sum_{C\in\mathcal{C}} t(C) + \sum_{e\in E(\boldsymbol{K})} t(e) + |E(\boldsymbol{U})|$, and $|E(\tilde{\boldsymbol{H}})| = |E(\boldsymbol{U})| + |E(\boldsymbol{K})|$.

Now, we account for the number of distinct vertices in $\boldsymbol{H}_{\vec{i}}$.

Claim 4.2 (Bound on size of vertex set)**.** *We can bound the size of the vertex set of $\boldsymbol{H}_{\vec{i}}$ by*

$$|V(\boldsymbol{H}_{\vec{i}})| = \ell - |\mathcal{C}| - |E(\tilde{\boldsymbol{H}})| + |V(\tilde{\boldsymbol{H}})|$$

Proof. We charge each contracted vertex to the cycle or edge in which it was contracted during the creation of $\tilde{\boldsymbol{H}}_{\vec{i}}$ from $\boldsymbol{H}_{\vec{i}}$. In particular, when a cycle C is contracted, every vertex save for the final vertex is removed, for a total of $t(C)-1$. When a path of t edges is contracted down to a single edge, $t-1$ vertices are removed. This amounts to a total of $\sum_{C\in\mathcal{C}}(t(C)-1) + \sum_{e\in E(\tilde{H})}(t(e)-1) = \ell - |\mathcal{C}| - |E(\tilde{\boldsymbol{H}})|$ vertices removed. Finally, we account for the vertices which remain in $\tilde{\boldsymbol{H}}$. ☐

We now have all of the ingredients with which to bound the trace. We partition the sum over weighted closed walks of length ℓ in $[n]$ according to the shape of the corresponding graph $\boldsymbol{H}$. For each $\boldsymbol{H}$ resulting from a closed walk of length ℓ, let $\mathrm{num}(\boldsymbol{H})$ be the number of sequences $\vec{i} = i_1,\dots,i_\ell$ which yield the graph $\boldsymbol{H}$. For each $s,m,v\in\mathbb{N}$, let $\mathcal{H}_{s,m,v}$ be the set of all possible unlabeled graphs $\boldsymbol{H}$ resulting from a closed walk of length ℓ for which $s = |\mathcal{C}|$ cycles are removed in the process of producing $\tilde{\boldsymbol{H}}$, and in which $|E(\tilde{\boldsymbol{H}})| = m$ and $|V(\tilde{\boldsymbol{H}})| = v$. Note that $s \leqslant \ell/2$ always, since each contracted cycle uses at least two edges, and similarly $v \leqslant m/2$, because every vertex left over in $\tilde{\boldsymbol{H}}$ cannot be contracted and therefore has degree at least 4. Then

$$
\begin{aligned}
\mathbf{E}(\mathrm{tr}(Q^\ell)) &= \sum_{\vec{i}\in[n]^\ell} \mathbf{E}\left(\prod_{t=1}^{\ell} Q_{i_t,i_{t+1}}\right) \\
&= \sum_{s=0}^{\ell/2}\sum_{m=0}^{\ell-2s}\sum_{v=0}^{m/2}\sum_{H\in\mathcal{H}_{s,m,v}} \mathrm{num}(\boldsymbol{H})\cdot\mathbf{E}\left(\prod_{e\in E(\boldsymbol{H})} Q_e\right)
\end{aligned}
$$

From Claim 4.2, if $\boldsymbol{H}\in\mathcal{H}_{s,m,v}$, $\mathrm{num}(\boldsymbol{H}) \leqslant n^{\ell-s-m+v}$ since we sample vertex labels from the set $[n]$ without replacement. In combination with Equation (6) this gives

$$
\begin{aligned}
&\leqslant \sum_{s=0}^{\ell/2}\sum_{m=0}^{\ell-2s}\sum_{v=0}^{m/2}\sum_{H\in\mathcal{H}_{s,m,v}} n^{\ell-s-m+v}\left(\left(c\log^5 n\right)^{m+s/2}(cp\tau^2)^{\ell-2s-m}p^{s+v-1} + \frac{n^2(2+\sqrt{\tilde{d}})^\ell}{n^{\log^3 n/3}}\right) \\
&\leqslant \sum_{s=0}^{\ell/2}\sum_{m=0}^{\ell-2s}\sum_{v=0}^{m/2} |\mathcal{H}_{s,m,v}|\cdot\left(n\cdot\left(c\log^5 n\right)^{m+s/2}(cnp\tau^2)^{\ell-2s-m}(np)^{s+v-1} + \frac{n^2(2+\sqrt{\tilde{d}})^\ell n^\ell}{n^{\log^3 n/3}}\right)
\end{aligned}
$$

Since $|\mathcal{H}_{s,m,v}|$ is upper bounded by the number of ℓ-vertex graphs with ℓ edges, which is at most $\ell^{2\ell}$,

$$\leqslant \ell^{2\ell} \sum_{s=0}^{\ell/2} \sum_{m=0}^{\ell-2s} \sum_{v=0}^{m/2} \left(n \cdot \left(c \log^5 n\right)^{m+s/2} (cnp\tau^2)^{\ell-2s-m} (np)^{s+v-1} + \frac{n^2(2+\sqrt{\tilde{d}})^\ell n^\ell}{n^{\log^3 n/3}} \right)$$
$$\leqslant \ell^{2\ell} \sum_{s=0}^{\ell/2} \sum_{m=0}^{\ell-2s} \ell \cdot \left(n \cdot \left(c \log^5 n\right)^{m+s/2} (cnp\tau^2)^{\ell-2s-m} (np)^{s+m/2-1} + \frac{n^2(2+\sqrt{\tilde{d}})^\ell n^\ell}{n^{\log^3 n/3}} \right)$$
$$\leqslant \ell^{2\ell+1} \sum_{s=0}^{\ell/2} \sum_{m=0}^{\ell-2s} \left(n \cdot \left(\log^5 n \frac{\sqrt{np}}{np\tau^2}\right)^{m+2s} (cnp\tau^2)^\ell + \frac{n^2(2+\sqrt{\tilde{d}})^\ell n^\ell}{n^{\log^3 n/3}} \right)$$

The maximum term in the summation is achieved either at $m+2s=0$ or $m+2s=\ell$, and there are at most ℓ^2 terms, so

$$\leqslant \ell^{2\ell+3} \cdot \left(n \cdot \max\left((c\log^5 n\sqrt{np})^\ell, (cnp\tau^2)^\ell\right) + \frac{n^2(2+\sqrt{\tilde{d}})^\ell n^\ell}{n^{\log^3 n/3}} \right).$$

Choosing $\ell = \log^2 n$, the second term is dwarfed by the first, and applying the trace method, we conclude that with high probability over the $\{\tau^{i,j}\}_{i,j\in[n]}$,

$$\Pr_{v_1,\ldots,v_n}\left(\|Q\| > e^\varepsilon \cdot 2\log^9 n \cdot \max\left(c\sqrt{np}, cnp\tau^2\right)\right) \leqslant \exp(-\varepsilon \log^2 n),$$

as desired.

4.3 Controlling expansion coefficients and contracted edge weights

Above, we relied on high-probability upper bounds on the contribution that each edge could make to the weight of a walk in order to bound the contribution of walks containing vertices of degree > 2. The main purpose of this section is to prove those bounds. We start by establishing a lemma that bounds the size of τ.

Lemma 4.8. *For $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$, $d > \log^2 n$, and $p \in [0, 1/2-\varepsilon]$, there are constants $c_1 > 0$ and $C_2 > 0$ such that for all n sufficiently large,*

$$c_1\sqrt{\frac{\log(1/p)}{d}} \leqslant \tau \leqslant C_2\sqrt{\frac{\log(1/p)}{d}}.$$

Proof. For each vector u_i, we write $u_i = (\mu S_i + N_i, w_i)$, where $N_i \sim \mathcal{N}(0, 1/d)$, $S_i \sim \mathrm{Unif}(\{\pm 1\})$, and $w_i \sim \mathcal{N}(0, I_{d-1}/d)$ independently. Then p can be rewritten as

$$p = \frac{1}{2}\Pr[\mu^2 + \mu(N_i+N_j) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau] + \frac{1}{2}\Pr[-\mu^2 + \mu(N_i - N_j) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau].$$

We bound τ in terms of p. First, an upper bound on p gives us the following relation:

$$\begin{aligned} p &\leqslant \Pr[\mu^2 + \mu(N_i+N_j) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau] \\ &\leqslant 2\max\left\{\Pr[N_i + N_j \geqslant \tfrac{\tau-\mu^2}{2\mu}], \Pr[N_iN_j + \langle w_i, w_j\rangle \geqslant \tfrac{\tau-\mu^2}{2}]\right\}. \end{aligned}$$

Write $\tau_\mu := (\tau - \mu^2)/(2\mu)$. Then we note that

$$\mathbf{Pr}[N_i + N_j \geqslant \tau_\mu] = \mathbf{Pr}[\sqrt{d/2}(N_i + N_j) \geqslant \sqrt{d/2}\tau_\mu] \leqslant \exp(-\tau_\mu d/2).$$

Further, we note that $N_iN_j + \langle w_i, w_j\rangle$ can be written as the difference of two independent normalized Chi-Squared random variables $N_iN_j + \langle w_i, w_j\rangle = (A_d - B_d)/(2d)$, where $A_d, B_d \sim \chi^2_d$. By the Laurent-Massart bound [LM00], we know that $\mathbf{Pr}[A_d - d \geqslant 2\sqrt{dx} + 2x] \leqslant \exp(-x)$, $\mathbf{Pr}[A_d - d \leqslant -2\sqrt{dx}] \leqslant \exp(-x)$ and so does B_d. This implies that

$$\begin{aligned}
&\mathbf{Pr}[N_iN_j + \langle w_i, w_j\rangle \geqslant (\tau - \mu^2)/2] = \mathbf{Pr}[(A_d - d)/(2d) - (B_d - d)/(2d) \geqslant (\tau - \mu^2)/2] \\
&\leqslant \mathbf{Pr}[(A_d - d)/(2d) \geqslant (\tau - \mu^2)/4] + \mathbf{Pr}[(B_d - d)/(2d) \leqslant -(\tau - \mu^2)/4] \\
&\leqslant 2\exp(-c(\tau - \mu^2)^2 d)
\end{aligned}$$

for some constant c. Combining the two estimates, we have that

$$p \leqslant 2\max\left\{\exp(-\tau_\mu d/2), 2\exp(-c(\tau - \mu^2)^2 d)\right\}.$$

This implies that

$$(\tau - \mu^2)^2 \geqslant \frac{\log(4/p)}{cd}. \tag{7}$$

To show that $\tau - \mu^2$ cannot be negative, we lower bound p. There exists a constant $\alpha > 0$ so that for for any $C > 0$ sufficiently large,

$$\begin{aligned}
\frac{1}{2} - \varepsilon \geqslant p &\geqslant \mathbf{Pr}[-\mu^2 + \mu(N_i - N_j) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau] \\
&\geqslant \mathbf{Pr}[-\mu^2 + \mu(N_i - N_j) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau \mid |N_i|, |N_j| \leqslant \tfrac{C}{\sqrt{d}}] \cdot (1 - 2e^{-\alpha C^2})^2 \\
&\geqslant \mathbf{Pr}[\langle w_i, w_j\rangle \geqslant \tau + \mu^2 + 2\tfrac{C}{\sqrt{d}}\mu + \tfrac{C^2}{d}] \cdot (1 - 2e^{-\alpha C^2})^2 \\
&= \mathbf{Pr}[\langle w_i, w_j\rangle \geqslant \tau + (\mu + \tfrac{C}{\sqrt{d}})^2] \cdot (1 - 2e^{-\alpha C^2})^2
\end{aligned}$$

Now, choose C to be a constant large enough such that $(\frac{1}{2} - \varepsilon)/(1 - 2e^{-\alpha C^2})^2 \leqslant \frac{1}{2} - \frac{\varepsilon}{2}$. By assumption, $\mu, \frac{1}{\sqrt{d}} \leqslant \frac{1}{d^{1/4}\log^{1/2} n}$, so for d sufficiently large the above implies that

$$\frac{1}{2} - \frac{1}{2}\varepsilon \geqslant \mathbf{Pr}[\langle w_i, w_j\rangle \geqslant \tau + \tfrac{(1+C)^2}{\sqrt{d}\log n}].$$

Note that if $\tau + \frac{(1+C)^2}{\sqrt{d}\log n} \leqslant 0$, then the symmetry of w_i implies that the probability on the right-hand side is at least $\frac{1}{2}$, a contradiction. Moreover, if $\tau + \frac{(1+C)^2}{\sqrt{d}\log n} \geqslant \frac{10}{\sqrt{d}}$, then for d sufficiently large $\tau \gg \frac{1}{\sqrt{d}\log n}$ and hence $\tau \gg \mu^2$, and there is nothing left to prove. If instead $\tau + \frac{(1+C)^2}{\sqrt{d}\log n} \leqslant \frac{10}{\sqrt{d}}$, then (by an approximation of the density of $\langle w_i, w_j\rangle$) there exists a constant c_2, such that

$$\frac{1}{2} - \frac{1}{2}\varepsilon \geqslant \mathbf{Pr}[\langle w_i, w_j\rangle \geqslant \tau + (\mu + \tfrac{C}{\sqrt{d}})^2] \geqslant \frac{1}{2} - c_2\sqrt{d}(\tau + \tfrac{(1+C)^2}{\sqrt{d}\log n}) \quad \Longrightarrow \quad \tau \geqslant \frac{\varepsilon}{c_2\sqrt{d}} - \frac{(1+C)^2}{\sqrt{d}\log n} \gg \mu^2.$$

So we conclude that $\tau \gg \mu^2$. Combining with Equation (7), our conclusion holds. □

We next present a lemma that bounds the size of $\tau^{i,j}$ and $\lambda_k^{i,j}$.

Lemma 4.9. *For any $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$, $d > \log^2 n$, there exists a constant C_ε such that with high probability with respect to the randomness of $\tau^{i,j}$, the following holds uniformly,*

$$
\begin{aligned}
&\tau^{i,j} \leqslant C_\varepsilon \sqrt{\frac{\log(1/p)}{d}},\\
&\lambda_0^{i,j} = \Pr_{\xi\sim\mathcal{D}_{\tilde d}}(\xi \geqslant \sqrt{\tilde d}\tau^{i,j}) \leqslant C_\varepsilon p,\\
&\lambda_1^{i,j} = \frac{1}{\sqrt{\tilde d}} \mathop{\mathbf{E}}_{\xi\sim\mathcal{D}_{\tilde d}}[\xi \mathbb{1}(\xi \geqslant \sqrt{\tilde d}\tau^{i,j})] \leqslant C_\varepsilon p_0^{i,j}\tau^{i,j} \leqslant C_\varepsilon^2 p\tau,\\
&\lambda_2^{i,j} = \frac{1}{\tilde d - 1} \mathop{\mathbf{E}}_{\xi\sim\mathcal{D}_{\tilde d}}[(\xi^2-1)\mathbb{1}(\xi \geqslant \sqrt{\tilde d}\tau^{i,j})] \leqslant C_\varepsilon p_0^{i,j}(\tau^{i,j})^2 \leqslant C_\varepsilon^2 p\tau^2.
\end{aligned}
$$

Proof. Recall the definition of $\tau^{i,j}$ in Equation (2) and the paragraph above. We will firstly show that $p = \Theta(\Psi(\tau\sqrt{d})) = \Theta(\exp(-\tau^2 d/2)/(\tau\sqrt{d}))$, where Ψ is 1−CDF of a standard Gaussian distribution. Lemma 4.8 implies that $\tau\mu^2 = O(1/(d\sqrt{\log n}))$. Therefore, $\Theta(\Psi(\tau\sqrt{d})) = \Theta(\Psi((\tau+\mu^2)\sqrt{d})) = \Theta(\Psi((\tau-\mu^2)\sqrt{d}))$. Write $\tau_\mu := (\tau-\mu^2)(1-\frac{1}{d^{1/4}})$. By an argument similar to the proof of Lemma 4.8, we have that

$$
\begin{aligned}
p &\leqslant \Pr\left[N_i + N_j \geqslant \tfrac{\tau-\mu^2}{\mu d^{1/4}}\right] + \Pr\left[N_iN_j + \langle w_i, w_j\rangle \geqslant (\tau-\mu^2)(1-\tfrac{1}{d^{1/4}})\right]\\
&\leqslant \Psi\left(\frac{\sqrt{2d}(\tau-\mu^2)}{\mu d^{1/4}}\right) + \Pr\left[N_iN_j + \langle w_i, w_j\rangle \geqslant \tau_\mu\right]\\
&\leqslant \Psi(\tau\sqrt{d}) + \Pr\left[N_iN_j + \langle w_i, w_j\rangle \geqslant \tau_\mu\right].
\end{aligned}
$$

Now, for the second term,

$$
\begin{aligned}
&\Pr\left[N_iN_j + \langle w_i, w_j\rangle = \tau_\mu\right] \leqslant \Pr_{\xi\sim\mathcal{D}_d}\left(\xi \geqslant \sqrt{d}\tau_\mu\right)\\
&= \frac{\Gamma(d/2)}{\sqrt{d\pi}\Gamma((d-1)/2)} \int_{\sqrt{d}\tau_\mu}^{\sqrt{d}} (1-\xi^2/d)^{\frac{d-3}{2}} d\xi\\
&\leqslant \frac{\Gamma(d/2)}{\sqrt{d\pi}\Gamma((d-1)/2)} \int_{\sqrt{d}\tau_\mu}^{\sqrt{d}} \frac{\xi}{\sqrt{d}\tau_\mu}(1-\xi^2/d)^{\frac{d-3}{2}} d\xi\\
&= \frac{\Gamma(d/2)}{\sqrt{d\pi}\Gamma((d-1)/2)} \frac{d}{d-1}\frac{1}{\sqrt{d}\tau_\mu}(1-\tau_\mu^2)^{\frac{d-3}{2}}\\
&= O(\Psi(\tau_\mu\sqrt{d})) = O(\exp(-\tau^2 d/2)/(\tau\sqrt{d})).
\end{aligned}
$$

For the lower tail, we use a similar technique as in the proof of Lemma 4.8 and have that for a constant $C_2, C_3 > 0$,

$$
\begin{aligned}
p &\geqslant \frac{1}{C_2}\Pr\left[\langle w_i, w_j\rangle \geqslant \tau + O(\tfrac{1}{\sqrt{d}\log n})\right]\\
&\geqslant \frac{1}{C_3}\Pr_{\xi\sim\mathcal{D}_{d-1}}(\xi \geqslant \sqrt{d-1}\tau) = \frac{\Gamma(d-1/2)}{C_3\sqrt{(d-1)\pi}\Gamma((d-2)/2)} \int_{\sqrt{d-1}\tau}^{\sqrt{d-1}} (1-\xi^2/(d-1))^{\frac{d-4}{2}} d\xi\\
&= \Omega(\Psi(\tau\sqrt{d})) = \Omega(\exp(-\tau^2 d/2)/(\tau\sqrt{d})).
\end{aligned}
$$

We apply similar arguments to bound $\tau^{i,j}$ and $\lambda^{i,j}$. Recall that we have defined $u_i = (a_i, \ell_i v_i)$, $\tau^{i,j} = (\tau - a_i a_j)/\ell_i\ell_j$ and $\lambda_k^{i,j} = \frac{1}{\sqrt{N_k}}\mathbf{E}_{\xi\sim\mathcal{D}_{\tilde d}}[q_k^{(\tilde d)}(\xi)\mathbf{1}(\xi \geqslant \sqrt{\tilde d}\tau^{i,j})]]$. Note that as $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$, we have $a_ia_j =$

$O(1/(\sqrt{d}\log n))$ and $\ell_i = 1 + o(\log n/\sqrt{d})$ with high probability. So

$$\tau^{i,j} = \frac{\tau - a_i a_j}{\ell_i \ell_j} = \tau + O(1/(\sqrt{d}\log n)).$$

Therefore we have that

$$p_0^{i,j} = \Pr_{\xi \sim \mathcal{D}_{\tilde{d}}}(\xi \geqslant \sqrt{\tilde{d}}\tau^{i,j}) = \frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}} (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi = \Theta(\Psi(\tau^{i,j}\sqrt{d})) = \Theta(\Psi(\tau\sqrt{d})) = \Theta(p).$$

This implies that $\tau^{i,j} = \Theta(\sqrt{\log(1/p^{i,j})}/\sqrt{d})$ with high probability. Similarly we have that $p_0 = \Theta(\Psi(\tau\sqrt{d})) = \Theta(p)$. Similarly, for $\lambda_1^{i,j}$, we have an explicit formula

$$\begin{aligned}
\lambda_1^{i,j} &= \frac{1}{\sqrt{\tilde{d}}} \mathop{\mathbf{E}}_{\xi \sim \mathcal{D}_{\tilde{d}}}[\xi \mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau^{i,j})] = \frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}} \xi (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi \\
&= -\frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} \frac{\tilde{d}}{\tilde{d}-1} (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-1}{2}} \Big|_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}} = \Theta(p_0^{i,j}\tau^{i,j}) = \Theta(p\tau).
\end{aligned}$$

Similarly for $\lambda_2^{i,j}$, we also have an explicit formula

$$\begin{aligned}
\lambda_2^{i,j} &= \frac{1}{\tilde{d}-1} \mathop{\mathbf{E}}_{\xi \sim \mathcal{D}_{\tilde{d}}}[(\xi^2 - 1)\mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau^{i,j})] = \frac{\Gamma(\tilde{d}/2)}{(\tilde{d}-1)\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}} (\xi^2 - 1)(1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi \\
&= -\frac{\Gamma(\tilde{d}/2)}{(\tilde{d}-1)\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} \xi (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-1}{2}} \Big|_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}} = \Theta(p_0^{i,j}(\tau^{i,j})^2) = \Theta(p\tau^2).
\end{aligned}$$

□

To prove Lemma 4.6, we need an L_∞ bound of the normalized Gegenbauer polynomials q_k.

Lemma 4.10. *There exists a fixed constant d_0 such that for any $\tilde{d} > d_0$, $k \geqslant 0$, and any $B \geqslant 3$,*

$$\sup_{x \in [-B,B]} |q_k(x)| \leqslant B^k.$$

Proof. The Gegenbauer polynomials can be defined using the following recurrence [Dai13, EF14]:

$$x \cdot q_k(x) = a_k \cdot q_{k+1}(x) + a_{k-1} \cdot q_{k-1}(x), \quad \text{where} \quad a_k = \sqrt{\tfrac{(k+1)(\tilde{d}+k-2)\tilde{d}}{(\tilde{d}+2k)(\tilde{d}+2k-2)}}.$$

We will prove that the supremum bound inductively. One can manually verify that the lemma holds for $k = 0, 1, 2$, noting that

$$q_0(x) = 1, \quad q_1(x) = x, \quad q_2(x) = \frac{1}{\sqrt{2}}\sqrt{\tfrac{\tilde{d}+2}{\tilde{d}-1}}(x^2 - 1).$$

Now, for the inductive step, the recurrence and the inductive hypothesis imply that

$$\sup_{|x| \leqslant B} |q_{k+1}(x)| \leqslant \frac{a_{k-1}}{a_k} \sup_{|x| \leqslant B} |q_{k-1}(x)| + \frac{1}{a_k} \sup_{|x| \leqslant B} |x| \cdot |q_k(x)| \leqslant \frac{a_{k-1}}{a_k} B^{k-1} + \frac{1}{a_k} B^{k+1}.$$

We'll show that so long as $d > d_0$, $k \geqslant 2$, and $B \geqslant 3$, $\frac{a_{k-1}}{a_k B^2} + \frac{1}{a_k} \leqslant 1$, which completes the proof of the lemma. We note that when $d, k \geqslant 2$,

$$\sqrt{k+1} \geqslant a_k \geqslant \min\left\{\sqrt{\tfrac{1}{8}\tilde{d}}, \sqrt{\tfrac{1}{2}(k+1)}\right\}, \qquad \frac{a_{k-1}}{a_k} = \sqrt{\frac{k(\tilde{d}+k-3)(\tilde{d}+2k)}{(k+1)(\tilde{d}+k-2)(\tilde{d}+2k-4)}} \leqslant \sqrt{\frac{\tilde{d}+2k}{\tilde{d}+2k-4}}.$$

Therefore, for any $k \geqslant 2$ and $\tilde{d}$ larger than some constant d_0, we have that

$$\frac{1}{a_k} + \frac{a_{k-1}}{a_k B^2} \leqslant \max\left(\sqrt{\tfrac{8}{\tilde{d}}}, \sqrt{\tfrac{2}{k+1}}\right) + \frac{1}{9}\sqrt{1 + \tfrac{4}{\tilde{d}+2k-4}} \leqslant 1,$$

as desired. $\square$

Now we provide another bound which will come in useful when k is large relative to d.

Claim 4.3. *There exists a fixed constant d_0 such that for any $\tilde{d} > d_0$, $k \geqslant d^{2/3}$, and any $0 \leqslant B < \sqrt{\tilde{d}}/2$,*

$$\sup_{x \in [-B,B]} |q_k(x)| \leqslant \sqrt{N_k}\sqrt{3}\tilde{d}^{1/4}\exp(B^2/2) \cdot \frac{k\Gamma(\frac{\tilde{d}-1}{2})\Gamma(k)}{\Gamma(\frac{\tilde{d}-1}{2}+k)}.$$

Proof. This is a consequence of the connection between Gegenbauer and Jacobi polynomials, combined with known bounds on Jacobi polynomials. The Jacobi polynomials $P_k^{(\alpha,\beta)}(x)$ are defined by

$$P_k^{(\alpha,\beta)}(x) = \frac{(-1)^k}{2^k k!}(1-x)^{-\alpha}(1+x)^{-\beta}\left(\frac{d}{dx}\right)^k \left((1-x)^{\alpha+k}(1+x)^{\beta+k}\right).$$

The Gegenbauer polynomial $q_k^{(\tilde{d})}$ is proportional to the Jacobi polynomial with $\alpha = \beta = (\tilde{d}-3)/2$,

$$\begin{aligned} q_k(x) &= \frac{k+\frac{\tilde{d}}{2}-1}{\frac{\tilde{d}}{2}-1}\frac{1}{\sqrt{N_k}}\frac{\Gamma(\frac{\tilde{d}}{2}-\frac{1}{2})\Gamma(k+\tilde{d}-2)}{\Gamma(\tilde{d}-2)\Gamma(k+\frac{\tilde{d}}{2}-\frac{1}{2})}P_k^{((\tilde{d}-3)/2,(\tilde{d}-3)/2)}\left(\frac{x}{\sqrt{\tilde{d}}}\right) \\ &= \sqrt{N_k}\frac{k\Gamma(\frac{\tilde{d}-1}{2})\Gamma(k)}{\Gamma(\frac{\tilde{d}-1}{2}+k)}P_k^{((\tilde{d}-3)/2,(\tilde{d}-3)/2)}\left(\frac{x}{\sqrt{\tilde{d}}}\right), \end{aligned} \tag{8}$$

for a reference, see [Dai13], equation (B.2.1) (noting that we normalize our Gegenbauer polynomial differently). By Theorem 2 in [Kra07], we have that when $k \geqslant 6$, and $\alpha, \beta \geqslant \frac{1+\sqrt{2}}{4}$, for any $x \in [-1,1]$,

$$(1-x)^{\alpha+\frac{1}{2}}(1+x)^{\beta+\frac{1}{2}}\left(P_k^{(\alpha,\beta)}(x)\right)^2 < 3\alpha^{1/3}(1+\frac{\alpha}{k})^{1/6}.$$

This implies that for any $0 \leqslant b < 1$,

$$\sup_{x\in[-b,b]}\left|P_k^{(\tilde{d}-3)/2,(\tilde{d}-3)/2)}(x)\right| < \sqrt{\frac{3(\frac{\tilde{d}-3}{2})^{1/3}(1+\frac{\tilde{d}-3}{2k})^{1/6}}{(1-b)^{\frac{\tilde{d}}{2}-1}(1+b)^{\frac{\tilde{d}}{2}-1}}}.$$

Combining with Equation (8), we have

$$\sup_{x\in[-B,B]} |q_k(x)| < \sqrt{N_k}\frac{k\Gamma(\frac{\tilde{d}-1}{2})\Gamma(k)}{\Gamma(\frac{\tilde{d}-1}{2}+k)}\sqrt{\frac{3(\frac{\tilde{d}-3}{2})^{1/3}(1+\frac{\tilde{d}-3}{2k})^{1/6}}{\left(1-\frac{B}{\sqrt{\tilde{d}}}\right)^{\tilde{d}/2-1}\left(1+\frac{B}{\sqrt{\tilde{d}}}\right)^{\tilde{d}/2-1}}}.$$

Now for any $k \geqslant d^{2/3}$, d sufficiently large, and any $0 \leqslant B < \sqrt{\tilde{d}}/2$, we have that

$$\sqrt{\frac{3(\frac{\tilde{d}-3}{2})^{1/3}(1+\frac{\tilde{d}-3}{2k})^{1/6}}{\left(1-\frac{B}{\sqrt{\tilde{d}}}\right)^{\tilde{d}/2-1}\left(1+\frac{B}{\sqrt{\tilde{d}}}\right)^{\tilde{d}/2-1}}} \leqslant \sqrt{\frac{3(\tilde{d})^{1/3}(d^{1/3})^{1/6}}{\left(1-\frac{B^2}{\tilde{d}}\right)^{\tilde{d}/2-1}}} \leqslant \sqrt{3}\tilde{d}^{1/4}\left(1+\frac{2B^2}{\tilde{d}}\right)^{\tilde{d}/4} \leqslant \sqrt{3}\tilde{d}^{1/4}\exp(B^2/2),$$

which completes the proof. ☐

In the proof of Lemma 4.6 we also need to control the decay of λ_k. Intuitively, when k is small, λ_k should be $O(p\tau^k) = O(\text{polylog}\,\frac{1}{p}\cdot pd^{-k/2})$. The two lemmas below give a very coarse bound on the decay of λ_k for any k.

Lemma 4.11. *Suppose $p = \Omega(\frac{1}{n})$. Then for any $k \geqslant 3$, $2 \leqslant t \leqslant \log^2 n$ and any i, j, we have that*

$$\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)|\sqrt{N_k}|\lambda_k^{i,j}|^t \leqslant \log^4 n\sqrt{N_2}(C_\varepsilon p_0\tau^2)^t/k^2$$

uniformly with high probability with respect to the randomness of $\tau^{i,j}$.

Proof. We divide our proof into two cases: relatively small k and large k. For $k \leqslant \log n$, we use a more refined estimate. We note that by Rodrigues' formula for Gegenbauer polynomials [Dai13], we have that

$$q_k^{(\tilde{d})}(\xi)\cdot(1-\frac{\xi^2}{\tilde{d}})^{(\tilde{d}-3)/2} = C_{k,\tilde{d}}\left(\frac{d}{d\xi}\right)^k\left(1-\frac{\xi^2}{\tilde{d}}\right)^{k+(\tilde{d}-3)/2},$$

where

$$C_{k,\tilde{d}} = \sqrt{N_k^{(\tilde{d})}}\left(-\frac{1}{2}\right)^k\frac{\Gamma(\frac{\tilde{d}-1}{2})}{\Gamma(k+\frac{\tilde{d}-1}{2})}\sqrt{\tilde{d}}^k.$$

A similar formula holds for $k-1$ and $\tilde{d}+2$, i.e.,

$$q_{k-1}^{(\tilde{d}+2)}(\xi)\cdot(1-\frac{\xi^2}{\tilde{d}+2})^{(\tilde{d}-1)/2} = C_{k-1,\tilde{d}+2}\left(\frac{d}{d\xi}\right)^{k-1}\left(1-\frac{\xi^2}{\tilde{d}+2}\right)^{k-1+(\tilde{d}-1)/2}.$$

Applying a change of variables, the anti-derivative of $q_k^{(\tilde{d})}(\xi)\cdot(1-\xi^2/\tilde{d})^{(\tilde{d}-3)/2}$ equals

$$f_{k,\tilde{d}}(\xi) := q_{k-1}^{(\tilde{d}+2)}\left(\frac{\sqrt{\tilde{d}+2}}{\sqrt{\tilde{d}}}\xi\right)(1-\xi^2/\tilde{d})^{(\tilde{d}-1)/2}\left(\frac{\sqrt{\tilde{d}+2}}{\sqrt{\tilde{d}}}\right)^{k-1}\frac{C_{k,\tilde{d}}}{C_{k-1,\tilde{d}+2}}.$$

By Lemma 4.10, we have that

$$
\begin{aligned}
|f_{k,\tilde{d}}(\xi)| &\leqslant \left(\frac{\sqrt{\tilde{d}+2}}{\sqrt{\tilde{d}}}|\xi|\right)^{k-1}\left(1-\tfrac{\xi^2}{\tilde{d}}\right)^{(\tilde{d}-1)/2}\left(\frac{\tilde{d}+2}{\tilde{d}}\right)^{(k-1)/2}\left|\frac{C_{k\tilde{d}}}{C_{k-1,\tilde{d}+2}}\right| \\
&= \left(\frac{\tilde{d}+2}{\tilde{d}}\right)^{(k-1)/2}|\xi|^{k-1}\left(1-\tfrac{\xi^2}{\tilde{d}}\right)^{(\tilde{d}-1)/2}\frac{\tilde{d}}{\sqrt{k(\tilde{d}+k-2)(\tilde{d}-1)}} \\
&\leqslant \left|\xi\sqrt{\tfrac{\tilde{d}+2}{\tilde{d}}}\right|^{k-1}\exp\left(-\xi^2\tfrac{\tilde{d}-1}{2\tilde{d}}\right),
\end{aligned}
$$

Note that since the probability distribution of $\xi \sim \mathcal{D}_{\tilde{d}}$ equals $\frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)}(1-\xi/\tilde{d})^{(\tilde{d}-3)/2}$ and $\frac{\Gamma(d/2)}{\sqrt{d\pi}\Gamma((d-1)/2)} \to 1/\sqrt{\pi}$, this implies that

$$
\begin{aligned}
|\lambda_k^{i,j}| &= \left|\frac{1}{\sqrt{N_k}}\mathop{\mathbf{E}}_{\xi\sim\mathcal{D}_{\tilde{d}}}[q_k(\xi)\mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau^{i,j})]\right| \leqslant \left|\frac{1}{\sqrt{N_k}}(f_{k,\tilde{d}}(\sqrt{\tilde{d}}\tau^{i,j}) - f_{k,\tilde{d}}(\sqrt{\tilde{d}}))\right| \\
&\leqslant \frac{1}{\sqrt{N_k}}|\sqrt{\tilde{d}+2}\cdot\tau^{i,j}|^{k-1}\exp\left(-(\tau^{i,j})^2\tfrac{\tilde{d}-1}{2}\right).
\end{aligned}
$$

By the proof of Lemma 4.9, we have that $\sqrt{\tilde{d}+2}\cdot\tau^{i,j} \leqslant C_\varepsilon\sqrt{\log\frac{1}{p}}$ for all i,j with high probability over the randomness of $\tau^{i,j}$, and that $\exp(-(\tau^{i,j})^2(\tilde{d}-1)/2) = O(\tau^{i,j}\sqrt{\tilde{d}}p^{i,j}) = O(\tau\sqrt{d}p)$. Therefore, we have that

$$
|\lambda_k^{i,j}| \leqslant \frac{1}{\sqrt{N_k}}(C_0\sqrt{\tilde{d}+2}\cdot\tau^{i,j})^{k-1}\tau^{i,j}\sqrt{\tilde{d}}p^{i,j} \leqslant \frac{p_0}{\sqrt{N_k}}(C\tau\sqrt{d})^k \leqslant p_0\frac{\sqrt{k!}}{\sqrt{d^k}}(C\tau\sqrt{d})^k \leqslant p_0\,(C\tau)^k\sqrt{k!} \tag{9}
$$

For $k \leqslant \log n$, we have that there is an absolute constant C such that

$$
\frac{\sqrt{N_k}|\lambda_k^{i,j}|^t}{\sqrt{N_2}(C_\varepsilon p_0\tau^2)^t} \leqslant \frac{d^{k/2}}{\sqrt{k!}d}\cdot\frac{(p_0(C\tau)^k\sqrt{k!})^t}{(p_0\tau^2)^t} \leqslant (C^2 2k\sqrt{d}\tau^2)^{k-2}\cdot((C\tau)^{k-2}\sqrt{k!})^{t-2}
$$

where we have used that $k! \leqslant (2k)^{k-2}$ for $k \geqslant 3$. By Lemma 4.10, we know that $\sup_{x\in[-\log^2 n,\log^2 n]}|q_k(x)| \leqslant \log^{2k} n$. Therefore, we have

$$
\sup_{x\in[-\log^2 n,\log^2 n]}\frac{|q_k(x)|}{\log^4 n}\frac{\sqrt{N_k}|\lambda_k^{i,j}|^t}{\sqrt{N_2}(Cp_0\tau^2)^t/k^2} \leqslant \left(k^2\left(C^2 2k\sqrt{d}\tau^2\log^2 n\right)^{k-2}\right)\cdot((C\tau)^{k-2}\sqrt{k!})^{t-2}.
$$

Since $d \gg \log^{12} n$, $3 \leqslant k \leqslant \log n$, and $\tau = \Theta(\sqrt{\log n/d})$, the first term is $\ll 1$ and the latter term is at most 1 for large enough n.

For $k \geqslant \log n$, we use a more direct bound. Since the q_k form an orthonormal basis, this implies that

$$
(\lambda_k^{i,j})^2 N_k \leqslant \left\|\mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau^{i,j})\right\|_2^2 = p_0^{i,j} \leqslant 1. \tag{10}
$$

Therefore, we have that

$$
\left|\sqrt{N_k}(\lambda_k^{i,j})^t\right| \leqslant N_k^{-(t-1)/2},
$$

and so

$$\sup_{x\in[-\log^2 n,\log^2 n]} \frac{|q_k(x)|}{\log^4 n} \frac{\sqrt{N_k}|\lambda_k^{i,j}|^t}{\sqrt{N_2}(Cp_0\tau^2)^t/k^2} \leqslant k^2 \cdot \sup_{|x|\leqslant \log^2 n} |q_k(x)| \cdot \sqrt{\frac{N_k}{N_2}} \frac{1}{(\sqrt{N_k}Cp_0\tau^2)^t}.$$

We note that by definition

$$N_k = \frac{2k+\tilde{d}-2}{k}\binom{k+\tilde{d}-3}{k-1} \geqslant \frac{\tilde{d}^k}{k^k}.$$

Thus, since the N_k are non-decreasing,

$$\sqrt{N_k}p_0\tau^2 \geqslant \frac{\sqrt{N_k}}{nd} \geqslant \frac{\sqrt{N_{\lfloor \log n\rfloor}}}{nd} \geqslant \left(\frac{\tilde{d}}{\lfloor \log n\rfloor}\right)^{\lfloor \log n\rfloor/2} \frac{1}{nd} \gg 1, \tag{11}$$

since $d \gg \log^2 n$. So to prove our statement for large k, it remains to check the case when $t = 2$, i.e., we need to show that

$$\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)| \ll \log^4 n\sqrt{N_k}\sqrt{N_2}(Cp_0\tau^2)^2/k^2$$

Now, if $k \leqslant d^{2/3}$, by Lemma 4.10, then our lower bound $d \gg \log^{12} n$ implies that

$$\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)| \leqslant (\log n)^{2k} \ll \frac{d^{k/2}}{k^{k/2}n^2d^2k^2} \ll \log^4 n\sqrt{N_k}\sqrt{N_2}(Cp_0\tau^2)^2/k^2,$$

where the second inequality follows by taking log on both sides and dividing by k. If $k \geqslant d^{2/3}$, by Claim 4.3,

$$\frac{\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)|}{\sqrt{N_k}} \leqslant \sqrt{3}\tilde{d}^{1/4}\exp(\log^4 n/2) \cdot \frac{k\Gamma(\frac{\tilde{d}-1}{2})\Gamma(k)}{\Gamma(\frac{\tilde{d}-1}{2}+k)}. \tag{12}$$

By Stirling's formula,

$$\frac{k\Gamma(\frac{\tilde{d}-1}{2})\Gamma(k)}{\Gamma(\frac{\tilde{d}-1}{2}+k)} \sim \frac{\sqrt{2\pi(\frac{\tilde{d}-1}{2}\cdot k^3)}}{\sqrt{\frac{\tilde{d}-1}{2}+k}} \left(\frac{\frac{\tilde{d}-1}{2}}{\frac{\tilde{d}-1}{2}+k}\right)^{\frac{\tilde{d}-1}{2}} \left(\frac{k}{\frac{\tilde{d}-1}{2}+k}\right)^k.$$

Taking logarithm of the right hand side of Equation (12), we have that

$$\begin{aligned}
&\log\left(\frac{\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)|}{\sqrt{N_k}}\right) \\
&\leqslant O(1) + O(\log(\tilde{d})) + \log^4 n/2 + O(\log k) - \frac{\tilde{d}-1}{2}\log\left(\frac{\frac{\tilde{d}-1}{2}+k}{\frac{\tilde{d}-1}{2}}\right) - k\log\left(\frac{\frac{\tilde{d}-1}{2}+k}{k}\right) \\
&< -2\log(n) - \log(\tilde{d}) - 2\log(k),
\end{aligned}$$

where the last inequality follows for all $d \geqslant \log^{12} n$ so long as n is sufficiently large, for instance by noting that in the case $k > d^{100}$, the subtracted terms have magnitude at least $\frac{1}{3}d\log k$, whereas in the case $k \leqslant d^{100}$ the subtracted terms have magnitude at least $k\log 2 \geqslant d^{2/3}\log 2$.

This implies that

$$\sup_{x\in[-\log^2 n,\log^2 n]} |q_k(x)| \leqslant \frac{\sqrt{N_k}}{n^2dk^2} \ll \log^4 n\sqrt{N_k}\sqrt{N_2}(Cp_0\tau^2)^2/k^2,$$

which completes the proof. ☐

We will make use of a second bound for the $\lambda_k^{i,j}$:

Lemma 4.12. *Suppose $p = \Omega(\frac{1}{n})$. Then for any $k \geqslant 3$, $3 \leqslant t \leqslant \log^2 n$ and any i, j, we have that*

$$\left|q_k(\sqrt{\tilde{d}})\sqrt{N_k}(\lambda_k^{i,j})^t\right| \leqslant \tilde{d}\sqrt{N_2}(C_\varepsilon p_0\tau^2)^t/k^2$$

uniformly with high probability with respect to the randomness of $\tau^{i,j}$.

Proof. The proof follows the same strategy as Lemma 4.11. When $k \leqslant \log n$, by Equation (9) and Lemma 4.9, we have that

$$\left|q_k(\sqrt{\tilde{d}})\sqrt{N_k}(\lambda_k^{i,j})^t\right| \leqslant d^{k/2}\sqrt{N_k}p_0^t(C\tau\sqrt{\log n})^{tk} \leqslant p_0^t d^k(C\log n)^{tk/2}\tau^{tk}$$

Comparing to the desired upper bound,

$$\frac{p_0^t d^k(C\log n)^{tk/2}\tau^{tk}}{\tilde{d}\sqrt{N_2}(Cp_0\tau^2)^t/k^2} \leqslant k^2\frac{(C'\log n)^{t(k-1)}}{d^{(t/2-1)(k-2)}} = k^2\left(\frac{(C'\log n)^{\frac{t}{(t/2-1)}\frac{k-1}{k-2}}}{d}\right)^{(t/2-1)(k-2)}$$

for C' some constant, where we have used that $\tau^2 = \Theta(\log n/d)$. Given that $k \geqslant 3, t \geqslant 3$, and $d \gg \log^{16} n$, the right-hand side above is $o(1)$ and the conclusion holds.

It remains to handle the case $k \geqslant \log n$. By the orthonormality of the spherical harmonics, $q_k(\sqrt{\tilde{d}}) = \sqrt{N_k}$. So we have that by Lemma 4.10 and Equation (10),

$$\left|q_k(\sqrt{\tilde{d}})\sqrt{N_k}(\lambda_k^{i,j})^t\right| \leqslant |\lambda_k^{i,j}|^{t-2} \leqslant N_k^{-t/2+1} \leqslant \tilde{d}\sqrt{N_2}(C_\varepsilon p_0\tau^2)^t k^{-2}\frac{N_k^{-t/2+1}}{\tilde{d}\sqrt{N_2}(C_\varepsilon p_0\tau^2)^t k^{-2}}.$$

To bound the fraction, by Equation (11), it is enough to check the case when $t = 3$. Indeed, in such situation, the fraction is bounded by

$$\frac{N_k^{-1/2}k^2}{d^2(Cp_0\tau^2)^3} \leqslant \frac{N_{\lfloor\log n\rfloor}^{-1/2}\log^2 n}{d^2(Cp_0\tau^2)^3} \leqslant \frac{(C\log n)^{\log n/2+2}n^3}{d^{\log n/2}} \ll 1.$$

☐

Combining Lemmas 4.9, 4.10, and 4.11, we now prove Lemma 4.6.

Proof of Lemma 4.6. For any edge $e \in E(\tilde{G}_{\vec{i}})$ which is the result of the contraction of a path $s_1, \dots, s_{t+1} \in G_{\vec{i}}$, recall that by definition, we have

$$\|\tilde{Q}_e\mathbf{1}(\mathcal{G}_e)\|_\infty = \left\|\sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}}\langle v_{s_1}, v_{s_{t+1}}\rangle)\cdot\mathbf{1}(\mathcal{G}_{(s_1,s_{t+1})})\cdot\sqrt{N_k}\cdot\prod_{a=1}^{t}\lambda_k^{s_a,s_{a+1}}\right\|_\infty$$

$$= \left\| \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}}\langle v_{s_1}, v_{s_{t+1}}\rangle) \cdot \mathbf{1}\left(|\langle v_{s_1}, v_{s_{t+1}}\rangle| < \frac{\log^2(n)}{\sqrt{\tilde{d}}} \right) \cdot \sqrt{N_k} \cdot \prod_{a=1}^{t} \lambda_k^{s_a,s_{a+1}} \right\|_\infty .$$

By Lemma 4.10 (taking B to be $\log^2(n)$) and Lemma 4.11, we have that for $t \geqslant 2$,

$$\begin{aligned}
\|\tilde{Q}_e \mathbf{1}(\mathcal{G}_e)\|_\infty &\leqslant \left(\log^2(n)\right)^2 \tilde{d} \prod_{a=1}^{t} \left(C_\varepsilon p_0^{s_a,s_{a+1}} (\tau^{s_a,s_{a+1}})^2 \right) + \sum_{k=3}^{\infty} \left(\log^2(n)\right)^2 \sqrt{N_2} (C_\varepsilon p_0 \tau^2)^t k^{-2} \\
&\leqslant \log^4(n) \tilde{d} (C''_{\varepsilon,3} p_0 \tau^2)^t,
\end{aligned}$$

uniformly with high probability with respect to the randomness of $\tau^{i,j}$. The bound as stated in Lemma 4.6 follows by applying Lemma 4.9 to eliminate the factor $\tilde{d}$ at the cost of a factor $\log \frac{1}{p}$. When $t = 1$, then $\tilde{Q}_e = Q_e$. Write $e = (i, j)$, then we have

$$\begin{aligned}
&|Q_e \mathbf{1}(\mathcal{G}_e)| \leqslant \mathbf{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j}) + p_0^{i,j} + |\tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle \mathbf{1}(\mathcal{G}_e)| \\
&\leqslant \mathbf{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j}) + C_\varepsilon p_0 + C_\varepsilon \tilde{d} p_0 \tau \frac{\log^2(n)}{\sqrt{\tilde{d}}} \leqslant \mathbf{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j}) + 2C_\varepsilon \log^3(n) p_0.
\end{aligned}$$

where the second-to-last inequality follows from Lemma 4.9. Furthermore, $\mathbf{E}[\mathbf{1}[\langle v_i, v_j\rangle \geqslant \tau^{i,j}]] \leqslant C_\varepsilon p_0$.

Now for any cycle $C = (s_1, \dots, s_{t+1} = s_1) \in \boldsymbol{G}$ that was contracted into a self-loop and removed in producing $\tilde{\boldsymbol{G}}$, recall that we have

$$\tilde{Q}_C = \sum_{k=2}^{\infty} q_k(\sqrt{\tilde{d}}) \cdot \sqrt{N_k} \cdot \prod_{a=1}^{t} \lambda_k^{s_a,s_{a+1}}.$$

For $t \geqslant 3$, by Lemma 4.12, we have that

$$\begin{aligned}
\tilde{Q}_C &\leqslant \tilde{d}\sqrt{N_2}(C_\varepsilon p \tau^2)^{t(C)} + \sum_{k=3}^{\infty} \tilde{d}\sqrt{N_2}(C_\varepsilon p_0 \tau^2)^{t(C)} k^{-2} \leqslant C\tilde{d}\sqrt{N_2}(C_\varepsilon p \tau^2)^{t(C)} \\
&\leqslant (p^2 \tau^4 \tilde{d} \sqrt{N_2}) \cdot (c_2 p \tau^2)^{t(C)-2} \leqslant (\log^2 n) p \cdot (c_3 p \tau^2)^{t(C)-2}.
\end{aligned}$$

When $t = 2$, we write $e = (s_1, s_2)$, then $\tilde{Q}_C = \mathbf{E}(Q_e^2)$. Therefore, by Lemma 4.9, we have that

$$\begin{aligned}
\tilde{Q}_C &= \mathbf{E}\left(\mathbf{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j}) - p_0^{i,j} - \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle \right)^2 \\
&= p_0^{i,j} + (p_0^{i,j})^2 + \tilde{d}^2 (\lambda_1^{i,j})^2 \, \mathbf{E}(\langle v_i, v_j\rangle^2) - 2(p_0^{i,j})^2 - 2\tilde{d}\lambda_1^{i,j} \, \mathbf{E}(\langle v_i, v_j\rangle \mathbf{1}(\langle v_i, v_j\rangle \geqslant \tau^{i,j})) \\
&\leqslant Cp + Cp^2 \log(1/p) \leqslant C_2 p.
\end{aligned}$$

Therefore, we have

$$\tilde{Q}_C \leqslant \log^2(n) \cdot p \cdot (cp\tau^2)^{t(C)-2},$$

where $t(C)$ is the number of edges in the cycle C. □

4.4 Relating the Gaussian mixture and the sphere

In this section, we will relate the matrix we subtracted from A in Section 4.2 to the matrix we wish to show is close to the top eigenspace of A, $p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top$. Recall the definition of ℓ_k, v_k, a_k and $\tau^{i,j}$ in Equation (1). First, we finally define $\tilde{\mathbb{1}}_n$. For each $k \in [n]$, define $L_k := \ell_k - 1$, and define $\tilde{\mathbb{1}}_n$ to be a length-n

vector where each entry equals

$$(\tilde{\mathbb{1}}_n)_k = 1 + \frac{L_k \lambda_1 \tilde{d}\tau}{p_0}.$$

We prove the following lemma and another lemma (Lemma 4.17) designed for larger μ. Let V be the n by d matrix where the i-th row equals v_i.

Lemma 4.13. *For $\mu \leqslant \tau$, we have that*

$$\|p_0^{i,j} \mathbb{1}_n \mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j} VV^\top - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top) - \mathrm{diag}\,\|_{\mathrm{op}} \ll np\tau^2 \log^4(n),$$

with high probability. Here diag *is the diagonal of* $p_0^{i,j} \mathbb{1}_n \mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j} VV^\top - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top)$.

Definition 4.1. We define (and recall that)

$$b_i := a_i/\sqrt{\tau}, \quad L_i := \ell_i - 1, \quad p_0 = \Pr_{\xi \in \mathcal{D}_{\tilde{d}}} (\xi \geqslant \sqrt{\tilde{d}}\tau), \quad \lambda_1 = \frac{1}{\sqrt{\tilde{d}}} \mathop{\mathbf{E}}_{\xi \in \mathcal{D}_{\tilde{d}}} [\xi \mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau)].$$

We further define the change of $\tau^{i,j}$, $p_0^{i,j}$ and $\lambda_1^{i,j}$ as follows.

$$\Delta\tau^{i,j} := \tau^{i,j} - \tau, \qquad \Delta p_0^{i,j} := p_0^{i,j} - p_0, \qquad \Delta\lambda_1^{i,j} := \lambda_1^{i,j} - \lambda_1.$$

To prove Lemma 4.13, we firstly prove a lemma bounding the fluctuations of $\tau^{i,j}$, $p_0^{i,j}$ and $\lambda_1^{i,j}$.

Lemma 4.14. *For $\mu \leqslant \tau$, the following holds uniformly with high probability*

$$\begin{aligned}
&L_i = o(\tfrac{\log n}{\sqrt{d}}), \quad \ell_i^{-1} = 1 - L_i + o(\tfrac{\log^2 n}{d}), \quad b_i b_j = o(\tfrac{\log n}{\sqrt{d}}), \quad \langle v_i, v_j \rangle = o(\tfrac{\log n}{\sqrt{d}}),\\
&\Delta\tau^{i,j} = -\tau(b_i b_j + L_i + L_j + o(\tfrac{\log^2 n}{d})),\\
&\Delta p_0^{i,j} = -\lambda_1 \tilde{d} \Delta\tau^{i,j}(1 + o(\tau^2 \sqrt{d} \log n)) = \lambda_1 \tilde{d}\tau(b_i b_j + L_i + L_j + o(\tfrac{\log^2 n}{d})),\\
&\Delta\lambda_1^{i,j} = -\lambda_1 \tilde{d}\tau \Delta\tau^{i,j}(1 + o(\tau^2 \sqrt{d} \log n)).
\end{aligned}$$

Proof. We firstly note that λ_1 can be computed explicitly as follows.

$$\begin{aligned}
\lambda_1 &= \frac{1}{\sqrt{\tilde{d}}} \mathop{\mathbf{E}}_{\xi \sim \mathcal{D}_{\tilde{d}}} [\xi \mathbb{1}(\xi \geqslant \sqrt{\tilde{d}}\tau)] = \frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau}^{\sqrt{\tilde{d}}} \xi (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi \\
&= -\frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} \frac{\tilde{d}}{\tilde{d}-1} (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-1}{2}} \Big|_{\sqrt{\tilde{d}}\tau}^{\sqrt{\tilde{d}}} = \frac{\Gamma(\tilde{d}/2)}{(\tilde{d}-1)\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} (1-\tau^2)^{\frac{\tilde{d}-1}{2}}.
\end{aligned}$$

The bounds on L_i, ℓ_i^{-1}, $b_i b_j$, $\langle v_i, v_j \rangle$, and $\Delta\tau^{i,j}$ follows directly from concentration inequalities and definitions. For $\Delta p_0^{i,j}$, we note that

$$\begin{aligned}
\Delta p_0^{i,j} &= \Pr_{\xi \sim \mathcal{D}_{\tilde{d}}} [\xi \in (\sqrt{\tilde{d}}\tau^{i,j}, \sqrt{\tilde{d}}\tau)] = \frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}\tau} (1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi \\
&= -\Delta\tau^{i,j} \sqrt{\tilde{d}} \frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)} (1-\tau^2)^{\frac{\tilde{d}-3}{2}} (1 + o(\tau^2 \sqrt{d} \log n)) = -\lambda_1 \tilde{d} \Delta\tau^{i,j}(1 + o(\tau^2 \sqrt{d} \log n)).
\end{aligned}$$

Similarly, for $\Delta\lambda_1^{i,j}$, we have that

$$\Delta\lambda_1^{i,j} = \frac{1}{\sqrt{\tilde{d}}} \mathop{\mathbf{E}}_{\xi\sim\mathcal{D}_{\tilde{d}}}[\xi\mathbf{1}\left(\xi \in (\sqrt{\tilde{d}}\tau^{i,j}, \sqrt{\tilde{d}}\tau)\right)] = \frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)} \int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}\tau} \xi(1-\xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}} d\xi$$

$$= -\Delta\tau^{i,j}\tilde{d}\frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)}\tau(1-\tau^2)^{\frac{\tilde{d}-3}{2}}(1+o(\tau^2\sqrt{d}\log n)) = -\lambda_1\tilde{d}\tau\Delta\tau^{i,j}(1+o(\tau^2\sqrt{d}\log n)). \qquad \square$$

Now we recall two facts that are useful in our proofs.

Fact 4.15. *For $M \in \mathbb{R}^{n\times n}$, define $\|M\|_\infty = \max_{i,j\in[n]} |M_{ij}|$. Then $\|M\|_{\mathrm{op}} \leqslant n\|M\|_\infty$.*

Proof. This is because $\|M\|_{\mathrm{op}} \leqslant \|M\|_{\mathrm{F}} \leqslant n\|M\|_\infty$. $\square$

Fact 4.16. *Let $w_1, \cdots, w_n \in \mathcal{S}^{d-1}$ be uniform random vectors in $\mathcal{S}^{d-1}$, with $n > d$. Then their gram matrix with the diagonal set to zero satisfies $\|[\langle w_i, w_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \leqslant O\left(\frac{n}{d}\right)$, with high probability.*

The proof of Fact 4.16 follows from standard matrix concentration results (see e.g. [Ver18], Theorem 4.6.1 and Theorem 3.4.6). Now we are ready to prove Lemma 4.13.

Proof of Lemma 4.13. We firstly note that $p_0 = \Theta(p)$ and $\lambda_1 = \Theta(p\tau)$. This follows from the proof of Lemma 4.9. For each $i, j \in [n]$ with $i \neq j$, we have that by Lemma 4.14,

$$\begin{aligned}
&p_0^{i,j} + \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle - p_0(1 + L_i\lambda_1\tilde{d}\tau/p_0)(1 + L_j\lambda_1\tilde{d}\tau/p_0) - \tilde{d}\lambda_1\langle u_i, u_j\rangle \\
&= (\Delta p_0^{i,j} + p) + \tilde{d}(\Delta\lambda_1^{i,j} + \lambda_1)\langle v_i, v_j\rangle - p_0(1 + L_i\lambda_1\tilde{d}\tau/p_0)(1 + L_j\lambda_1\tilde{d}\tau/p_0) - \tilde{d}\lambda_1(\tau b_i b_j + \ell_i\ell_j\langle v_i, v_j\rangle) \\
&= -\lambda_1\tilde{d}\tau o(\tfrac{\log^2 n}{d}) - \lambda_1^2\tau^2\tilde{d}^2 L_i L_j/p_0 + \tilde{d}\Delta\lambda_1\langle v_i, v_j\rangle - \tilde{d}\lambda_1(\ell_i\ell_j - 1)\langle v_i, v_j\rangle \\
&=: f_1(i,j) + f_2(i,j) + f_3(i,j) + f_4(i,j).
\end{aligned}$$

For simplicity, again, we adopt a notation and write $[a_{i,j}]_{0,n\times n}$ to denote the $n\times n$ matrix where each off-diagonal entry equals $a_{i,j}$ and the diagonal equals 0. Similarly, we write $[a_{i,j}]_{n\times n}$ as the $n\times n$ matrix where each entry equals $a_{i,j}$. Then we can rewrite our goal as to show that

$$\|[p_0^{i,j} - \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle - p_0(1 + L_i\lambda_1\tilde{d}\tau/p_0)(1 + L_j\lambda_1\tilde{d}\tau/p_0) + \tilde{d}\lambda_1\langle u_i, u_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \ll np\tau^2\log^4(n).$$

Now according to the above computation, the left hand side can be reduced to

$$\|[f_1(i,j)]_{0,n\times n} + [f_2(i,j)]_{0,n\times n} + [f_3(i,j)]_{0,n\times n} + [f_4(i,j)]_{0,n\times n}\|_{\mathrm{op}}.$$

Note that by Lemma 4.14 and Fact 4.15, we have that

$$\|[f_1(i,j)]_{0,n\times n}\|_{\mathrm{op}} \leqslant n\lambda_1\tilde{d}\tau o(\tfrac{\log^2 n}{d}) \leqslant o\left(np\tau^2\log^2 n\right).$$

Similarly, by Lemma 4.14 and Fact 4.15, we have that

$$\|[f_2(i,j)]_{0,n\times n}\|_{\mathrm{op}} \leqslant n\frac{\lambda_1^2\tau^2\tilde{d}^2}{p_0}o(\tfrac{\log^2 n}{d}) \leqslant o\left(np\tau^4 d\log^2 n\right).$$

Furthermore, by Lemma 4.14, Fact 4.15, and Fact 4.16, we have that

$$\begin{aligned}
\|[f_3(i,j)]_{0,n\times n}\|_{\mathrm{op}} &= \|[\lambda_1\tilde{d}^2\tau^2(b_ib_j + L_i + L_j + o(\tfrac{\log^2 n}{d}))\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant \|[\lambda_1\tilde{d}^2\tau^2 b_ib_j\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + 2\|[\lambda_1\tilde{d}^2\tau^2 L_i\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + n\lambda_1\tilde{d}^2\tau^2 o(\tfrac{\log^2 n}{d})o(\tfrac{\log n}{\sqrt{d}})
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \lambda_1 \tilde{d}^2\tau^2 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{d}\right) + \lambda_1 \tilde{d}^2\tau^2 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{d}\right) + n\lambda_1 \tilde{d}^2\tau^2 o(\tfrac{\log^2 n}{d}) o(\tfrac{\log n}{\sqrt{d}}) \\
&\leqslant o\left(n\lambda_1 \tilde{d}^2\tau^2 \tfrac{\log^3 n}{d\sqrt{d}}\right) \leqslant o\left(np\sqrt{d}\tau^3 \log^3 n\right).
\end{aligned}
$$

Finally, by Lemma 4.14, Fact 4.15, and Fact 4.16, we have that

$$
\begin{aligned}
&\|[f_4(i,j)]_{0,n\times n}\|_{\mathrm{op}} = \|[\tilde{d}\lambda_1(\ell_i\ell_j - 1)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} = \|[\tilde{d}\lambda_1(L_i + L_j + o(\tfrac{\log^2 n}{d}))\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&= \|[\tilde{d}\lambda_1(L_i + L_j)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + \|[\tilde{d}\lambda_1 o(\tfrac{\log^2 n}{d})\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant \tilde{d}\lambda_1 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{d}\right) + n\tilde{d}\lambda_1 o\left(\tfrac{\log^2 n}{d}\right) o\left(\tfrac{\log n}{\sqrt{d}}\right) \leqslant o\left(np\tau \log^3 n/\sqrt{d}\right).
\end{aligned}
$$

Combining the bounds for f_1, f_2, f_3, and f_4 together, we have that

$$
\begin{aligned}
&\|[f_1(i,j)]_{0,n\times n} + [f_2(i,j)]_{0,n\times n} + [f_3(i,j)]_{0,n\times n} + [f_4(i,j)]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant o\left(np\sqrt{d}\tau^3 \log^3 n\right) \leqslant o\left(np\tau^2 \log^4 n\right).
\end{aligned}
$$

□

In the rest of the subsection, we will prove the following lemma for relatively larger μ.

Lemma 4.17. *For $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, we have that*

$$
\|p_0^{i,j}\mathbb{1}_n\mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j}VV^\top - (p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top) - \mathrm{diag}\,\|_{\mathrm{op}} \ll npd\mu^4\log^5(n),
$$

with high probability. Here diag *is the diagonal of the matrix on the left hand side.*

Adopting the same notation, we have the following lemma.

Lemma 4.18. *For $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, the following holds uniformly with high probability*

$$
\begin{aligned}
&L_i = o(\tfrac{\log n}{\sqrt{d}}), \quad \ell_i^{-1} = 1 - L_i + o(\tfrac{\log^2 n}{d}), \quad a_i a_j = O(\mu^2), \quad \langle v_i, v_j\rangle = o(\tfrac{\log n}{\sqrt{d}}), \\
&\Delta\tau^{i,j} = -a_i a_j(1 - L_i - L_j + o(\tfrac{\log^2 n}{d})) - \tau(L_i + L_j + o(\tfrac{\log^2 n}{d})), \\
&\Delta p_0^{i,j} = -\lambda_1 \tilde{d}\Delta\tau^{i,j}(1 + O(\Delta\tau^{i,j}\tau d)) \\
&\qquad = \lambda_1\tilde{d}\left(a_i a_j(1 - L_i - L_j) + \tau(L_i + L_j) + o(\tau^3\log^2 n) + O(\mu^4\tau d)\right), \\
&\Delta\lambda_1^{i,j} = -\lambda_1\tilde{d}\tau\Delta\tau^{i,j}(1 + O(\Delta\tau^{i,j}\tau d)) \\
&\qquad = \lambda_1\tilde{d}\tau\left(a_i a_j(1 - L_i - L_j) + \tau(L_i + L_j) + o(\tau^3\log^2 n) + O(\mu^4\tau d)\right).
\end{aligned}
$$

Proof. The bounds on L_i, ℓ_i^{-1}, $b_i b_j$, $\langle v_i, v_j\rangle$, and $\Delta\tau^{i,j}$ follows directly from concentration inequalities and definitions. For $\Delta p_0^{i,j}$, we note that

$$
\begin{aligned}
\Delta p_0^{i,j} &= \Pr_{\xi\sim\mathcal{D}_{\tilde{d}}}[\xi \in (\sqrt{\tilde{d}}\tau^{i,j}, \sqrt{\tilde{d}}\tau)] = \frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)}\int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}\tau}(1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}}d\xi \\
&= -\Delta\tau^{i,j}\sqrt{\tilde{d}}\frac{\Gamma(\tilde{d}/2)}{\sqrt{\tilde{d}\pi}\Gamma((\tilde{d}-1)/2)}(1-\tau^2)^{\frac{\tilde{d}-3}{2}}(1 + O(\Delta\tau^{i,j}\tau d)) = -\lambda_1\tilde{d}\Delta\tau^{i,j}(1 + O(\Delta\tau^{i,j}\tau d)).
\end{aligned}
$$

Similarly, for $\Delta\lambda_1^{i,j}$, we have that

$$
\Delta\lambda_1^{i,j} = \frac{1}{\sqrt{\tilde{d}}}\mathop{\mathbf{E}}_{\xi\sim\mathcal{D}_{\tilde{d}}}[\xi\mathbf{1}\left(\xi \in (\sqrt{\tilde{d}}\tau^{i,j}, \sqrt{\tilde{d}}\tau)\right)] = \frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)}\int_{\sqrt{\tilde{d}}\tau^{i,j}}^{\sqrt{\tilde{d}}\tau}\xi(1 - \xi^2/\tilde{d})^{\frac{\tilde{d}-3}{2}}d\xi
$$

$$= -\Delta\tau^{i,j}\tilde{d}\frac{\Gamma(\tilde{d}/2)}{\tilde{d}\sqrt{\pi}\Gamma((\tilde{d}-1)/2)}\tau(1-\tau^2)^{\frac{\tilde{d}-3}{2}}(1+O(\Delta\tau^{i,j}\tau d)) = -\lambda_1\tilde{d}\tau\Delta\tau^{i,j}(1+O(\Delta\tau^{i,j}\tau d)). \qquad \square$$

Proof of Lemma 4.17. For each $i, j \in [n]$ with $i \neq j$, we have that by Lemma 4.18

$$\begin{aligned}
&p_0^{i,j} + \tilde{d}\lambda_1^{i,j}\langle v_i, v_j\rangle - p_0(1+L_i\lambda_1\tilde{d}\tau/p_0)(1+L_j\lambda_1\tilde{d}\tau/p_0) - \tilde{d}\lambda_1\langle u_i, u_j\rangle \\
&= (\Delta p_0^{i,j} + p_0) + \tilde{d}(\Delta\lambda_1^{i,j} + \lambda_1)\langle v_i, v_j\rangle - p_0(1+L_i\lambda_1\tilde{d}\tau/p_0)(1+L_j\lambda_1\tilde{d}\tau/p_0) - \tilde{d}\lambda_1(a_ia_j + \ell_i\ell_j\langle v_i, v_j\rangle) \\
&= \tilde{d}\lambda_1(o(\tau^3\log^2 n) + O(\mu^4\tau d)) - \lambda_1^2\tau^2\tilde{d}^2L_iL_j/p_0 + \tilde{d}\Delta\lambda_1\langle v_i, v_j\rangle - \tilde{d}\lambda_1(\ell_i\ell_j - 1)\langle v_i, v_j\rangle \\
&=: f_1(i,j) + f_2(i,j) + f_3(i,j) + f_4(i,j).
\end{aligned}$$

We follow the same proof strategy as in Lemma 4.13, with slightly different bounds. Note that by Lemma 4.14 and Fact 4.15, we have that

$$\|[f_1(i,j)]_{0,n\times n}\|_{\mathrm{op}} \leqslant n\tilde{d}\lambda_1 o(\tau^3\log^2 n) + n\tilde{d}\lambda_1 O(\mu^4\tau d) \leqslant o\left(npd\tau^4\log^2 n\right) + O(npd^2\tau^2\mu^4).$$

Similarly, by Lemma 4.14 and Fact 4.15, we have that

$$\|[f_2(i,j)]_{0,n\times n}\|_{\mathrm{op}} \leqslant n\frac{\lambda_1^2\tau^2\tilde{d}^2}{p_0}o(\tfrac{\log^2 n}{d}) \leqslant o\left(np\tau^4 d\log^2 n\right).$$

Furthermore, by Lemma 4.14, Fact 4.15, and Fact 4.16, we have that

$$\begin{aligned}
&\|[f_3(i,j)]_{0,n\times n}\|_{\mathrm{op}} = \|[\lambda_1\tilde{d}^2\tau\left(a_ia_j(1-L_i-L_j) + \tau(L_i+L_j) + o(\tau^3\log^2 n) + O(\mu^4\tau d)\right)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant \|[\lambda_1\tilde{d}^2\tau a_ia_j\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + 2\|[\lambda_1\tilde{d}^2\tau a_ia_jL_i\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + 2\|[\lambda_1\tilde{d}^2\tau^2L_i\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&\qquad + n\lambda_1\tilde{d}^2\tau o(\tau^3\log^2 n)o(\tfrac{\log n}{\sqrt{d}}) + n\lambda_1\tilde{d}^2\tau o(\mu^4\tau d)O(\tfrac{\log n}{\sqrt{d}}) \\
&\leqslant \lambda_1\tilde{d}^2\tau\mu^2 O\left(\tfrac{n}{d}\right) + \lambda_1\tilde{d}^2\tau\mu^2 o\left(\tfrac{\log n}{\sqrt{d}}\right)O\left(\tfrac{n}{d}\right) + \lambda_1\tilde{d}^2\tau^2 o\left(\tfrac{\log n}{\sqrt{d}}\right)O\left(\tfrac{n}{d}\right) \\
&\qquad + n\lambda_1\tilde{d}^2\tau o(\tau^3\log^2 n)o(\tfrac{\log n}{\sqrt{d}}) + n\lambda_1\tilde{d}^2\tau o(\mu^4\tau d)O(\tfrac{\log n}{\sqrt{d}}) \\
&\leqslant o\left(np\tau^5 d^{3/2}\log^3 n + np\tau^3 d^{5/2}\mu^4\right) = o\left(np\tau^2\log^{9/2} n + npd\mu^4\log^{5/2} n\right).
\end{aligned}$$

Finally, by Lemma 4.14, Fact 4.15, and Fact 4.16, we have that

$$\begin{aligned}
&\|[f_4(i,j)]_{0,n\times n}\|_{\mathrm{op}} = \|[\tilde{d}\lambda_1(\ell_i\ell_j - 1)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} = \|[\tilde{d}\lambda_1(L_i + L_j + o(\tfrac{\log^2 n}{d}))\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&= \|[\tilde{d}\lambda_1(L_i+L_j)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} + \|[\tilde{d}\lambda_1 o(\tfrac{\log^2 n}{d})\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant \tilde{d}\lambda_1 o\left(\tfrac{\log n}{\sqrt{d}}\right)O\left(\tfrac{n}{d}\right) + n\tilde{d}\lambda_1 o\left(\tfrac{\log^2 n}{d}\right)o\left(\tfrac{\log n}{\sqrt{d}}\right) \leqslant o\left(np\tau\log^3 n/\sqrt{d}\right).
\end{aligned}$$

Combining the bounds for f_1, f_2, f_3, and f_4 together, we have that

$$\begin{aligned}
&\|[f_1(i,j)]_{0,n\times n} + [f_2(i,j)]_{0,n\times n} + [f_3(i,j)]_{0,n\times n} + [f_4(i,j)]_{0,n\times n}\|_{\mathrm{op}} \\
&\leqslant o\left(np\tau^2\log^{9/2} n + npd\mu^4\log^{5/2} n\right) = o(npd\mu^4\log^5 n). \qquad \square
\end{aligned}$$

4.5 Accounting for the diagonal

In this section, we will prove Proposition 2.2 by combining Proposition 4.1 with Lemma 4.13, and prove Proposition 2.4 by combining Proposition 4.1 with Lemma 4.17. Directly combining Proposition 4.1 with

Lemma 4.13, we have that for $\mu \leqslant \tau$,

$$\|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top) - \operatorname{diag}\|_{\mathrm{op}} \overset{\text{w.h.p.}}{\ll} \log^9(n) \max(np\tau^2, \sqrt{np}).$$

And similarly, directly combining Proposition 4.1 with Lemma 4.17, we have that for $\tau < \mu \leqslant d^{-1/4} \log^{-1/2}(n)$,

$$\|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top) - \operatorname{diag}\|_{\mathrm{op}} \overset{\text{w.h.p.}}{\ll} \log^9(n) \max(npd\mu^4, \sqrt{np}).$$

The remaining of the task is to bound the diagonal of $A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top)$. Note that we have that with high probability,

$$p_0(\tilde{\mathbb{1}}_n)_i^2 = p_0(1 + L_i \lambda_1 \tilde{d}\tau / p_0)^2 = O(p_0).$$

Furthermore, with high probability,

$$\tilde{d}\lambda_1 \langle u_i, u_i \rangle = \tilde{d}\lambda_1 (b_i^2 + \ell_i^2) = O(dp\tau).$$

Define diag to be the diagonal (matrix) of $A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top)$. Therefore, for $\mu \leqslant \tau$, with high probability,

$$\begin{aligned}
&\|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top)\|_{\mathrm{op}} \\
&\leqslant \|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top) - \operatorname{diag}\|_{\mathrm{op}} + \|\operatorname{diag}(A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top))\|_{\mathrm{op}} \\
&\leqslant o(\log^4(n) \max(np\tau^2, \sqrt{np})) + O(p_0) + O(dp\tau) \\
&\leqslant o(\log^4(n) \max(np\tau^2, \sqrt{np})).
\end{aligned}$$

And similarly, for $\tau < \mu \leqslant d^{-1/4} \log^{-1/2}(n)$, we have that with high probability,

$$\begin{aligned}
&\|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top)\|_{\mathrm{op}} \\
&\leqslant \|A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top) - \operatorname{diag}\|_{\mathrm{op}} + \|\operatorname{diag}(A - (p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top))\|_{\mathrm{op}} \\
&\leqslant o(\log^5(n) \max(npd\mu^4, \sqrt{np})) + O(p_0) + O(dp\tau) \\
&\leqslant o(\log^5(n) \max(npd\mu^4, \sqrt{np})).
\end{aligned}$$

4.6 Hypothesis testing

We begin by recalling our hypothesis testing algorithm. Define τ' to be the connectivity threshold for the one-community model $\boldsymbol{G}_{n,d}(p, 0)$. Correspondingly, as in equation Equation (3), we define λ_k' to be the normalized Gegenbauer polynomial expansion coefficient of $\mathbb{1}(x \geqslant \sqrt{\tilde{d}}\tau')$. For the task of hypothesis testing, we check if the second largest eigenvalue of A satisfies

$$\eta_1 > n\lambda_1' \left(1 + \frac{1}{2} \max\left\{\sqrt{\frac{\log 1/p}{d}}, \sqrt{\frac{d}{np \log \frac{1}{p}}}\right\} \log^9 n\right).$$

If so, we declare the model to be the separated mixture model. Otherwise, we say the model is the one-community model.

Theorem (Restatement of Theorem 1.5)**.** *Define the one-community model to be the null hypothesis* $H_0 =$

$\mathbf{G}_{n,d}(p,0)$ and the separated mixture model to be the alternative hypothesis $H_1 = \mathbf{G}_{n,d}(p,\mu)$. If d, n, μ satisfy

$$\mu^2 \geqslant \max\left\{\sqrt{\frac{\log 1/p}{d^3}}, \sqrt{\frac{1}{npd\log\frac{1}{p}}}\right\}\log^9 n, \qquad \log^{16} n \ll d < n, \qquad pn \gg 1, \qquad p \in [0, 1/2-\varepsilon],$$

then if we run the spectral algorithm described above on input graph G we have that

$$\min\{\mathbf{Pr}(\textit{accept } H_0 \mid G \sim H_0), \mathbf{Pr}(\textit{reject } H_0 \mid G \sim H_1)\} \geqslant 1 - o_n(1),$$

In other words, both type 1 error and type 2 error go to zero as n goes to infinity.

In the rest of the subsection, we will firstly prove Theorem 2.1 and Theorem 2.3. This will allow us prove Theorem 1.5.

As the vector $\tilde{\mathbb{1}}_n$ is not necessarily orthogonal to columns of U, we next prove a proposition that shows that they are not far away from othogononality, and in fact the same results hold for the projection. Define a projected matrix as

$$\mathcal{P}(UU^\top) := \left(I - \frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2}\right)UU^\top\left(I - \frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2}\right).$$

Proposition 4.19. *For $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$, we have that the following holds with high probability,*

$$\|A - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1\mathcal{P}(UU^\top)\|_{\mathrm{op}} \ll \log^9(n)\max\left\{np\tau^2, \sqrt{np}\right\}, \quad \text{if } \mu \leqslant \tau,$$
$$\|A - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1\mathcal{P}(UU^\top)\|_{\mathrm{op}} \ll \log^9(n)\max\left\{npd\mu^4, \sqrt{np}\right\}, \quad \text{if } \tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n).$$

Proof of Proposition 4.19. We will show that with high probability,

$$\|UU^\top - \mathcal{P}(UU^\top)\|_{\mathrm{op}} = O(\log(n)). \tag{13}$$

Together with Proposition 2.2 and Proposition 2.4, this would imply the proposition, since $\tilde{d}\lambda_1\log n$ is much smaller than the right hand sides of the proposition. To bound the operator norm, we note that

$$UU^\top - \mathcal{P}(UU^\top) = -\frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2}UU^\top - UU^\top\frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2} + \frac{\|\tilde{\mathbb{1}}_n^\top U\|^2}{\|\tilde{\mathbb{1}}_n\|^4}\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top.$$

We will bound the operator norm of each of the summands above separately. By definition, U can be written as

$$U_{k,1} = \mu S_i + N_{k,1}, \qquad U_{k,j} = N_{k,j} \quad \text{for } j \in \{2, \cdots, d\},$$

where all $N_{k,j}$ are i.i.d. $\mathcal{N}(0, 1/d)$ and S_i represents the community information, which are i.i.d. samples from $\{\pm 1\}$ with probability $1/2$ each. Recall that $L_k = \ell_k - 1$ and ℓ_k is defined to be the length of w_i, which equals $\sqrt{U_{k,2}^2 + \cdots U_{k,d}^2}$. We note that, L_k is independent of S_k and $N_{k,1}$. Therefore, by Lemma 4.14, Lemma 4.18 and concentration inequalities for subgaussian random variables,

$$\left|(\tilde{\mathbb{1}}^\top U)_1\right| = \left|\sum_{k=1}^n (1 + L_k\lambda_1\tilde{d}\tau/p_0)(\mu S_k + N_{k,1})\right| = O\left(\sqrt{\log(n)n\left(\frac{1}{d} + \mu^2\right)}\right),$$

with high probability. Now for general $j \in \{2, \cdots, d\}$, we have that

$$\left|(\tilde{\mathbb{1}}^\top U)_j\right| = \left|\sum_{k=1}^{n}(1 + L_k\lambda_1\tilde{d}\tau/p_0)N_{k,j}\right| \leqslant \left|\sum_{k=1}^{n} N_{k,j}\right| + \lambda_1\tilde{d}\tau/p_0\left|\sum_{k=1}^{n}(\ell_k - 1)\,N_{k,j}\right|.$$

The first term is again bounded by $O(\sqrt{\log(n)n/d})$. For the second term, we note that by Lemma 4.14 and Lemma 4.18, the term $\ell_k - 1 = o(\log n/\sqrt{d})$ uniformly with high probability. Therefore, by concentration inequalities for subgaussian random variables again, we have that

$$\lambda_1\tilde{d}\tau/p_0\left|\sum_{k=1}^{n}(\ell_k - 1)\,N_{k,j}\right| \leqslant \lambda_1\tilde{d}\tau/p_0\sqrt{\frac{n\log^3(n)}{d^2}} \ll \sqrt{\frac{\log(n)n}{d}},$$

with high probability uniformly in j. Therefore we have that

$$\|\tilde{\mathbb{1}}_n^\top U\|^2 = (\tilde{\mathbb{1}}^\top U)_1^2 + \sum_{j=2}^{d}(\tilde{\mathbb{1}}^\top U)_j^2 = O(n\mu^2\log n) + O(n\log n) = O\left(n\log n\right),$$

with high probability, provided that $\mu^2 = O(1)$. This implies that with high probability,

$$\left\|\frac{\|\tilde{\mathbb{1}}_n^\top U\|^2}{\|\tilde{\mathbb{1}}_n\|^4}\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top\right\|_{\mathrm{op}} = \frac{\|\tilde{\mathbb{1}}_n^\top U\|^2}{\|\tilde{\mathbb{1}}_n\|^4}\left\|\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top\right\|_{\mathrm{op}} = \frac{\|\tilde{\mathbb{1}}_n^\top U\|^2}{\|\tilde{\mathbb{1}}_n\|^2} = O(\log(n)).$$

Similarly, we have that with high probability,

$$\left\|\frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2}UU^\top\right\|_{\mathrm{op}} \leqslant \sum_{j=1}^{d}\left\|\frac{\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top}{\|\tilde{\mathbb{1}}_n\|^2}(U^j)(U^j)^\top\right\|_{\mathrm{op}} = \sum_{j=1}^{d}\frac{(\tilde{\mathbb{1}}^\top U)_j^2}{\|\tilde{\mathbb{1}}_n\|^2} = O(\log(n)),$$

where U^j denotes the j-th column of U. Putting the above estimates together, we have that

$$\left\|UU^\top - \mathcal{P}(UU^\top)\right\|_{\mathrm{op}} = O(\log(n)).$$

□

To facilitate the proof, we cite a useful result concerning the spectrum of $UU^\top$.

Lemma 4.20 (Spectrum of $UU^\top$)**.** *Let $\lambda_1(UU^\top)$ be the largest eigenvalue of $UU^\top$, where each row is sampled from the Gaussian mixture distribution $\frac{1}{2}\mathcal{N}(-\mu\cdot e_1, \frac{1}{d}\mathbb{1}_d) + \frac{1}{2}\mathcal{N}(\mu\cdot e_1, \frac{1}{d}\mathbb{1}_d)$. Then we have that*

$$\left|\lambda_1(UU^\top) - \left(\mu^2 n + \frac{n}{d}\right)\right| \leqslant O\left(\sqrt{\frac{d}{n}}\left(\mu^2 n + \frac{n}{d}\right)\right),$$

with high probability. Similarly, if where each row is sampled from the Gaussian distribution $\mathcal{N}(0, \frac{1}{d}\mathbb{1}_d)$, then we have that

$$\left|\lambda_1(UU^\top) - \left(\frac{n}{d}\right)\right| \leqslant O\left(\sqrt{\frac{n}{d}}\right),$$

with high probability.

Proof. This follows from Theorem 4.6.1 in [Ver18]. □

Proof of Theorem 2.1. Recall that (see the definitions before Theorem 2.1) we used η_0 and w_0 to denote the first eigenvalue and eigenvector of the adjacency matrix A respectively. Similarly, as before, we use $\tilde{d}\lambda_1 \mathsf{U}\mathsf{U}^\top$ to denote the projection of A onto the subspace spanned by its second to $d+1$-th eigenvector. By Proposition 2.2, we have that

$$\|A - \eta_0 w_0 w_0^\top - \tilde{d}\lambda_1 \mathsf{U}\mathsf{U}^\top\|_{\text{op}} \ll \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}.$$

Combining with Proposition 4.19, we have that

$$\|\eta_0 w_0 w_0^\top + d\lambda_1 \mathsf{U}\mathsf{U}^\top - p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 \mathcal{P}(UU^\top)\|_{\text{op}} \ll \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}.$$

We note that the first eigenvalue of $p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 \mathcal{P}(UU^\top)$ is $p_0 \|\tilde{\mathbb{1}}_n\|^2 = \Theta(pn)$ and the second eigenvalue of it satisfies

$$\lambda_2(p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 \mathcal{P}(UU^\top)) = \tilde{d}\lambda_1 \cdot \lambda_1(\mathcal{P}(UU^\top)) \leqslant \tilde{d}\lambda_1 O\left(\mu^2 n + \frac{n}{d} + \sqrt{\frac{d}{n}}\left(\mu^2 n + \frac{n}{d}\right) + \log(n)\right) \ll pn,$$

by Equation (13) and Lemma 4.20. Thus, by the Davis-Kahan $\sin\theta$ theorem (Theorem 4.5.5 in [Ver18]), we have that there exist $\theta \in \{\pm 1\}$, such that

$$\left\|\theta w_0 - \frac{\tilde{\mathbb{1}}_n}{\|\tilde{\mathbb{1}}_n\|}\right\|_2 \ll \frac{\log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}}{pn}. \tag{14}$$

Without loss of generality, assume $\theta = 1$. Therefore by Equation (14), we have that

$$\begin{aligned}
\|\eta_0 w_0 w_0^\top - p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top\|_{\text{op}} &\leqslant |\eta_0 - p_0 \|\tilde{\mathbb{1}}_n\|^2| + p_0 \|\tilde{\mathbb{1}}_n\|^2 \|(w_0 w_0^\top - \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top / \|\tilde{\mathbb{1}}_n\|^2)\|_{\text{op}} \\
&\leqslant |\eta_0 - p_0 \|\tilde{\mathbb{1}}_n\|^2| + 2p_0 \|\tilde{\mathbb{1}}_n\|^2 \|w_0 - \tilde{\mathbb{1}}_n / \|\tilde{\mathbb{1}}_n\|\| \ll \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}.
\end{aligned}$$

Therefore, combine the above with the fact that $\|UU^\top - \mathcal{P}(UU^\top)\|_{\text{op}} = O(\log(n))$, we have that

$$\|d\lambda_1 \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top\|_{\text{op}} \ll \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}.$$

As $|\tilde{d}\lambda_1 - d\lambda_1| = \lambda = O(p\tau)$, we thus have that

$$\begin{aligned}
\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\text{op}} &\leqslant \frac{1}{dp\tau} \|(dp\tau \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top)\|_{\text{op}} + \frac{1}{dp\tau} \|(dp\tau - \tilde{d}\lambda_1) UU^\top)\|_{\text{op}} \\
&= o\left(\frac{1}{dp\tau} \log^9(n) \max\left\{np\tau^2, \sqrt{np}\right\}\right) + O\left(\frac{1}{dp\tau} p\tau \left(\frac{n}{d} + \mu^2 n\right)\right) \\
&= o\left(\max\left\{\frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p} d\tau}\right\} \log^9(n)\right).
\end{aligned}$$

□

Proof of Theorem 2.3. We use the same proof idea for large μ. By Proposition 2.4, we have that

$$\|A - \eta_0 w_0 w_0^\top - \tilde{d}\lambda_1 \mathsf{U}\mathsf{U}^\top\|_{\text{op}} \ll \log^9(n) \max\left\{npd\mu^4, \sqrt{np}\right\}.$$

Combining with Proposition 4.19, we have that

$$\|\eta_0 w_0 w_0^\top + d\lambda_1 \mathsf{U}\mathsf{U}^\top - p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 \mathcal{P}(UU^\top)\|_{\text{op}} \ll \log^9(n) \max\left\{npd\mu^4, \sqrt{np}\right\}.$$

We note that the first eigenvalue of $p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 \mathcal{P}(UU^\top)$ is $p_0\|\tilde{\mathbb{1}}_n\|^2 = \Theta(pn)$ and the second eigenvalue of it satisfies

$$\lambda_2(p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 \mathcal{P}(UU^\top)) = \tilde{d}\lambda_1 \cdot \lambda_1(\mathcal{P}(UU^\top)) \leqslant \tilde{d}\lambda_1 O\left(\mu^2 n + \frac{n}{d} + \sqrt{\frac{d}{n}}\left(\mu^2 n + \frac{n}{d}\right) + \log(n)\right) \ll pn,$$

by Equation (13) and Lemma 4.20. Thus, by the Davis-Kahan $\sin\theta$ theorem (Theorem 4.5.5 in [Ver18]), we have that there exist $\theta \in \{\pm 1\}$, such that

$$\left\| \theta w_0 - \frac{\tilde{\mathbb{1}}_n}{\|\tilde{\mathbb{1}}_n\|}\right\|_2 \ll \frac{\log^9(n) \max\left\{npd\mu^4, \sqrt{np}\right\}}{pn}. \tag{15}$$

Without loss of generality, assume $\theta = 1$. Therefore by Equation (15), we have that

$$\begin{aligned}\|\eta_0 w_0 w_0^\top - p_0 \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top\|_{\mathrm{op}} &\leqslant |\eta_0 - p_0\|\tilde{\mathbb{1}}_n\|^2| + p_0\|\tilde{\mathbb{1}}_n\|^2 \|(w_0 w_0^\top - \tilde{\mathbb{1}}_n \tilde{\mathbb{1}}_n^\top / \|\tilde{\mathbb{1}}_n\|^2)\|_{\mathrm{op}} \\ &\leqslant |\eta_0 - p_0\|\tilde{\mathbb{1}}_n\|^2| + 2p_0\|\tilde{\mathbb{1}}_n\|^2 \|w_0 - \tilde{\mathbb{1}}_n/\|\tilde{\mathbb{1}}_n\|\| \ll \log^9(n)\max\left\{npd\mu^4, \sqrt{np}\right\}.\end{aligned}$$

Therefore, combine the above with the fact that $\|UU^\top - \mathcal{P}(UU^\top)\|_{\mathrm{op}} = O(\log(n))$, we have that

$$\|d\lambda_1 \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{op}} \ll \log^9(n)\max\left\{npd\mu^4, \sqrt{np}\right\}.$$

As $|\tilde{d}\lambda_1 - d\lambda_1| = \lambda_1 = O(p\tau)$, we thus have that

$$\begin{aligned}\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{op}} &\leqslant \frac{1}{dp\tau}\|(dp\tau \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top)\|_{\mathrm{op}} + \frac{1}{dp\tau}\|(dp\tau - \tilde{d}\lambda_1)UU^\top)\|_{\mathrm{op}} \\ &= o\left(\frac{1}{dp\tau}\log^9(n)\max\left\{npd\mu^4, \sqrt{np}\right\}\right) + O\left(\frac{1}{dp\tau}p\tau\left(\frac{n}{d} + \mu^2 n\right)\right) \\ &= o\left(\max\left\{\frac{n\mu^4}{d\tau}, \frac{\sqrt{n}}{\sqrt{p}d\tau}\right\}\log^9(n)\right).\end{aligned}$$

□

Proof of Theorem 1.5. We use subscripts g and m to denote terms in the Gaussian model and in the mixture model respectively. In the Gaussian model $\boldsymbol{G}_{n,d}(0,\tau')$, by combining Theorem 2.1 with Lemma 4.20, we have that the largest eigenvalue of $\mathsf{U}_{\mathrm{g}}\mathsf{U}_{\mathrm{g}}^\top$ satisfies

$$\lambda_1(\mathsf{U}_{\mathrm{g}}\mathsf{U}_{\mathrm{g}}^\top) \leqslant \frac{n}{d} + O\left(\sqrt{\frac{n}{d}}\right) + o\left(\max\left\{\frac{n\tau'}{d}, \frac{\sqrt{n}}{\sqrt{p}d\tau'}\right\}\log^9(n)\right)$$

with high probability. Therefore, in the testing problem,

$$\begin{aligned}\lambda_2(A_{\mathrm{g}}) = \tilde{d}\lambda_1' \cdot \lambda_1(\mathsf{U}_{\mathrm{g}}\mathsf{U}_{\mathrm{g}}^\top) &= \frac{n\tilde{d}\lambda_1'}{d} + o\left(\tilde{d}\lambda_1' \max\left\{\frac{n\tau'}{d}, \frac{\sqrt{n}}{\sqrt{p}d\tau'}\right\}\log^9(n)\right) \\ &= n\lambda_1' + o\left(d\lambda_1' \max\left\{\frac{n\tau'}{d}, \frac{\sqrt{n}}{\sqrt{p}d\tau'}\right\}\log^9(n)\right) \\ &= n\lambda_1'\left(1 + o\left(\max\left\{\tau', \frac{1}{\sqrt{np}\tau'}\right\}\log^9(n)\right)\right)\end{aligned}$$

$$< n\lambda_1' \left(1 + \frac{1}{2} \max\left\{ \sqrt{\frac{\log 1/p}{d}}, \sqrt{\frac{d}{np \log \frac{1}{p}}} \right\} \log^9 n \right),$$

where the last inequality follows from the fact that $\tau' = \Theta(\sqrt{\log(1/p)/d})$, by Lemma 4.8. Therefore, $\mathbf{Pr}(\text{accept } H_0 \mid G \sim H_0) \geqslant 1 - o_n(1)$ holds. Now in the mixture model $\boldsymbol{G}_{n,d}(\mu, \tau)$, by combining Theorem 2.1 with Lemma 4.20, we have that when $\mu \leqslant \tau$,

$$\begin{aligned} \lambda_1(\mathsf{U}_{\mathrm{m}}\mathsf{U}_{\mathrm{m}}^\top) &\geqslant \mu^2 n + \frac{n}{d} + O\left(\sqrt{\frac{d}{n}} \left(\mu^2 n + \frac{n}{d} \right) \right) - o\left(\max\left\{ \frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right) \\ &= \mu^2 n + \frac{n}{d} - o\left(\max\left\{ \frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right), \end{aligned}$$

with high probability. For $\mu \leqslant \tau$ satisfying the conditions in Theorem 1.5, we have that

$$\begin{aligned} \lambda_2(A_{\mathrm{m}}) = \tilde{d}\lambda_1' \cdot \lambda_1(\mathsf{U}_{\mathrm{m}}\mathsf{U}_{\mathrm{m}}^\top) &= \tilde{d}\lambda_1'\mu^2 n + \frac{n\tilde{d}\lambda_1'}{d} - o\left(\tilde{d}\lambda_1' \max\left\{ \frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right) \\ &\geqslant n\lambda_1' \left(\tilde{d}\mu^2 + 1 - o\left(\max\left\{ \frac{\tau}{d}, \frac{1}{\sqrt{np}\tau} \right\} \log^9(n) \right) \right) \\ &\geqslant n\lambda_1' \left(\tilde{d} \max\left\{ \sqrt{\frac{\log 1/p}{d^3}}, \sqrt{\frac{1}{npd \log \frac{1}{p}}} \right\} \log^9 n + 1 - o\left(\max\left\{ \frac{\tau}{d}, \frac{1}{\sqrt{np}\tau} \right\} \log^9(n) \right) \right) \\ &> n\lambda_1' \left(1 + \frac{1}{2} \max\left\{ \sqrt{\frac{\log 1/p}{d}}, \sqrt{\frac{d}{np \log \frac{1}{p}}} \right\} \log^9 n \right), \end{aligned}$$

where again the last inequality follows from the fact that $\tau = \Theta(\sqrt{\log(1/p)/d})$, by Lemma 4.8. Therefore, we have $\mathbf{Pr}(\text{reject } H_0 \mid G \sim H_1) \geqslant 1 - o_n(1)$. Now $\mu > \tau$, by combining Theorem 2.3 with Lemma 4.20, we have that,

$$\begin{aligned} \lambda_1(\mathsf{U}_{\mathrm{m}}\mathsf{U}_{\mathrm{m}}^\top) &\geqslant \mu^2 n + \frac{n}{d} + O\left(\sqrt{\frac{d}{n}} \left(\mu^2 n + \frac{n}{d} \right) \right) - o\left(\max\left\{ \frac{n\mu^4}{d\tau}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right) \\ &= \mu^2 n + \frac{n}{d} - o\left(\max\left\{ \frac{n\mu^4}{d\tau}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right), \end{aligned}$$

with high probability. Therefore, for $\mu > \tau$ satisfying the conditions in Theorem 1.5, we have that

$$\begin{aligned} \lambda_2(A_{\mathrm{m}}) &= \tilde{d}\lambda_1' \cdot \lambda_1(\mathsf{U}_{\mathrm{m}}\mathsf{U}_{\mathrm{m}}^\top) \\ &= \tilde{d}\lambda_1'\mu^2 n + \frac{n\tilde{d}\lambda_1'}{d} - o\left(\tilde{d}\lambda_1' \max\left\{ \frac{n\mu^4}{d\tau}, \frac{\sqrt{n}}{\sqrt{p} d\tau} \right\} \log^9(n) \right) \\ &\geqslant n\lambda_1' \left(\tilde{d}\mu^2 + 1 - o\left(\max\left\{ \frac{\mu^4}{d\tau}, \frac{1}{\sqrt{np}\tau} \right\} \log^9(n) \right) \right) \\ &\geqslant n\lambda_1' \left(1 + \frac{1}{3}\tilde{d}\mu^2 + \frac{1}{2}\tilde{d} \max\left\{ \sqrt{\frac{\log 1/p}{d^3}}, \sqrt{\frac{1}{npd \log \frac{1}{p}}} \right\} \log^9 n \right) \\ &> n\lambda_1' \left(1 + \frac{1}{2} \max\left\{ \sqrt{\frac{\log 1/p}{d}}, \sqrt{\frac{d}{np \log \frac{1}{p}}} \right\} \log^9 n \right), \end{aligned}$$

where again we used the fact that $\tau = \Theta(\sqrt{\log(1/p)/d})$, by Lemma 4.8. Therefore, we have $\mathbf{Pr}(\text{reject } H_0 \mid G \sim H_1) \geqslant 1 - o_n(1)$. Therefore, the statement follows. □

4.7 Latent vector embedding

In this subsection, we prove our latent vector embedding results. We re-state our theorem in terms of τ for convenience.

Theorem (Restatement of Theorem 1.4). *Suppose that $n, d \in \mathbb{Z}_+$ and $\mu \in \mathbb{R}_+$, and $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, satisfy the conditions $\log^{16} n \ll d < n$, $\mu^2 \leqslant 1/(\sqrt{d}\log n)$, and $pn \gg 1$. Then given $G \sim \boldsymbol{G}_{n,d}(\mu, \tau)$ generated by latent vectors $u_1, \dots, u_n \in \mathbb{R}^d$, the spectral algorithm described before produces vectors $\hat{u}_1, \dots, \hat{u}_n$ which satisfy*

$$\mathop{\mathbf{E}}_{i,j\sim[n]} |\langle \hat{u}_i, \hat{u}_j\rangle - \langle u_i, u_j\rangle| \ll \max\left\{\tau, \frac{\mu^2}{\tau}, \frac{1}{\sqrt{np}\tau}\right\} \log^9 n \mathop{\mathbf{E}}_{i,j\sim[n]} |\langle u_i, u_j\rangle|,$$

with high probability as n goes to infinity.

We state another approximation theorem in terms of the spectral distance between the matrices.

Theorem 4.21. *Suppose that $n, d \in \mathbb{Z}_+$ and $\mu \in \mathbb{R}_+$, and $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, satisfy the conditions $\log^{16} n \ll d < n$, $\mu^2 \leqslant 1/(\sqrt{d}\log n)$, and $pn \gg 1$, we have that*

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{op}} \ll \frac{1}{1+\mu^2 d} \max\left\{\tau, \frac{\mu^4}{d\tau}, \frac{1}{\sqrt{np}\tau}\right\} \log^9(n) \|UU^\top\|_{\mathrm{op}},$$

with high probability as n goes to infinity.

Proof of Theorem 4.21. This theorem follows directly by Theorem 2.1, Theorem 2.3, and Lemma 4.20. □

In the rest of the subsection, we prove Theorem 1.4. We firstly prove a proposition that bounds the Frobenius norm of $\mathsf{U}\mathsf{U}^\top - UU^\top$.

Proposition 4.22. *We have that for any $\mu \leqslant d^{-1/4}\log^{-1/2}(n)$,*

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{F}} \overset{\text{w.h.p.}}{\ll} \sqrt{d} \max\left\{\frac{n\tau}{d}, \frac{n\mu^2}{d\tau}, \frac{\sqrt{n}}{\sqrt{p}d\tau}\right\} \log^9(n).$$

Proof. When $\mu \leqslant \tau$, by Theorem 2.1, we have that

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{F}} \leqslant \sqrt{2d}\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{op}} \overset{\text{w.h.p.}}{\ll} \sqrt{d} \max\left\{\frac{n\tau}{d}, \frac{\sqrt{n}}{\sqrt{p}d\tau}\right\} \log^9(n).$$

When $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, by a similar proof as that of Lemma 4.17, we have that with high probability

$$\begin{aligned}
&\|p_0^{i,j}\mathbb{1}_n\mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j}VV^\top - (p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top) - \mathrm{diag}\|_{\mathrm{F}} \\
&\leqslant \|[f_1(i,j)]_{0,n\times n} + [f_2(i,j)]_{0,n\times n} + [f_3(i,j)]_{0,n\times n} + [f_4(i,j)]_{0,n\times n}\|_{\mathrm{F}}.
\end{aligned}$$

For the first term, we have that $\|[f_1(i,j)]_{0,n\times n}\|_{\mathrm{F}} \leqslant o\left(npd\tau^4\log^2 n\right) + O(npd^2\tau^2\mu^4)$. And for the second term, similarly, we have that $\|[f_2(i,j)]_{0,n\times n}\|_{\mathrm{F}} \leqslant o\left(np\tau^4 d\log^2 n\right)$. For the third term, we have that

$$\begin{aligned}
&\|[f_3(i,j)]_{0,n\times n}\|_{\mathrm{F}} = \|[\lambda_1\tilde{d}^2\tau\left(a_ia_j(1 - L_i - L_j) + \tau(L_i + L_j) + o(\tau^3\log^2 n) + O(\mu^4\tau d)\right)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} \\
&\leqslant \|[\lambda_1\tilde{d}^2\tau a_ia_j\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} + 2\|[\lambda_1\tilde{d}^2\tau a_ia_jL_i\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} + 2\|[\lambda_1\tilde{d}^2\tau^2 L_i\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} \\
&\qquad + n\lambda_1\tilde{d}^2\tau o(\tau^3\log^2 n)o(\tfrac{\log n}{\sqrt{d}}) + n\lambda_1\tilde{d}^2\tau o(\mu^4\tau d)O(\tfrac{\log n}{\sqrt{d}})
\end{aligned}$$

$$\leqslant \lambda_1 \tilde{d}^2 \tau \mu^2 O\left(\tfrac{n}{\sqrt{d}}\right) + \lambda_1 \tilde{d}^2 \tau \mu^2 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{\sqrt{d}}\right) + \lambda_1 \tilde{d}^2 \tau^2 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{\sqrt{d}}\right)$$
$$+ n\lambda_1 \tilde{d}^2 \tau o(\tau^3 \log^2 n) o(\tfrac{\log n}{\sqrt{d}}) + n\lambda_1 \tilde{d}^2 \tau o(\mu^4 \tau d) O(\tfrac{\log n}{\sqrt{d}})$$
$$\leqslant o\left(np\tau^5 d^2 \log^3 n + np\sqrt{d}\mu^2 \log n\right).$$

Furthermore,

$$\begin{aligned}
&\|[f_4(i,j)]_{0,n\times n}\|_{\mathrm{F}} = \|[\tilde{d}\lambda_1(\ell_i\ell_j - 1)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} = \|[\tilde{d}\lambda_1(L_i + L_j + o(\tfrac{\log^2 n}{d}))\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} \\
&= \|[\tilde{d}\lambda_1(L_i + L_j)\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} + \|[\tilde{d}\lambda_1 o(\tfrac{\log^2 n}{d})\langle v_i, v_j\rangle]_{0,n\times n}\|_{\mathrm{F}} \\
&\leqslant \tilde{d}\lambda_1 o\left(\tfrac{\log n}{\sqrt{d}}\right) O\left(\tfrac{n}{\sqrt{d}}\right) + n\tilde{d}\lambda_1 o\left(\tfrac{\log^2 n}{d}\right) o\left(\tfrac{\log n}{\sqrt{d}}\right) \leqslant o\left(np\tau \log n\right).
\end{aligned}$$

Putting all the inequalities together, we have that

$$\|p_0^{i,j}\mathbb{1}_n\mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j}VV^\top - (p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top) - \operatorname{diag}\|_{\mathrm{F}} = o(np\tau\log^5 n + np\sqrt{d}\mu^2 \log n).$$

We also note that by the same argument as in the proof of Proposition 2.2 and Proposition 2.4,

$$\|\operatorname{diag}\|_{\mathrm{F}} = O(\sqrt{n}dp\tau) = o(\sqrt{dnp}\log^4(n)).$$

Therefore, we know that

$$\|p_0^{i,j}\mathbb{1}_n\mathbb{1}_n^\top + \tilde{d}\lambda_1^{i,j}VV^\top - (p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top + \tilde{d}\lambda_1 UU^\top)\|_{\mathrm{F}} = o(np\tau\log^5 n + np\sqrt{d}\mu^2 \log n).$$

By Proposition 4.1, we further have that with high probability,

$$\begin{aligned}
&\|\eta_0 w_0 w_0^\top + d\lambda_1 \mathsf{U}\mathsf{U}^\top - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{F}} \\
&= o\left(\max\left\{np\tau^2, \sqrt{np}\right\}\sqrt{d}\log^9(n)\right) + o(np\tau\log^5 n + np\sqrt{d}\mu^2\log n) \\
&\ll \max\left\{np\tau^2, \sqrt{np}, np\mu^2\right\}\sqrt{d}\log^9(n).
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
&\|d\lambda_1 \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{F}} \\
&\leqslant \|\eta_0 w_0 w_0^\top + d\lambda_1 \mathsf{U}\mathsf{U}^\top - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{F}} + \|\eta_0 w_0 w_0^\top - p_0\tilde{\mathbb{1}}_n\tilde{\mathbb{1}}_n^\top\|_{\mathrm{F}} \\
&\ll \max\left\{np\tau^2, \sqrt{np}, np\mu^2\right\}\sqrt{d}\log^9(n).
\end{aligned}$$

We thus have that

$$\begin{aligned}
d\lambda_1\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{F}} &\leqslant \|d\lambda_1 \mathsf{U}\mathsf{U}^\top - \tilde{d}\lambda_1 UU^\top\|_{\mathrm{F}} + |\tilde{d}\lambda_1 - d\lambda_1|\|UU^\top\|_{\mathrm{F}} \\
&\ll \max\left\{np\tau^2, \sqrt{np}, np\mu^2\right\}\sqrt{d}\log^9(n).
\end{aligned}$$

As $|\tilde{d}\lambda_1 - d\lambda_1| = O(\lambda_1) = O(p\tau)$, we thus have that

$$\|\mathsf{U}\mathsf{U}^\top - UU^\top\|_{\mathrm{F}} \ll \frac{1}{d\lambda_1}\max\left\{np\tau^2, \sqrt{np}, np\mu^2\right\}\sqrt{d}\log^9(n) = O\left(\max\left\{\frac{n\tau}{d}, \frac{n\mu^2}{d\tau}, \frac{\sqrt{n}}{\sqrt{p}\sqrt{d}\tau}\right\}\log^9(n)\right).$$

□

Proof of Theorem 1.4. We note that by Cauchy–Schwarz inequality and Proposition 4.22,

$$\frac{1}{n^2}\sum_{i,j\in[n]}|\langle\hat{u}_i,\hat{u}_j\rangle-\langle u_i,u_j\rangle|\leqslant\sqrt{\frac{1}{n^2}\sum_{i,j\in[n]}|\langle \mathsf{u}_i,\mathsf{u}_j\rangle-\langle u_i,u_j\rangle|^2}=\frac{1}{n}\|\mathsf{U}\mathsf{U}^\top-UU^\top\|_{\mathrm{F}}$$
$$\ll\frac{\sqrt{d}}{n}\max\left\{\frac{n\tau}{d},\frac{n\mu^2}{d\tau},\frac{\sqrt{n}}{\sqrt{p}d\tau}\right\}\log^9(n).$$

Therefore, our goal remains to show that

$$\frac{1}{n^2}\sum_{i,j\in[n]}|\langle u_i,u_j\rangle|=\Omega\left(\frac{1}{\sqrt{d}}\right).$$

Notice that for any fixed vector v, we can write

$$\langle v,u_j\rangle=N_j\|v\|+\mu S_jv_1,$$

where N_j follows $\mathcal{N}(0,1/d)$ and S_j equals to $\{\pm1\}$ with probability $1/2$. Therefore, we have that

$$\mathbf{Pr}(|\langle v,u_j\rangle|\geqslant\frac{1}{2\sqrt{d}})\geqslant\mathbf{Pr}(N_j\|v\|\geqslant\frac{1}{2\sqrt{d}})=\mathbf{Pr}(\sqrt{d}N_j\geqslant\frac{1}{2\|v\|})=\Psi(\frac{1}{2\|v\|}),$$

where Ψ is 1−CDF of standard Gaussian distribution. For each $i\in[n]$ with $\|u_i\|\geqslant1/2$, if we fix u_i, then we have that $\{|\langle u_i,u_j\rangle|\}_{j\neq i}$ are all independent and satisfy that $\mathbf{Pr}(|\langle u_i,u_j\rangle|\geqslant1/(2\sqrt{d}))\geqslant1/2$. Therefore, for any $i\in[n]$ with $\|u_i\|\geqslant1/2$,

$$\mathbf{Pr}\left(\frac{1}{n-1}\sum_{j\neq i}|\langle u_i,u_j\rangle|\geqslant\frac{1}{8\sqrt{d}}\right)\geqslant1-\exp(-n/128).$$

Notice that with high probability there are at least $n/2$ of $i\in[n]$ with $\|u_i\|\geqslant1/2$. So by a union bound over all $i\in[n]$ with $\|u_i\|\geqslant1/2$, we have that with high probability,

$$\frac{1}{n^2}\sum_{i,j\in[n]}|\langle u_i,u_j\rangle|=\Omega\left(\frac{1}{\sqrt{d}}\right).\qquad\square$$

4.8 Spectral clustering

In this section, we will prove our clustering result. The theorem will follow from an analysis of a basic spectral clustering algorithm for a mixture of two Gaussians in the absence of a perturbation to the data matrix which is bounded in operator norm. The argument is certainly not novel, but we could not find a statement in the literature which matched our precise needs and so we include this appendix for completeness.

Algorithm 4.23 (Spectral clustering)**.** On input $M\in\mathbb{R}^{n\times n}$, compute the top right singular vector a of M and then output $y=\mathrm{sign}(a)$, applying the sign function entrywise and breaking ties arbitrarily if $a_i=0$.

Proposition 4.24. *Suppose $U\in\mathbb{R}^{d\times n}$ with $d\leqslant n$ has columns $u_1,\ldots,u_n$ which are sampled independently from the Gaussian mixture $\frac{1}{2}\mathcal{N}(-\theta,\frac{1}{d}\mathbb{1}_d)+\frac{1}{2}\mathcal{N}(-\theta,\frac{1}{d}\mathbb{1}_d)$ with $\|\theta\|=\mu$, and let $x\in\{\pm1\}^n$ denote the vector of the component labels of the u_i, so that $x_i=+1$ if and only if u_i was sampled from the mixture component with mean $+\theta$.*

Then with probability $1 - o(1)$ over the matrix U, when Algorithm 4.23 is run on an $n \times n$ matrix $M = U^\top U + \Delta$ for Δ an arbitrary matrix with $\|\Delta\| \leqslant \eta$, then the output of Algorithm 4.23 is a vector $y \in \{\pm 1\}^n$ with

$$\frac{|\langle x, y\rangle|}{n} \geqslant 1 - O\left(\frac{1}{\mu\sqrt{d}}\right) - O\left(\sqrt{\frac{\eta}{\mu^2 n}}\right),$$

that is, Algorithm 4.23 clusters at most a $O(\frac{1}{\mu\sqrt{d}}) + O(\sqrt{\frac{\eta}{\mu^2 n}})$-fraction of the columns of U incorrectly.

We re-state our theorem for convenience.

Theorem (Restatement of Theorem 1.6)**.** *Suppose that $n, d \in \mathbb{Z}_+$ and $\mu \in \mathbb{R}_+$, and $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$, satisfy the conditions $d^{-1/2} \ll \mu \leqslant d^{-1/4}\log^{-1/2} n$, $\log^{16} n \ll d < n$ and $pn \gg 1$. If $G \sim \boldsymbol{G}_{n,d}(p, \mu)$, then with high probability Algorithm 4.23 on input $\mathsf{U}\mathsf{U}^\top$ correctly labels (up to a global sign flip) a*

$$1 - O\left(\frac{1}{\mu\sqrt{d}} + \sqrt{\max\left\{\frac{\mu^2}{\tau}, \frac{1}{d\tau\mu^2\sqrt{np}}\right\}\log^9 n}\right)\text{-fraction of the vertices.}$$

Remark 4.25. Similar to the case in latent vector recovery, potentially we could remove the second error term $O(\sqrt{\mu^2/\tau})$ by choosing $\hat{u}_i$ in a slightly different way that accounts for changes in the connecting probability as μ gets close to its upper limit of $d^{-1/4}\log^{-1/2} n$.

Proof of Theorem 1.6. For $\tau < \mu \leqslant d^{-1/4}\log^{-1/2}(n)$, we prove the theorem by combining Proposition 4.24 with Theorem 2.3. □

Proof of Proposition 4.24. Our proof proceeds in two steps: first we will show that with high probability, the top eigenvector of $U^\top U$ is well-correlated with x. Then we will apply a matrix perturbation result to argue that the same is true for $M = U^\top U + \Delta$.

We let $\lambda_i(\cdot)$ denote the ith-largest eigenvalue, and $\sigma_i(\cdot)$ denote the ith-largest singular value. We also define $\alpha := \sqrt{\frac{1}{d}\frac{1}{\mu}}$ and assume for the remainder of the proof that $\alpha \ll 1$ (which we may do without loss of generality since otherwise the claim of the theorem is vacuous).

Claim 4.4. *With probability $1 - o(1)$, $\lambda_2(U^\top U) \leqslant \frac{n}{d}(1 + O(\sqrt{\frac{d}{n}}))$. Further, if $\alpha = \sqrt{\frac{1}{d}\frac{1}{\mu}} \ll 1$, then with probability $1 - o(1)$, $\lambda_1(U^\top U)^2 = \mu^2 n(1 \pm O(\alpha))$.*

Proof of Claim 4.4. We will work with $UU^\top$ rather than $U^\top U$; both matrices have the same spectrum but $UU^\top$ is more convenient to work with because it has full rank. We decompose each $u_i = z_i + x_i\theta$, for $z_i \sim \mathcal{N}(0, \frac{1}{d}\mathbb{1}_d)$. Let Z be the $d \times n$ matrix whose columns are the z_i. Then $U = Z + \theta x^\top$, and

$$UU^\top = ZZ^\top + \theta x^\top Z^\top + Zx\theta^\top + n \cdot \theta\theta^\top.$$

Invoking standard matrix concentration results (see e.g. [Ver18], Theorem 4.6.1), with probability at least $1 - o(1)$ over the choice of U, $\|\frac{d}{n}ZZ^\top - \mathbb{1}_d\| \leqslant C\sqrt{\frac{d}{n}}$ for a universal constant C.

Conditioning on this event, for any unit vector $w \perp \theta$, $w^\top UU^\top w = \frac{n}{d}(1 \pm C\sqrt{\frac{d}{n}})$, and so defining v_1 to be the top eigenvector of $UU^\top$, from the variational characterization of eigenvalues,

$$\lambda_2(UU^\top) = \max_{\substack{\|w\|=1\\ w\perp v_1}} w^\top UU^\top w \leqslant \max_{\substack{\|w\|=1\\ w\perp \theta}} w^\top UU^\top w \leqslant \frac{n}{d}\left(1 + C\sqrt{\frac{d}{n}}\right).$$

Now, we lower bound the maximum eigenvalue. Let $\bar{\theta}$ be the unit vector in the direction of θ. We have

$$\lambda_1(UU^\top) \geqslant \bar{\theta}^\top UU^\top \bar{\theta} = \bar{\theta}^\top ZZ^\top \bar{\theta} + 2\|\theta\| \cdot x^\top Z^\top \bar{\theta} + n\|\theta\|^2 = \frac{n}{d}\left(1 \pm C\sqrt{\tfrac{d}{n}}\right) + 2\mu \cdot x^\top Z^\top \bar{\theta} + n\mu^2$$

and because $x^\top Z^\top \bar{\theta} \sim \mathcal{N}(0, \frac{n}{d})$, with probability $1 - o(1)$, $|x^\top Z^\top \bar{\theta}| \leqslant n\sqrt{\frac{\log n}{dn}}$. Therefore using that $\alpha = \sqrt{\frac{1}{d}}/\mu \ll 1$ and that $d \leqslant n$,

$$\lambda_1(UU^\top) \geqslant \frac{n}{d}\left(1 - C\sqrt{\tfrac{d}{n}}\right) - 2\mu n\sqrt{\frac{\log n}{dn}} + n\mu^2 \geqslant n\mu^2(1 - 2\alpha - (C+1)\alpha^2)$$

with high probability.

Finally, we also need an upper bound on $\lambda_1(UU^\top)$. For this we can use the above concentration results; decomposing any unit $w \in \mathbb{R}^d$ into the sum $w = c\bar{\theta} + w_\perp$ for $c = \langle \bar{\theta}, w\rangle$ and $w_\perp$ the orthogonal component,

$$\begin{aligned}
\lambda_1(UU^\top) &= \max_{\|w\|=1} w^\top UU^\top w \\
&= \max_{\substack{c\in[0,1] \\ w_\perp \perp \bar{\theta}, \|w_\perp\|^2 = 1-c^2}} c^2 \bar{\theta}^\top UU^\top \bar{\theta} + w_\perp^\top UU^\top w_\perp \\
&\leqslant \max_{c\in[0,1]} c^2 \left(\frac{n}{d}\left(1 + C\sqrt{\tfrac{d}{n}}\right) + 2\mu n\sqrt{\tfrac{\log n}{dn}} + n\mu^2\right) + (1-c^2)\left(\frac{n}{d}\left(1 + C\sqrt{\tfrac{d}{n}}\right)\right) \\
&\leqslant \frac{n}{d}\left(1 + C\sqrt{\tfrac{d}{n}}\right) + 2\mu n\sqrt{\tfrac{\log n}{dn}} + n\mu^2 \\
&\leqslant n\mu^2(1 + 2\alpha + (C+1)\alpha^2)
\end{aligned}$$

□

From Claim 4.4, we can show that the unit vector in the direction of x, $\bar{x}$, is well-correlated with the top right singular vector of U.

Claim 4.5. *Let a_1 be the top unit right eigenvector of $U^\top U$. If $\alpha = \sqrt{\frac{1}{d}}\frac{1}{\mu} \ll 1$, then with high probability,*

$$|\langle \bar{x}, a_1\rangle| \geqslant 1 - O(\alpha).$$

Proof of Claim 4.5. By direct calculation,

$$\bar{x}^\top U^\top U\bar{x} = \|Z\bar{x}\|^2 + 2\|x\| \cdot \theta^\top Z\bar{x} + \|\theta\|^2\|x\|^2 = \|Z\bar{x}\|^2 + 2\sqrt{n}\theta^\top Z\bar{x} + \mu^2 n.$$

We now argue that the first two terms concentrate: since $\|Z\bar{x}\|^2$ is a Chi-squared random variable, with probability at least $1 - o(1)$, $\|Z\bar{x}\|^2 = 1 \pm \sqrt{\frac{C}{d}}$ for a universal constant C. And since $\theta^\top Z\bar{x} \sim \mathcal{N}(0, \frac{\mu^2}{d})$, with probability at least $1 - o(1)$, $|\theta^\top Z\bar{x}| \leqslant \mu\sqrt{\frac{\log n}{d}}$. Hence with high probability,

$$\bar{x}^\top U^\top U\bar{x} \geqslant 1 - \sqrt{\tfrac{C}{d}} - 2\mu n\sqrt{\tfrac{\log n}{dn}} + n\mu^2 \geqslant n\mu^2\left(1 - \left(2 + \tfrac{C}{n}\right)\alpha\right).$$

Now, write $\bar{x} = ca_1 + \bar{x}_\perp$ for a_1 the top right unit singular vector of U and $c = \langle a_1, \bar{x}\rangle$. With Claim 4.4's upper bound on $\lambda_1(U^\top U)$ and $\lambda_2(U^\top U)$ we have that

$$\begin{aligned}
\mu^2 n\left(1 - \left(2 + \tfrac{C}{n}\right)\alpha\right) \leqslant \bar{x}^\top U^\top U\bar{x} &\leqslant c^2\sigma_1^2 + (1-c^2)\sigma_2^2 \\
&\leqslant c^2\mu^2 n(1 + 2\alpha + (C+1)\alpha^2) + \frac{n}{d}\left(1 + C\sqrt{\tfrac{d}{n}}\right)
\end{aligned}$$

$$\leqslant c^2\mu^2 n(1 + 2\alpha + 2(C+1)\alpha^2),$$

and now simplifying the above, so long as $\alpha \ll 1$, we have that

$$|\langle a_1, \bar{x}\rangle| = |c| \geqslant \sqrt{\frac{1-(2+\frac{C}{n})\alpha}{1+2\alpha+(C+1)\alpha^2}} = 1 - O(\alpha). \qquad \square$$

Finally, we use a spectral perturbation bound to argue that the top singular vector of $M = U^\top U + \Delta$ is not too far from the top eigenvector of $U^\top U$. Here we will apply the classic Davis-Kahan $\sin\theta$ theorem ([Ver18], Theorem 4.5.5):

Theorem (Corollary of the Davis-Kahan $\sin\theta$ theorem)**.** *Let $A, B \in \mathbb{R}^{n\times n}$. Let a_i be the ith unit eigenvector of A and let b_i be the ith unit eigenvector of B. Then there exists a sign $s \in \{\pm 1\}$ so that*

$$\|a_i - s\cdot b_i\|_2 \leqslant \frac{2^{3/2}\|A-B\|}{\min_{j\neq i}|\lambda_i(A)-\lambda_j(A)|}$$

From this, Claim 4.4, and Claim 4.5, we conclude that the top singular vector a of M has

$$|\langle a, a_1\rangle| \geqslant 1 - \sqrt{2}\frac{\|\Delta\|}{\lambda_1(U^\top U)-\lambda_2(U^\top U)} \geqslant 1 - \sqrt{2}\frac{\eta}{\mu^2 n(1-O(\alpha))},$$

and hence

$$|\langle a, \bar{x}\rangle| \geqslant |\langle\langle a, a_1\rangle\cdot a_1, \bar{x}\rangle| - \sqrt{1-\langle a, a_1\rangle^2} \geqslant (1-O(\alpha))\cdot\left(1 - O\left(\frac{\eta}{\mu^2 n}\right)\right) - O\left(\sqrt{\frac{\eta}{\mu^2 n}}\right) \geqslant 1 - O(\alpha) - O\left(\sqrt{\frac{\eta}{\mu^2 n}}\right)$$

From Claim 4.5 we conclude that a must agree with x on at least $n(1-O(\alpha)-O(\sqrt{\frac{\eta}{\mu^2 n}}))$ of the entry signs. To see why, note that up to sign we can write

$$a = \bar{x} + \delta.$$

Since each entry of $\bar{x}$ has absolute value $\frac{1}{\sqrt{n}}$, the vector δ has to have magnitude at least $\frac{1}{\sqrt{n}}$ in a coordinate to flip a's sign to be opposite of $\bar{x}$'s. But from Claim 4.5, $\|\delta\| = O(\alpha) + O(\sqrt{\frac{\eta}{\mu^2 n}})$ so δ can flip at most $(O(\alpha) + O(\sqrt{\eta/\mu^2 n}))n$ signs. This completes the proof. $\square$

Acknowledgements

T.S. would like to thank Samory Kpotufe for suggesting the question of spectral clustering in the GBM, and would like to thank Sidhanth Mohanty and Sam Hopkins for inspiring conversations about the work and its relation to recovering embeddings in other models. The authors also wish to thank Kiril Bangachev for pointing out two errors in the original trace method proof.

References

[ABARS20] Emmanuel Abbe, Enric Boix-Adsera, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157, 2020. 5

[Abb17] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017. 1, 8

[ABD21] Konstantin Avrachenkov, Andrei Bobu, and Maximilien Dreveton. Higher-order spectral clustering for geometric graphs. *Journal of Fourier Analysis and Applications*, 27(2):22, 2021. 6

[ABFX08] Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008. 6

[ABS21] Emmanuel Abbe, Francois Baccelli, and Abishek Sankararaman. Community detection on euclidean random graphs. *Information and Inference: A Journal of the IMA*, 10(1):109–160, 2021. 6

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory: 18th Annual Conference on Learning Theory, Bertinoro, Italy, June 27-30, 2005. Proceedings 18*, pages 458–469. Springer, 2005. 9

[AS15] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE Computer Society, 2015. 1, 7

[AVY19] Ernesto Araya Valdivia and De Castro Yohann. Latent distance estimation for random geometric graphs. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[Ban18] Debapratim Banerjee. Contiguity and non-reconstruction results for planted partition models: the dense case. *Electron. J. Probab*, 23(18):1–28, 2018. 7

[BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56:89–113, 2004. 1

[BM17] Debapratim Banerjee and Zongming Ma. Optimal hypothesis testing for stochastic block models with growing degrees. *arXiv preprint arXiv:1705.05305*, 2017. 7

[BMR21] Jess Banks, Sidhanth Mohanty, and Prasad Raghavendra. Local statistics, semidefinite programming, and community detection. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms*, pages 1298–1316. SIAM, 2021. 1

[BMV+18] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018. 48, 49

[Bop87] Ravi B Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pages 280–285. IEEE, 1987. 7

[Bor13] Charles Bordenave. On Euclidean random matrices in high dimension. *Electronic Communications in Probability*, 18:1–8, 2013. 6

[CKK+06] Shuchi Chawla, Robert Krauthgamer, Ravi Kumar, Yuval Rabani, and D Sivakumar. On the hardness of approximating multicut and sparsest-cut. *computational complexity*, 15:94–114, 2006. 1

[CS13] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013. 6

[CWA12] David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012. 7

[Dai13] Feng Dai. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013. 13, 22, 23, 24

[DdNS22] Jingqiu Ding, Tommaso d'Orsi, Rajai Nasser, and David Steurer. Robust recovery for stochastic block models. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science*, pages 387–394. IEEE, 2022. 1

[DKK⁺18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018. 7

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011. 8

[DV13] Yen Do and Van Vu. The spectrum of random kernel matrices: universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 2(03):1350005, 2013. 6

[EF14] Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014. 13, 22

[EK10] Noureddine El Karoui. The spectrum of kernel random matrices. *Annals of Statistics*, 38(1):1–50, 2010. 6

[EMP22] Ronen Eldan, Dan Mikulincer, and Hester Pieters. Community detection and percolation of information in a geometric setting. *Combinatorics, Probability and Computing*, 31(6):1048–1069, 2022. 1, 5, 6, 8

[FM19] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, 2019. 6, 7

[GMMM21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2), 2021. 7

[GMPS18] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 5

[GMPS19] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. Connectivity of random annulus graphs and the geometric block model. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2019. 5

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018. 7

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 1

[HS17] Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science FOCS*, pages 379–390. IEEE, 2017. 6, 8

[Jan95] Svante Janson. Random regular graphs: asymptotic distributions and contiguity. *Combinatorics, Probability and Computing*, 4(4):369–405, 1995. 7

[KG00] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, pages 113–167, 2000. 6

[Kra07] Ilia Krasikov. An upper bound on jacobi polynomials. *Journal of Approximation Theory*, 149(2):116–130, 2007. 23

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018. 7

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000. 20, 50

[LMSY22a] Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Local and global expansion in random geometric graphs. *arXiv preprint arXiv:2210.00158*, 2022. 7

[LMSY22b] Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Testing thresholds for high-dimensional sparse random geometric graphs. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 672–677, 2022. 5

[LY22] Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices. *arXiv preprint arXiv:2205.06308*, 2022. 6

[Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014. 1

[McS01] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001. 7

[MMM21] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418. PMLR, 2021. 7

[MMM22] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. 7

[MNS14] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5), 2014. 7

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015. 1, 7

[MNS16] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. *The Annals of Applied Probability*, pages 2211–2256, 2016. 1

[MPW16] Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841, 2016. 1

[MS90] David W Matula and Farhad Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1-2):113–123, 1990. 1

[MS16] 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, 2016. 7

[MSS22] Elchanan Mossel, Allan Sly, and Youngtak Sohn. Exact phase transitions for stochastic block models and reconstruction on trees. *arXiv preprint arXiv:2212.03362*, 2022. 8

[Nda22] Mohamed Ndaoud. Sharp optimal recovery in the two component Gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022. 3, 4, 5, 7

[RCY11] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):596, 2011. 7

[SB18] Abishek Sankararaman and François Baccelli. Community detection on euclidean random graphs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2181–2200. SIAM, 2018. 6

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 30, 35, 36, 37, 42, 44

[Vu18] Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018. 7

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004. 7

[YP14] Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *Conference on learning theory*, pages 138–175. PMLR, 2014. 7

A Lower bound for hypothesis testing when the embedding is known

In this appendix we give a lower bound for hypothesis testing. We assume that we observe $u_1, \cdots, u_n$ and either $u_1, \ldots, u_n \sim H_0 = \mathcal{N}(0, \frac{1}{d}\mathbb{1})$ or $u_1, \ldots, u_n \sim H_1 = \frac{1}{2}\mathcal{N}(-\mu\theta, \frac{1}{d}\mathbb{1}) + \frac{1}{2}\mathcal{N}(\mu\theta, \frac{1}{d}\mathbb{1})$, where $\theta \sim \mathrm{Unif}(\mathcal{S}^{d-1})$.

Our proof is similar to the proof of Theorem 3 in [BMV⁺18], with two differences. In [BMV⁺18], it is assumed that $d/n = \Theta(1)$ whereas for us, d/n can approach 0 or ∞. Also, the signal vector in [BMV⁺18] follows gaussian distribution, while the signal vector in our setting is either θ or $-\theta$. The proof idea is to use the second moment computation to show that the two distributions are contiguous when the separation μ is small.

We change our notation slightly to align with the notation in [BMV⁺18]. Define X to be a n by d matrix with *i.i.d.* $\mathcal{N}(0, 1)$ entries. Let $\mathbb{P}$ be the distribution of X. Define S to be a n by d random matrix where each row equals $\pm\theta$ with probability $1/2$ independently, with $\theta \sim \mathrm{Unif}(\mathcal{S}^{d-1})$. Let $\mathbb{Q}$ be the distribution of $X + \mu\sqrt{d}S$. Then testing H_0 versus H_1 is the same as testing $\mathbb{P}$ versus $\mathbb{Q}$.

Claim A.1. *If $\mu < (\sqrt{2} - \varepsilon)(nd)^{-1/4}$ for arbitrary constant $\varepsilon > 0$, then $\mathbb{P}$ is contiguous to $\mathbb{Q}$. Thus detection is impossible.*

Proof. By Lemma 1 in [BMV+18], we compute the second moment

$$\mathop{\mathbf{E}}_{X\sim\mathbb{P}}\left[\left(\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}\right)^2\right] = \mathop{\mathbf{E}}_{S,T}\exp(\langle S,T\rangle) = \mathop{\mathbf{E}}_{N}\mathop{\mathbf{E}}_{\theta,\theta'}\exp\left(\mu^2 d\left(N-\frac{n}{2}\right)\langle\theta,\theta'\rangle\right),$$

where T is an independent copy of S with the same distribution, $N\sim\mathrm{Bin}(n,1/2)$, and θ and θ' are independent with the same distribution $\mathrm{Unif}(\mathcal{S}^{d-1})$. We now give a bound for $\mathbf{E}_{\theta,\theta'}(\exp(t\langle\theta,\theta'\rangle))$. Using the density function of $\langle\theta,\theta'\rangle$, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\theta,\theta'}(\exp(t\langle\theta,\theta'\rangle)) &= \int_{-1}^{1}\exp(tx)\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}(1-x^2)^{\frac{d-3}{2}}dx\\
&\leqslant \int_{-1}^{1}\exp(tx)\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}\exp(-x^2(d-3)/2)dx\\
&\leqslant \sqrt{\frac{2\pi}{d-3}}\exp\left(\frac{t^2}{2(d-3)}\right)\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)} = (1+o_d(1))\exp\left(-\frac{t^2}{2(d-3)}\right),
\end{aligned}$$

where in the last equality, we used $\Gamma(x+1/2)/\Gamma(x) = \sqrt{x}(1+o_x(1))$, for x large. With this inequality, we get that

$$\mathop{\mathbf{E}}_{X\sim\mathbb{P}}\left[\left(\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}\right)^2\right] \leqslant \mathop{\mathbf{E}}_{N}\exp\left(\mu^4\frac{d^2}{2(d-3)}\left(N-\frac{n}{2}\right)^2\right)(1+o_d(1)).$$

Now we give a bound for $\mathbf{E}_N\exp\left(t\left(N-n/2\right)^2\right)$. Note that for any $\varepsilon>0$, when $t<\frac{2(1-\varepsilon)}{n}$, we have that

$$\mathop{\mathbf{E}}_{N}\exp\left(t\left(N-\frac{n}{2}\right)^2\right) = \sum_{k=0}^{n}\binom{n}{k}2^{-n}\exp(t(k-n/2)^2) = \sum_{\ell=-n/2}^{n/2}2^{-n}\binom{n}{n/2+\ell}\exp(t\ell^2).$$

For $\ell<n^{2/3}$, we have that

$$\begin{aligned}
2^{-n}\binom{n}{n/2+\ell} &= (1+o_n(1))\frac{\sqrt{2\pi n}}{\sqrt{2\pi(n/2+\ell)}\sqrt{2\pi(n/2+\ell)}}\left(1-\frac{2\ell}{n}\right)^{-(n/2-\ell)}\left(1+\frac{2\ell}{n}\right)^{-(n/2+\ell)}\\
&= (1+o_n(1))\frac{\sqrt{2\pi n}}{\sqrt{2\pi(n/2+\ell)}\sqrt{2\pi(n/2+\ell)}}\left(1-\frac{4\ell^2}{n}\right)^{-(n/2-\ell)}\left(\frac{1-2\ell/n}{1+2\ell/n}\right)^{\ell}\\
&\leqslant \frac{C}{\sqrt{n}}\exp\left(2\ell^2/n\right)\exp\left(-(1-n^{-1/8})4\ell^2/n\right) \leqslant \frac{C}{\sqrt{n}}\exp\left(-(1-o_n(1))2\ell^2/n\right).
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\sum_{\ell=-n/2}^{n/2}2^{-n}\binom{n}{n/2+\ell}\exp(t\ell^2) = \sum_{|\ell|<n^{2/3}}2^{-n}\binom{n}{n/2+\ell}\exp(t\ell^2) + 2^{-n}\sum_{|\ell|\geqslant n^{2/3}}\binom{n}{n/2+\ell}\exp(t\ell^2)\\
&\leqslant \sum_{|\ell|<n^{2/3}}\frac{C}{\sqrt{n}}\exp(t\ell^2)\exp\left(-(1-o_n(1))2\ell^2/n\right) + n\frac{C}{\sqrt{n}}\exp(2(1-\varepsilon)n^{1/3})\exp\left(-(2-o_n(1))n^{1/3}\right)\\
&\leqslant \sum_{|\ell|<n^{2/3}}\frac{C}{\sqrt{n}}\exp(-\varepsilon\ell^2/n) + o_n(1)\\
&\leqslant C_\varepsilon.
\end{aligned}$$

Putting together the computations, we showed that as long as

$$\mu^4 \frac{d^2}{2(d-3)} < \frac{2(1-\varepsilon)}{n},$$

we have that the second moment

$$\mathop{\mathbf{E}}_{X\sim\mathbb{P}}\left[\left(\frac{\mathbb{P}(X)}{\mathbb{Q}(X)}\right)^2\right] \leqslant C_\varepsilon.$$

The above condition is equivalent to

$$\mu < (1-\varepsilon)\left(\frac{4}{dn}\right)^{1/4},$$

for arbitrarily small constant ε.

□

B Distinct connected components when the separation is large

In this appendix we show that when μ is large, each community corresponds to a distinct connected component in the graph.

For $u_1, \dots, u_n \sim \frac{1}{2}\mathcal{N}(-\mu \cdot e_1, \frac{1}{d}\mathbb{1}) + \frac{1}{2}\mathcal{N}(\mu \cdot e_1, \frac{1}{d}\mathbb{1})$, recall that we say that vertex $i \in [n]$ comes from community $+1$ if u_i comes from the component in the mixture with mean $\mu \cdot e_i$; otherwise we say that vertex i comes from community -1. We define C_+ to be the set of label $i \in [n]$ if u_i comes from community $+1$ and C_- to be the set of label $i \in [n]$ if u_i comes from community -1.

Claim B.1. *Suppose $p \in [0, 1/2 - \varepsilon]$ for any constant $\varepsilon > 0$ and $\mu \gg (\frac{1}{d}\log n)^{1/4}$, then there is no crossing edges with high probability.*

Proof. For $i \in C_+$ and $j \in C_-$, we write $u_i = (\mu + N_i, w_i)$, where $N_i \in \mathbb{R}$ and $w_i \in \mathbb{R}^{d-1}$ and similarly write $u_j = (-\mu + N_j, w_j)$, where $N_j \in \mathbb{R}$ and $w_j \in \mathbb{R}^{d-1}$. Here note that $N_i \sim \mathcal{N}(0, 1/d)$, $N_j \sim \mathcal{N}(0, 1/d)$, $w_i \sim \mathcal{N}(0, I_{d-1}/d)$, $w_j \sim \mathcal{N}(0, I_{d-1}/d)$ and they are all independent.

$$\begin{aligned}
&\mathbf{Pr}[i \sim j] = \mathbf{Pr}[(\mu + N_i)(-\mu + N_j) + \langle w_i, w_j\rangle \geqslant \tau] = \mathbf{Pr}[\mu(N_j - N_i) + N_iN_j + \langle w_i, w_j\rangle \geqslant \tau + \mu^2]\\
&\leqslant \mathbf{Pr}[N_j - N_i \geqslant \tau/(2\mu) + \mu/2] + \mathbf{Pr}[N_iN_j + \langle w_i, w_j\rangle \geqslant \tau/2 + \mu^2/2]\\
&\leqslant \mathbf{Pr}[N_j - N_i \geqslant \mu/2] + \mathbf{Pr}[N_iN_j + \langle w_i, w_j\rangle \geqslant \mu^2/2],
\end{aligned}$$

since $\tau > 0$. We note that

$$\mathbf{Pr}[N_j - N_i \geqslant \mu/2] = \mathbf{Pr}[\sqrt{d/2}(N_j - N_i) \geqslant \sqrt{d/2}\cdot\mu/2] \leqslant \exp(-\mu^2 d/8) \ll \exp(-\sqrt{d}).$$

Further, we note that $N_iN_j + \langle w_i, w_j\rangle$ can be written as the difference of two independent normalized Chi-Squared random variables $N_iN_j + \langle w_i, w_j\rangle = (A_d - B_d)/(2d)$, where $A_d, B_d \sim \chi_d^2$. By the Laurent-Massart bound [LM00], we know that $\mathbf{Pr}[A_d - d \geqslant 2\sqrt{dx} + 2x] \leqslant \exp(-x)$, $\mathbf{Pr}[A_d - d \leqslant -2\sqrt{dx}] \leqslant \exp(-x)$ and so does B_d. This implies that there exists a constant c such that

$$\mathbf{Pr}[N_iN_j + \langle w_i, w_j\rangle \geqslant \mu^2/2] \leqslant \exp(-cd\min(\mu^4, 1)) \ll n^{-10}.$$

By a union bound over all vertices $i \in C_+$ and $j \in C_-$, we have the claim. □