THE GEOMETRY OF KERNELIZED SPECTRAL CLUSTERING

BY GEOFFREY SCHIEBINGER[1], MARTIN J. WAINWRIGHT[2] AND BIN YU[3]

University of California, Berkeley

Clustering of data sets is a standard problem in many areas of science and engineering. The method of spectral clustering is based on embedding the data set using a kernel function, and using the top eigenvectors of the normalized Laplacian to recover the connected components. We study the performance of spectral clustering in recovering the latent labels of i.i.d. samples from a finite mixture of nonparametric distributions. The difficulty of this label recovery problem depends on the overlap between mixture components and how easily a mixture component is divided into two nonoverlapping components. When the overlap is small compared to the indivisibility of the mixture components, the principal eigenspace of the population-level normalized Laplacian operator is approximately spanned by the square-root kernelized component densities. In the finite sample setting, and under the same assumption, embedded samples from different components are approximately orthogonal with high probability when the sample size is large. As a corollary we control the fraction of samples mislabeled by spectral clustering under finite mixtures with nonparametric components.

1. Introduction. In the past decade, spectral methods have emerged as a powerful collection of nonparametric tools for unsupervised learning, or clustering. How can we recover information about the geometry or topology of a distribution from its samples? Clustering algorithms attempt to answer the most basic form of this question. One way in which to understand spectral clustering is as a relaxation of the NP-hard problem of searching for the best graph-cut. Spectral graph partitioning—using the eigenvectors of a matrix to find graph cuts—originated in the early 1970s with the work of Fiedler [5] and of Donath and Hoffman [4]. Spectral clustering was introduced in machine learning, with applications to clustering data sets and computing image segmentations (e.g., [9, 10, 14]). The past decade has witnessed an explosion of different spectral clustering algorithms. One point of variation is that some use the eigenvectors of the kernel matrix [4, 7, 15], or

adjacency matrix in the graph setting, whereas others use the eigenvectors of the normalized Laplacian matrix [5, 9, 10, 14]. This division goes all the way back to the work of Donath and Hoffman, who proposed using the adjacency matrix, and of Fiedler, who proposed using the normalized Laplacian matrix.

In its modern and most popular form, the spectral clustering algorithm [10, 14] involves two steps: first, the eigenvectors of the normalized Laplacian are used to embed the dataset, and second, the K-means clustering algorithm is applied to the embedded dataset. The normalized Laplacian embedding is an attractive preprocessing step because the transformed clusters tend to be linearly separable. Ng et al. [10] show that, under certain conditions on the empirical kernel matrix, an embedded dataset will cluster tightly around well-separated points on the unit sphere. Their results apply to a fixed dataset, and do not model the underlying distribution of the data. Recently Yan et al. [18] derived an expression for the fraction of data misclustered by spectral clustering by computing an analytical expression for the second eigenvector of the Laplacian. They assumed that the similarity matrix is a small perturbation away from the ideal block diagonal case.

The embedding defined by the normalized Laplacian has also been studied in the context of manifold learning, where the primary focus has been convergence of the underlying eigenvectors. This work is motivated in part by the fact that spectral properties of the limiting Laplace–Beltrami operator have long been known to shed light on the connectivity of a manifold [8]. The Laplacian eigenmaps of Belkin and Niyogi [2] reconstruct Laplace–Beltrami eigenfunctions from sampled data. Koltchinskii and Giné [6] analyze the convergence of the empirical graph Laplacian to the Laplace–Beltrami operator at a fixed point in the manifold. von Luxburg and Belkin [17] establish consistency for the embedding in as much as the eigenvectors of the Laplacian matrix converge uniformly to the eigenfunctions of the Laplacian operator. Rosasco et al. [12] provide simpler proofs of this convergence, and in part, our work sharpens these results by removing an unnecessary smoothness assumption on the kernel function.

In this paper, we study spectral clustering in the context of a nonparametric mixture model. The study of spectral clustering under nonparametric mixtures was initiated by Shi et al. [15]. One of their theorems characterizes the top eigenfunction of a kernel integral operator, showing that it does not change sign. One difficulty in using the eigenfunctions of a kernel integral operator to separate mixture components is that several of the top eigenfunctions may correspond to a single mixture component (e.g., one with a larger mixture weight). They propose that eigenfunctions of the kernel integral operator that approximately do not change sign correspond to different mixture components. However, their analysis does not deal with finite datasets nor does it provide bounds on the fraction of points misclustered.

The main contribution of this paper is an analysis of the normalized Laplacian embedding of i.i.d. samples from a finite mixture with nonparametric components. We begin by providing a novel and useful characterization of the principal

eigenspace of the population-level normalized Laplacian operator: more precisely, when the mixture components are indivisible and have small overlap, the eigenspace is close to the span of the square root kernelized component densities. We then use this characterization to analyze the geometric structure of the embedding of a finite set of i.i.d. samples. Our main result is to establish a certain geometric property of nonparametric mixtures referred to as *orthogonal cone structure*. In particular, we show that when the mixture components are indivisible and have small overlap, embedded samples from different components are almost orthogonal with high probability. We then prove that this geometric structure allows K-means to correctly label most of the samples. Our proofs rely on techniques from operator perturbation theory, empirical process theory and spectral graph theory.

The remainder of this paper is organized as follows. In Section 2, we set up the problem of separating the components of a mixture distribution. We state our main results and explore some of their consequences in Section 3. We prove our main results in Section 4, deferring the proofs of several supporting lemmas to the supplementary material [13].

Notation. For a generic distribution $\mathbb{P}$ on a measurable space $\mathcal{X}$, we denote the Hilbert space of real-valued square integrable functions on $\mathcal{X}$ by $L^2(\mathbb{P})$. The $L^2(\mathbb{P})$ inner product is given by $\langle f, g \rangle_{\mathbb{P}} = \int f(x)g(x)\,d\mathbb{P}(x)$, and it induces the norm $\|f\|_{\mathbb{P}}$. The norm $\|f\|_{\infty}$ is the supremum of the function f, up to sets of measure zero, where the relevant measure is understood from context. The Hilbert–Schmidt norm of an operator $\mathbf{T}: L^2(\mathbb{P}) \to L^2(\mathbb{P})$ is $\|\|\mathbf{T}\|\|_{\mathrm{HS}}$, and the operator norm is $\|\|\mathbf{T}\|\|_{\mathrm{op}}$. The complement of a set B is denoted by B^c. See Appendix D (supplementary material [13]) for an additional list of symbols.

2. Background and problem set-up.
We begin by introducing the family of nonparametric mixture models analyzed in this paper, and then provide some background on kernel functions, spectral clustering and Laplacian operators.

2.1. *Nonparametric mixture distributions.* For some integer $K \geq 2$, let $\{\mathbb{P}_m\}_{m=1}^{K}$ be a collection of probability measures on a compact space $\mathcal{X}$, and let the weights $\{w_m\}_{m=1}^{K}$ belong to the relative interior of the probability simplex in $\mathbb{R}^K$—that is, $w_m \in (0, 1)$ for all $m = 1, \ldots, K$, and $\sum_{m=1}^{K} w_m = 1$. This pair specifies a *finite nonparametric mixture distribution* via the convex combination

$$(2.1) \qquad \bar{\mathbb{P}} := \sum_{m=1}^{K} w_m \mathbb{P}_m.$$

We refer to $\{\mathbb{P}_m\}_{m=1}^{K}$ and $\{w_m\}_{m=1}^{K}$ as the mixture components and mixture weights, respectively. The family of models (2.1) is *nonparametric*, because the mixture components are not constrained to any particular parametric family.

A random variable $\bar{X} \sim \bar{\mathbb{P}}$ can be obtained by first drawing a multinomial random variable $Z \sim \text{Multinomial}(w_1, \ldots, w_K)$, and conditioning on the event $\{Z = m\}$, drawing a variable from mixture component $\mathbb{P}_m$. Consequently, given a collection of samples $\{X_i\}_{i=1}^n$ drawn i.i.d. from $\bar{\mathbb{P}}$, there is an underlying set of *latent labels* $\{Z_i\}_{i=1}^n$. Thus in the context of a mixture distribution, the clustering problem can be formalized as recovering these latent labels based on observing only the unlabeled samples $\{X_i\}_{i=1}^n$.

Of course, this clustering problem is ill defined whenever $\mathbb{P}_j = \mathbb{P}_k$ for some $j \neq k$. More generally, recovery of labels becomes more difficult as the overlap of any pair $\mathbb{P}_j$ and $\mathbb{P}_k$ increases, or if it is "easy" to divide any component into two nonoverlapping distributions. This intuition is formalized in our definition of the overlap and indivisibility parameters in Section 3.1 to follow.

2.2. *Kernels and spectral clustering.*   We now provide some background on spectral clustering methods and the normalized Laplacian embedding. A kernel k associated with the space $\mathcal{X}$ is a symmetric, continuous function $k : \mathcal{X} \times \mathcal{X} \to (0, \infty)$. A kernel is said to be positive semidefinite if for any integer $n \geq 1$ and elements $x_1, \ldots, x_n \in \mathcal{X}$, the kernel matrix $A \in \mathbb{R}^{n \times n}$ with entries $A_{ij} = k(x_i, x_j)/n$ is positive semidefinite. Throughout we consider a fixed but arbitrary positive semidefinite kernel function. In application to spectral clustering, one purpose of a kernel function is to provide a measure of the similarity between data points. A canonical example is the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2)$; it is close to 1 for vectors x and x' that are relatively close, and decays to zero for pairs that are far apart.

Let us now describe the normalized Laplacian embedding, which is a standard part of many spectral clustering routines. Given n i.i.d. samples $\{X_i\}_{i=1}^n$ from $\bar{\mathbb{P}}$, the associated kernel matrix $A \in \mathbb{R}^{n \times n}$ has entries $A_{ij} = \frac{1}{n} k(X_i, X_j)$. The normalized Laplacian matrix[4] is obtained by rescaling the kernel matrix by its row sums, namely

$$(2.2) \qquad\qquad L = D^{-1/2} A D^{-1/2},$$

where D is a diagonal matrix with entries $D_{ii} = \sum_{j=1}^n A_{ij}$. Since L is a symmetric matrix by construction, it has an orthonormal basis of eigenvectors, and we let $\{v_1, \ldots, v_K\}$ denote the eigenvectors corresponding to the largest K eigenvalues of L. The *normalized Laplacian embedding* is defined on the basis of these eigenvectors: it is the map $\Phi_{\mathcal{V}} : \{X_1, \ldots, X_n\} \to \mathbb{R}^K$ defined by

$$(2.3) \qquad\qquad \Phi_{\mathcal{V}}(X_i) := (v_{1i}, \ldots, v_{Ki}).$$

A typical form of spectral clustering consists of the following two steps. First, compute the normalized Laplacian, and map each data point X_i to a K-vector via

---

[4]To be precise, the matrix $I - L$ is actually the normalized graph Laplacian matrix. However, the eigenvectors of L are identical to those of $I - L$, and we find it simpler to work with L.

the embedding (2.3). The second step is to apply a standard clustering method (such as K-means clustering) to the embedded data points. The conventional rationale for the second step is that the embedding step typically helps reveal cluster structure in the data set, so that it can be found by a relatively simple algorithm. The goal of this paper is to formalize the sense in which the normalized Laplacian embedding (2.3) has this desirable property.

We do so by first analyzing the population operator that underlies the normalized Laplacian matrix. It is defined by the *normalized kernel function*

$$\bar{k}(x, y) := \frac{1}{\bar{q}(x)} k(x, y) \frac{1}{\bar{q}(y)}, \tag{2.4}$$

where $\bar{q}(y) = \sqrt{\int k(x, y) \, d\bar{\mathbb{P}}(x)}$. Note that this kernel function can be seen as a continuous analog of the normalized Laplacian matrix (2.2).

The normalized kernel function in conjunction with the mixture defines the *normalized Laplacian operator* $\bar{\mathbf{T}} : L^2(\bar{\mathbb{P}}) \to L^2(\bar{\mathbb{P}})$ given by

$$(\bar{\mathbf{T}} f)(\cdot) := \int \bar{k}(\cdot, y) f(y) \, d\bar{\mathbb{P}}(y). \tag{2.5}$$

Under suitable regularity conditions (see Appendix C.1 (supplementary material [13]) for details), this operator has an orthonormal set of eigenfunctions—with eigenvalues in $[0, 1]$—and our main results relate these eigenfunctions to the underlying mixture components $\{\mathbb{P}_m\}_{m=1}^K$.

3. Analysis of the normalized Laplacian embedding. This section is devoted to the statement of our main results, and discussion of their consequences. These results involve a few parameters of the mixture distribution, including its overlap and indivisibility parameters, which are defind in Section 3.1. Our first main result (Theorem 1 in Section 3.2) characterizes the principal eigenspace of the population-level normalized Laplacian operator (2.5), showing that it approximately spanned by the square root kernelized densities of the mixture components, as defined in Section 3.1. Our second main result (Theorem 2 in Section 3.3) provides a quantitative description of the angular structure in the normalized Laplacian embedding of a finite sample from a mixture distribution.

3.1. *Cluster similarity, coupling and indivisibility parameters.* In this section, we define some parameters associated with any nonparametric mixture distribution, as viewed through the lens of a given kernel. These quantities play an important role in our main results, as they reflect the intrinsic difficulty of the clustering problem.

Our first parameter is the *similarity index* of the mixture components $\{\mathbb{P}_m\}_{m=1}^K$. For any pair of distinct indices $\ell \neq m$, the ratio

$$\mathcal{S}(\mathbb{P}_\ell, \mathbb{P}_m) := \frac{\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \, d\mathbb{P}_m(x) \, d\mathbb{P}_\ell(y)}{\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \, d\bar{\mathbb{P}}(x) \, d\mathbb{P}_\ell(y)}$$

is a kernel-dependent measure of the expected similarity between the clusters indexed by $\mathbb{P}_\ell$ and $\mathbb{P}_m$, respectively. Note that $\mathcal{S}$ is not symmetric in its arguments. The *maximum similarity* over all mixture components

$$(3.1) \qquad \mathcal{S}_{\max}(\bar{\mathbb{P}}) := \max_{\substack{\ell, m=1,\ldots,K \\ \ell \neq m}} \mathcal{S}(\mathbb{P}_\ell, \mathbb{P}_m)$$

measures the overlap between mixture components with respect to the kernel k.

Our second parameter, known as the coupling parameter, is defined in terms of the square root kernelized densities of the mixture components. More precisely, given any distribution $\mathbb{P}$, its *square root kernelized density* is the function $q \in L^2(\mathbb{P})$ given by

$$(3.2) \qquad q(x) := \sqrt{\int k(x, y) \, d\mathbb{P}(y)}.$$

In particular, we denote the square root kernelized density of the mixture distribution $\bar{\mathbb{P}}$ by $\bar{q}$, and those of the mixture components $\{\mathbb{P}_m\}_{m=1}^K$ by $\{q_m\}_{m=1}^K$. In analogy with the normalized kernel function $\bar{k}$ from equation (2.4), we also define a normalized kernel for each mixture component, namely

$$(3.3) \qquad k_m(x, y) := \frac{k(x, y)}{q_m(x) q_m(y)} \qquad \text{for } m = 1, \ldots, K.$$

The *coupling parameter*

$$(3.4) \qquad \mathcal{C}(\bar{\mathbb{P}}) := \max_{m=1,\ldots,K} \| k_m - w_m \bar{k} \|_{\mathbb{P}_m \otimes \mathbb{P}_m}^2$$

measures the coupling of the spaces $L^2(\mathbb{P}_m)$ with respect to $\bar{\mathbf{T}}$. In particular, when $\mathcal{C}(\bar{\mathbb{P}}) = 0$, then the normalized Laplacian can be decomposed as the sum

$$(3.5) \qquad \bar{\mathbf{T}} = \sum_{m=1}^K w_m \mathbf{T}_m,$$

where $(\mathbf{T}_m f)(y) = \int f(x) k_m(x, y) \, d\mathbb{P}_m(x)$ is the operator defined by the normalized kernel k_m. When the coupling parameter is no longer exactly zero but still small, then decomposition (3.5) still holds in an approximate sense.

Our final parameter measures how easy or difficult it is to "split" any given mixture component $\mathbb{P}_m$ into two or more parts. If this splitting can be done easily for any component, then the mixture distribution will be hard to identify, since there is an ambiguity as to whether $\mathbb{P}_m$ defines one component or multiple components. In order to formalize this intuition, for a distribution $\mathbb{P}$ and for a measurable subset $S \subset \mathcal{X}$, we introduce the shorthand notation $p(S) = \int_S \int_{\mathcal{X}} k(x, y) \, d\mathbb{P}(x) \, d\mathbb{P}(y)$. With this notation, the indivisibility of $\mathbb{P}$ is

$$(3.6) \qquad \Gamma(\mathbb{P}) := \inf_S \frac{p(\mathcal{X}) \int_S \int_{S^c} k(x, y) \, d\mathbb{P}(x) \, d\mathbb{P}(y)}{p(S) p(S^c)},$$

where the infimum is taken over all measurable subsets S such that $p(S) \in (0, 1)$. The *indivisibility parameter* $\Gamma_{\min}(\bar{\mathbb{P}})$ of a mixture distribution $\bar{\mathbb{P}}$ is the minimum indivisibility of its mixture components

$$(3.7) \qquad \Gamma_{\min}(\bar{\mathbb{P}}) := \min_{m=1,\ldots,K} \Gamma(\mathbb{P}_m).$$

Our results in the next section apply when the similarity $\mathcal{S}_{\max}(\bar{\mathbb{P}})$ and coupling $\mathcal{C}(\bar{\mathbb{P}})$ are small compared to the indivisibility $\Gamma_{\min}(\bar{\mathbb{P}})$. Some examples help illustrate when this is the case.

EXAMPLE 1. Consider the one-dimensional triangular density function

$$g_{\mathbb{T}_\mu}(x) := \begin{cases} x - \mu + 1, & \text{if } x \in (\mu - 1, \mu); \\ -x + \mu + 1, & \text{if } x \in (\mu, \mu + 1); \\ 0, & \text{otherwise,} \end{cases}$$

with location $\mu > 0$. We denote corresponding distribution by $\mathbb{T}_\mu$. In this example we calculate the similarity, coupling and indivisibility parameters for the mixture of triangular distributions $\bar{\mathbb{T}} := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$ and the uniform kernel $k_\nu(x, y) = \frac{1}{2\nu}\mathbf{1}\{|x - y| \le \nu\}$ with bandwidth $\nu \in (0, 1)$.[5]

Similarity. It is straightforward to calculate the similarity parameter $\mathcal{S}_{\max}(\bar{\mathbb{T}})$ by solving a few simple integrals. We find that

$$\mathcal{S}_{\max}(\bar{\mathbb{T}}) = \frac{2(2 + \nu - \mu)_+^4}{\nu(16 - 8\nu^2 + 3\nu^3) + (2 + \nu - \mu)_+^4}.$$

Coupling. To compute the coupling parameter, we must compute the kernelized densities of $\mathbb{T}_0$ and $\mathbb{T}_\mu$, and the normalized kernel functions $k_1(x, y)$ and $\bar{k}(x, y)$. Some calculation yields the following equation for the kernelized density:

$$(3.8) \qquad q_1^2(x) = \begin{cases} \dfrac{(1 + \nu - x)^2}{4\nu}, & \text{if } x \in (-1 - \nu, -1 + \nu), \\[2mm] 1 - \dfrac{\nu}{2} - \dfrac{x^2}{2\nu}, & \text{if } x \in (-\nu, \nu), \\[2mm] \dfrac{(1 + \nu + x)^2}{4\nu}, & \text{if } x \in (1 - \nu, 1 + \nu), \\[2mm] g_{\mathbb{T}_0}(x), & \text{otherwise.} \end{cases}$$

As can be seen in Figure 1, the kernelized density $q_1^2(x)$ of $\mathbb{T}_0$ is a smoothed version of $g_{\mathbb{T}_0}(x)$ that interpolates quadratically around the nondifferentiable points of $g_{\mathbb{T}_0}(x)$.

---

[5]This is not a positive semidefinite kernel function, but it helps to build intuition our intuition in a case where all the integrals are easy.
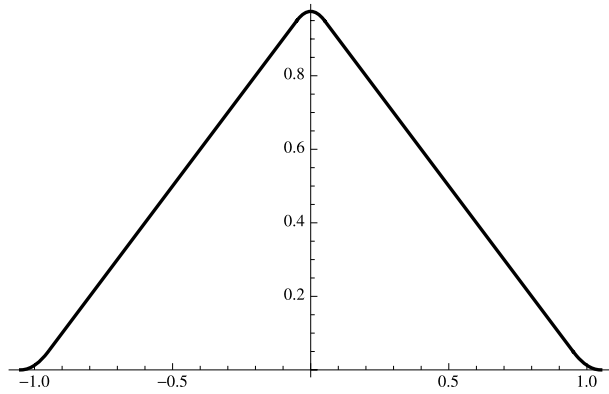
FIG. 1.　*The kernelized density $q_1^2(x)$ of equation* (3.8) *with $\nu = 0.05$.*

The kernelized density of $\mathbb{T}_\mu$ has the same shape as that of $\mathbb{T}_0$ but is shifted by μ. In particular,

$$q_2^2(x) = q_1^2(x - \mu).$$

Therefore the normalized kernels satisfy $k_1(x, y) = \frac{1}{2}\bar{k}(x, y)$ for $x, y \in (-1, \mu - 1 - \nu)$. By upper bounding the integrand over the remaining region, we find that

$$\mathcal{C}(\bar{\mathbb{T}})^2 = \iint (k_1 - \bar{k}/2)^2 \, d\mathbb{P}_1(x, y) \leq 2(2 + \nu - \mu)_+.$$

Indivisibility.　It is straightforward to calculate the indivisibility $\Gamma(\mathbb{T}_\mu)$. For any $\nu \in (0, 1)$ and $\mu \in \mathbb{R}$, the set S defining $\Gamma(\mathbb{T}_\mu)$ is $S = (\mu, \infty)$. Hence by solving a few simple integrals, we find that

$$\Gamma(\mathbb{T}_\mu) = \frac{2(6 - \nu)(2 - \nu)\nu}{16 - 8\nu^2 + 3\nu^3}.$$

Note that $\Gamma(\mathbb{T}_\mu)$ does not depend on μ. Therefore the indivisibility of $\bar{\mathbb{T}} := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$ is

$$\Gamma_{\min}(\bar{\mathbb{T}}) = \Gamma(\mathbb{T}_\mu) = \Gamma(\mathbb{T}_0).$$

It is instructive to consider the indivisibility of the following poorly defined two-component mixture:

$$\bar{\mathbb{T}}_{\text{bad}} := \frac{1}{2}\mathbb{P}_{\text{bad}}(\mu) + \frac{1}{2}\mathbb{T}_{2\mu},$$

where $\mathbb{P}_{\text{bad}}(\mu)$ is the bimodal component $\mathbb{P}_{\text{bad}}(\mu) := \frac{1}{2}\mathbb{T}_0 + \frac{1}{2}\mathbb{T}_\mu$. It is easy to verify that for any $\nu \in (0, 1)$ and $\mu > 2 + \nu$, the indivisibility of the bimodal component is $\Gamma(\mathbb{P}_{\text{bad}}(\mu)) = 0$, and therefore $\Gamma_{\min}(\bar{\mathbb{T}}_{\text{bad}}) = 0$.

From Example 1 we learn that the similarity $\mathcal{S}_{\max}(\bar{\mathbb{T}})$ and coupling $\mathcal{C}(\bar{\mathbb{T}})$ parameters decrease as the offset μ increases. Together, these two parameters measure the overlap which our intuition tells us should decrease as μ increases. On the other hand, the indivisibility parameter $\Gamma_{\min}(\bar{\mathbb{T}})$ is independent of μ.

Our next example is more realistic in the sense that the kernel and mixture components do not have bounded support, and the kernel function is positive semidefinite.

EXAMPLE 2. In this example we calculate the similarity, coupling and indivisibility parameters for the mixture of Gaussians $\bar{\mathbb{N}} = \frac{1}{2}\mathbb{N}(0, 1) + \frac{1}{2}\mathbb{N}(\mu, 1)$ equipped with the Gaussian kernel $k_\nu(x, y) = \frac{1}{\sqrt{2\pi}\nu} \exp[-\frac{|x-y|^2}{2\nu^2}]$.

Similarity. As in the previous example, it is straightforward to calculate the maximal intercluster similarity $\mathcal{S}_{\max}(\bar{\mathbb{N}})$ by solving a handful of Gaussian integrals. We find that

$$(3.9) \qquad \mathcal{S}_{\max}(\bar{\mathbb{N}}) = \frac{2\exp(-\mu^2/(2\nu^2 + 4))}{1 + \exp(-\mu^2/(2\nu^2 + 4))} \le 4e^{-\mu^2/(2\nu^2+4)}.$$

Coupling. The kernelized density of $\mathbb{N}(0, 1)$ is

$$q_1^2(x) = \frac{1}{\sqrt{\nu^2 + 1}} \exp\left[\frac{-x^2}{2(\nu^2 + 1)}\right],$$

and the kernelized density of $\mathbb{N}(\mu, 1)$ is simply the translation $q_2^2(x) = q_1^2(x - \mu)$. We can bound the coupling parameter $\mathcal{C}(\bar{\mathbb{N}})$ by upper bounding the integrand $k_1 - \bar{k}$ over a high-probability compact set (a modification of the trick from Example 1). We show the resulting bound in Figure 2. This (albeit loose) bound captures the exponential decay of $\mathcal{C}(\bar{\mathbb{N}})$ with μ.
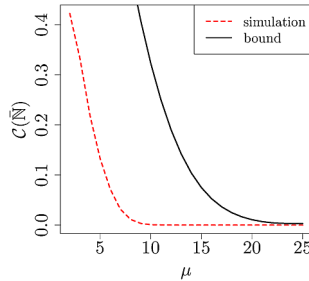

FIG. 2. *The coupling parameter $\mathcal{C}(\bar{\mathbb{N}})$ for the mixture of Gaussians with Gaussian kernel and $\nu = 2$. The red line displays the simulated value of $\mathcal{C}(\bar{\mathbb{N}})$ as a function of the offset μ between mixture components. The black line displays our analytical bound.*

Indivisibility. It is straightforward to compute the indivisibility of the unit-variance normal distribution $\mathbb{N}(\mu, 1)$ with location $\mu \in \mathbb{R}$. The set defining the indivisibility is $S = (\mu, \infty)$. Solving a handful of Gaussian integrals yields

$$(3.10) \qquad \Gamma_{\min}\big(\mathbb{N}(\mu, 1)\big) = \frac{2}{\pi} \arctan\big(\nu\sqrt{2 + \nu^2}\big).$$

We conclude with a counter example, showing that the similarity parameter is *not* relatively small for the linear kernel $k(x, y) = x \cdot y$.

EXAMPLE 3. Consider the mixture $\bar{\mathbb{P}} = \frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$ with components $\mathbb{P}_1$ uniform over $(1, 2)$ and $\mathbb{P}_2$ uniform over $(2 + \delta, 3 + \delta)$. With the linear kernel $k(x, y) = x \cdot y$, we have

$$\mathcal{S}_{\max}(\bar{\mathbb{P}}) = \frac{\iint xy \, d\mathbb{P}_1(x) \, d\mathbb{P}_2(y)}{\iint xy \, d\mathbb{P}_1(x) \, d\bar{\mathbb{P}}(y)} = \frac{\int y \, d\mathbb{P}_2(y)}{\int y \, d\bar{\mathbb{P}}(y)} \geq \frac{1}{2}.$$

Since $\Gamma_{\min}(\bar{\mathbb{P}})$ is always between 0 and 1, this calculation demonstrates that the similarity parameter $\mathcal{S}_{\max}(\bar{\mathbb{P}})$ is never small compared to $\Gamma_{\min}(\bar{\mathbb{P}})$.

3.2. *Population-level analysis.* In this section, we present our population-level analysis of the normalized Laplacian embedding. Consider the following two subspaces of $L^2(\bar{\mathbb{P}})$:

- the subspace $\mathcal{R} \subset L^2(\bar{\mathbb{P}})$ spanned by the top K eigenfunctions of the normalized Laplacian operator $\bar{\mathbf{T}}$ from equation (2.5) and
- the span $\mathcal{Q} = \text{span}\{q_1, \ldots, q_K\} \subset L^2(\bar{\mathbb{P}})$ of the square root kernelized densities; see equation (3.2).

The subspace $\mathcal{Q}$ can be used to define a map $\Phi_{\mathcal{Q}} : \mathcal{X} \to \mathbb{R}^K$ known as the *square-root kernelized density embedding*, given by

$$(3.11) \qquad \Phi_{\mathcal{Q}}(x) := \big(q_1(x), \ldots, q_K(x)\big).$$

This map is relevant to clustering, since the vector $\Phi_{\mathcal{Q}}(x)$ encodes sufficient information to perform a likelihood ratio test (based on the kernelized densities) for labeling data points.

On the other hand, the subspace $\mathcal{R}$ is important because it is the population-level quantity that underlies spectral clustering. As described in Section 2.2, the first step of spectral clustering involves embedding the data using the eigenvectors of the Laplacian matrix. This procedure can be understood as a way of estimating the *population-level Laplacian embedding*: more precisely, the map $\Phi_{\mathcal{R}} : \mathcal{X} \to \mathbb{R}^K$ given by

$$(3.12) \qquad \Phi_{\mathcal{R}}(x) := \big(r_1(x), \ldots, r_K(x)\big),$$

where $\{r_m\}_{m=1}^K$ are the top K eigenfunctions of the kernel operator $\bar{\mathbf{T}}$.

To build intuition, imagine for the moment varying the kernel function so that the kernelized densities converge to the true densities. For example, imagine sending the bandwidth of a Gaussian kernel to 0. While the kernelized densities approach the true densities, the subspace $\mathcal{R}$ is only a well defined mathematical object for kernels with nonzero bandwidth. Indeed, as the bandwidth shrinks to zero, the eigengap separating the principal eigenspace $\mathcal{R}$ of $\bar{\mathbf{T}}$ from its lower eigenspaces vanishes. For this reason, we analyze an arbitrary but fixed kernel function, and we discuss kernel selection in Section 5.

The goal of this section is to quantify the difference between the two mappings $\Phi_{\mathcal{Q}}$ and $\Phi_{\mathcal{R}}$, or equivalently between the underlying subspaces $\mathcal{Q}$ and $\mathcal{R}$. We assume that the square root kernelized densities $q_1, \ldots, q_K$ are linearly independent so that $\mathcal{Q}$ has the same dimension, K, as $\mathcal{R}$. This condition is very mild when the overlap parameters $\mathcal{S}_{\max}(\bar{\mathbb{P}})$ and $\mathcal{C}(\bar{\mathbb{P}})$ are small. We measure the distance between these subspaces by the Hilbert–Schmidt norm[6] applied to the difference between their orthogonal projection operators,

$$(3.13) \qquad \rho(\mathcal{Q}, \mathcal{R}) := \|\!|\Pi_{\mathcal{Q}} - \Pi_{\mathcal{R}}|\!\|_{\mathrm{HS}}.$$

Recall the similarity parameter $\mathcal{S}_{\max}(\bar{\mathbb{P}})$, coupling parameter $\mathcal{C}(\bar{\mathbb{P}})$ and indivisibility parameter $\Gamma_{\min}(\bar{\mathbb{P}})$, as previously defined in equations (3.1), (3.4) and (3.7), respectively. Our main results involve a function of these three parameters and the minimum $w_{\min} := \min_{m=1,\ldots,K} w_m$ of the mixture weights, given by

$$(3.14) \qquad \varphi(\bar{\mathbb{P}}; k) := \frac{\sqrt{K}[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\min}\Gamma_{\min}^2(\bar{\mathbb{P}})}.$$

Our first main theorem guarantees that as long as the mixture is relatively well separated, as measured by the *difficulty function* φ, then the ρ-distance (3.13) between $\mathcal{R}$ and $\mathcal{Q}$ is proportional to $\varphi(\bar{\mathbb{P}}; k)$. Our theorem also involves the quantity

$$b_{\max} := \max_{m=1,\ldots,K} \left\| \int k_m(x, y)\, d\mathbb{P}_m(y) \right\|_{\infty}^2.$$

Note that this is simply a constant whenever the kernels k_m are bounded.

THEOREM 1 (Population control of subspaces). *For any finite mixture $\bar{\mathbb{P}}$ with difficulty function bounded as $\varphi(\bar{\mathbb{P}}; k) \leq [576\sqrt{12 + b_{\max}}]^{-1}\Gamma_{\min}^2(\bar{\mathbb{P}})$, the distance between subspaces $\mathcal{Q}$ and $\mathcal{R}$ is bounded as*

$$(3.15) \qquad \rho(\mathcal{Q}, \mathcal{R}) \leq 16\sqrt{12 + b_{\max}}\,\varphi(\bar{\mathbb{P}}; k).$$

Relationship (3.15) is easy to understand in the context of translated copies of identical mixture components. Consider the mixture of Gaussians with Gaussian

---

[6]Recall that the Hilbert–Schmidt norm of an operator is the infinite dimensional analogue of the Frobenius norm of a matrix.
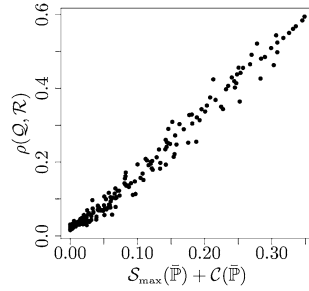
FIG. 3. *The ρ-distance between $\mathcal{Q}$ and $\mathcal{R}$ scales linearly with $\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})$. Each point corresponds to a different offset μ between the two Gaussian mixture components from Example* 3.

kernel setup in Example 3. Recall from equation (3.10) that the indivisibility parameter is independent of the offset μ. Hence in this setting relationship (3.15) simplifies to

$$\rho(\mathcal{Q}, \mathcal{R}) \leq c\big[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})\big]^{1/2}.$$

Figure 3 shows a clear linear relationship between $\rho(\mathcal{Q}, \mathcal{R})$ and $\mathcal{S}_{\max}(\bar{\bar{\mathbb{P}}}) + \mathcal{C}(\bar{\mathbb{P}})$, suggesting that it might be possible to remove the square root in the clustering difficulty (3.14).

One important consequence of relationship (3.15) stems from geometric structure in the square root kernelized density embedding. When there is little overlap between mixture components with respect to the kernel, the square root kernelized densities are not simultaneously large; that is, $\Phi_{\mathcal{Q}}(X)$ will have at most one component much different from zero. Therefore the data will concentrate in tight spikes about the axes. This is illustrated in Figure 4.
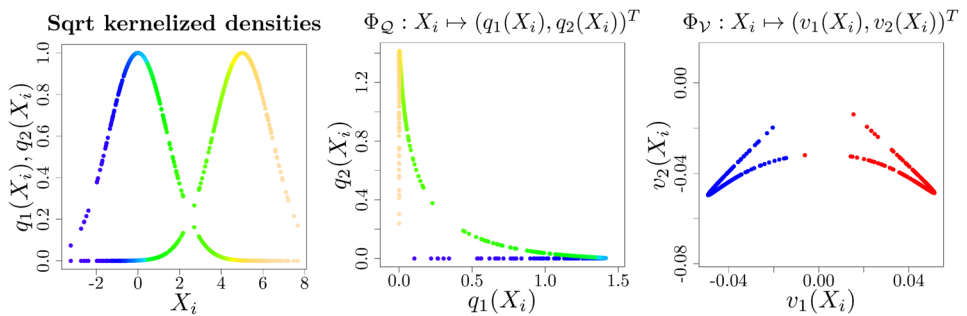

FIG. 4. *Geometric structure in the square root kernelized density embedding.* (*Left*) *square root kernelized densities for a mixture of Gaussians with Gaussian kernel. The color of the ith dot indicates the likelihood ratio of the mixture components at X_i.* (*Center*) *data embedded under the square root kernelized density embedding $\Phi_{\mathcal{Q}}$, colored by likelihood ratio.* (*Right*) *normalized Laplacian embedding of the samples, colored by latent label.*

3.3. *Finite sample analysis.* Thus far, our analysis has been limited to the population level, corresponding to the ideal case of infinitely many samples. We now turn to the case of finite samples. Here an additional level of analysis is required, in order to relate empirical versions (based on the finite collection of samples) to their population analogues. Doing so allows us to show that under suitable conditions, the Laplacian embedding applied to i.i.d. samples drawn from a finite mixture satisfies a certain geometric property, which we call *orthogonal cone structure*, or OCS for short.

We begin by providing a precise definition of when an embedding $\Phi : \mathcal{X} \to \mathbb{R}^K$ reveals orthogonal cone structure. Given a collection of labeled samples $\{X_i, Z_i\}_{i=1}^n \subset \mathcal{X} \times [K]$ drawn from a K-component mixture distribution, we let $\mathcal{Z}_m = \{i \in [n] | Z_i = m\}$ denote the subset of samples drawn from mixture component $m = 1, \ldots, K$. For any set $\mathcal{Z} \subseteq [n] = \{1, 2, \ldots, n\}$, we use $|\mathcal{Z}|$ to denote its cardinality. For vectors $u, v \in \mathbb{R}^n$, we use $\text{angle}(u, v) = \arccos \frac{\langle u,v \rangle}{\|u\|_2 \|v\|_2}$ to denote the angle between them. With this notation, we have the following:

DEFINITION 1 [Orthogonal cone structure (OCS)]. Given parameters $\alpha \in (0, 1)$ and $\theta \in (0, \frac{\pi}{4})$, the embedded data set $\{\Phi(X_i), Z_i\}_{i=1}^n$ has (α, θ)-OCS if there is an orthogonal basis $\{e_1, \ldots, e_K\}$ of $\mathbb{R}^K$ such that

$$\big|\{i \in [n] | \text{angle}(\Phi(X_i), e_m) < \theta\} \cap \mathcal{Z}_m\big| \geq (1 - \alpha)|\mathcal{Z}_m| \qquad \text{for all } m = 1, \ldots, K.$$

In words, a labeled dataset has orthogonal cone structure if most pairs of embedded data points with distinct labels are almost orthogonal. See Figure 5 for an illustration of this property.
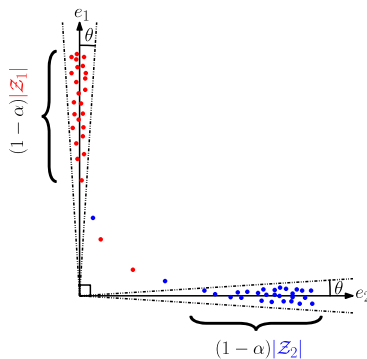

FIG. 5. *Visualizing (α, θ)-OCS: the labeled set of points plotted above has (α, θ) orthogonal cone structure with respect to its labeling. The color of each dot indicates the value of the corresponding label $Z_i \in \{1, 2\}$, where 1 corresponds to red and 2 to blue. This set of points has (α, θ)-orthogonal cone structure because a fraction $1 - \alpha$ of the red points (for which $Z_i = 1$) lie with an angle θ of e_1, a fraction $1 - \alpha$ blue points (for which $Z_i = 2$) lie with an angle θ of e_2 and e_1 is orthogonal to e_2.*

Our main theorem in the finite sample setting establishes that under suitable conditions, the normalized Laplacian embedding has orthogonal cone structure. In order to state this result precisely, we require a few additional conditions.

Kernel parameters. As a consequence of the compactness of $\mathcal{X}$, the kernel function is b-bounded, meaning that $k(x, x') \in (0, b)$ for all $x, x' \in \mathcal{X}$. As another consequence, the kernelized densities are lower bounded as $q_m(X^m) \geq r > 0$ with $\bar{\mathbb{P}}$-probability one. In the following statements, we use $c, c_0, c_1, \ldots$ to denote quantities that may depend on b, and r but are otherwise independent of the mixture distribution.

Tail decay. The tail decay of the mixture components enters our finite sample result through the function $\psi : (0, \infty) \to [0, 1]$, defined by

$$(3.16) \qquad \psi(t) := \sum_{m=1}^{K} \mathbb{P}_m \left[\frac{q_m^2(X)}{\|q_m\|_{\bar{\mathbb{P}}}^2} < t \right].$$

Note that ψ is an increasing function with $\psi(0) = 0$. The rate of increase of ψ roughly measures the tail decay of the square root kernelized densities. Intuitively, perturbations to the square root kernelized density embedding will have a greater effect on points closer to the origin.

Recall the population level clustering difficulty parameter $\varphi(\bar{\mathbb{P}}; k)$ previously defined in equation (3.14). Our theory requires that there is some $\delta > 0$ such that

$$(3.17) \qquad \underbrace{\left[\varphi(\bar{\mathbb{P}}; k) + \frac{1}{\Gamma_{\min}^2(\bar{\mathbb{P}})} \left(\frac{1}{\sqrt{n}} + \delta \right) \right]}_{\varphi_n(\delta)} \leq c \Gamma_{\min}^2(\bar{\mathbb{P}}).$$

In essence, we assume that the indivisibility of the mixture components is not too small compared to the clustering difficulty.

With this notation, the following result applies to i.i.d. labeled samples $\{(X_i, Z_i)\}_{i=1}^{n}$ from a K-component mixture $\bar{\mathbb{P}}$.

THEOREM 2 (Finite-sample angular structure). *There are numbers c, c_0, c_1, c_2 depending only on b and r such that for any $\delta \in (0, \frac{\|k\|_{\bar{\mathbb{P}}}}{b\sqrt{2\pi}})$ satisfying condition (3.17) and any $t > \frac{c_0}{w_{\min}^3} \sqrt{\varphi_n(\delta)}$, the embedded data set $\{\Phi_{\mathcal{V}}(X_i), Z_i\}_{i=1}^{n}$ has (α, θ)-OCS with*

$$(3.18) \quad |\cos\theta| \leq \frac{c_0 \sqrt{\varphi_n(\delta)}}{w_{\min}^3 t - c_0 \sqrt{\varphi_n(\delta)}} \quad \text{and} \quad \alpha \leq \frac{c_1}{(w_{\min})^{3/2}} \varphi_n(\delta) + \psi(2t),$$

and this event holds probability at least $1 - 8K^2 \exp(\frac{-c_2 n \delta^4}{\delta^2 + \mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})})$.
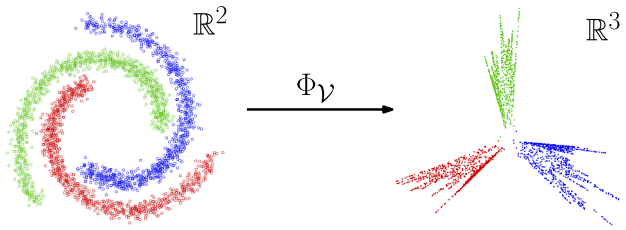
FIG. 6. *According to Theorem 2, the normalized Laplacian embedding of i.i.d. samples from a nonparametric mixture with small overlap, indivisible components and large enough sample size, has (α, θ)-OCS with $\alpha \ll 1$ and $\theta \ll 1$. The left plot shows i.i.d. samples in $\mathbb{R}^2$, and the right plot displays the image (in $\mathbb{R}^3$) of these data under the normalized Laplacian embedding, $\Phi_{\mathcal{V}}$. The embedding was performed using a regularized Gaussian kernel. The color of each point indicates the latent label of that point.*

Theorem 2 establishes that the embedding of i.i.d. samples from a finite mixture $\bar{\mathbb{P}}$ has orthogonal cone structure (OCS) *if* the components have small overlap and good indivisibility. This result holds with high probability on the sampling from $\bar{\mathbb{P}}$. See Figure 6 for an illustration of the theorem.

The tail decay of the mixture components enters the bounds on α and θ in different ways: the bound on θ is inversely proportional to t, but the bound on α is tighter for smaller t. Depending on how quickly ψ increases with t, it may very well be the dominant term in the bound on α. For example, if there is a $\gamma > 0$ such that $\psi(t) \leq t^{\gamma}$ for all $t \in (0, 1)$, and we set $t = \varphi_n^{\beta}$ for some $\beta \in (0, 1)$, then we obtain the simplified bounds

$$|\cos \theta| \leq \frac{c}{w_{\min}^3} \varphi_n^{1/2 - \beta} \quad \text{and} \quad \alpha \leq \frac{c}{(w_{\min})^{3/2}} \big(\varphi_n + \varphi_n^{\gamma \beta} \big).$$

Indeed, we find that whenever $\gamma \beta < 2$, the tail decay function is the dominant term in the bound on α. Note that this power law tail decay is easy to verify for the Gaussian distribution with Gaussian kernel from Example 2; see Figure 7.

Finally, the numbers c, c_0, c_1, c_2 increase as the kernel bound b increases and as r decreases. This is where we need the tail truncation condition $r > 0$. This as-
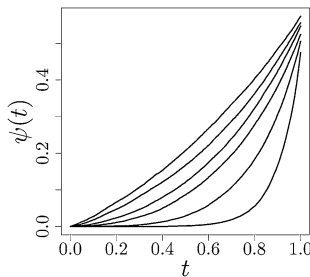

FIG. 7. *The tail decay function $\psi(t)$ roughly follows a power law for the standard Gaussian distribution and Gaussian kernel with bandwidths $v \in \{0.15, 0.45, 0.75, 1, 1.5, 2.5\}$.*

sumption is common in the literature; see Cao and Chen [3], for example. Both von Luxburg, Belkin and Bousquet [17] and Rosasco, Belkin and De Vito [12] assume $k(x, y) \geq r > 0$, which is more restrictive. Note that this automatically holds if we add a positive constant to any kernel. This is sometimes called *regularization* and can significantly increase the performance of spectral clustering in practice [1].

3.4. *Algorithmic consequences.* In this section we apply our theory to study the performance of spectral clustering. The standard spectral clustering algorithm applies K-means to the embedded dataset. For completeness, we give pseudo code for the update step of K-means in Algorithm 1 below.

In practice, we have found that applying K-means to an embedded dataset works well if the underlying orthogonal cone structure is "nice enough." The following proposition provides a quantitative characterization of this phenomenon. It applies to an embedded data set $\{\Phi_{\mathcal{V}}(X_i), Z_i\}_{i=1}^n$ with (α, θ)-OCS, and an initialization of $\mathbf{a}_1, \ldots, \mathbf{a}_K$ as uniformly random orthonormal vectors. Recall the notation $\mathcal{Z}_m = \{i \in [n] | Z_i = m\}$.

PROPOSITION 1. *Suppose θ and α are sufficiently small that*

(3.19)
$$\frac{\alpha n + (1 - \alpha)|\mathcal{Z}_m| \sin \theta}{(1 - \alpha)|\mathcal{Z}_m|} \leq \sin \frac{\pi}{8} \quad and$$

$$\frac{(1 - \alpha)|\mathcal{Z}_m| \cos \theta - \alpha n}{|\mathcal{Z}_m| + \alpha n} \geq \frac{1}{2}, \qquad m = 1, \ldots, K.$$

Then there is a constant c_K such that with probability at least $1 - \frac{4c_K \theta}{2\pi}$ over the random initialization, the K-means algorithm misclusters at most αn points. When $K = 2$, we have $c_K = 1$.

Algorithm 1 K-means update

Input: Normalized embedded data $y_i := \frac{\Phi_{\mathcal{V}}(X_i)}{\|\Phi_{\mathcal{V}}(X_i)\|}$ for $i = 1, \ldots, n$, and mean vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$

for $m \in \{1, \ldots, K\}$ **do**

$$\hat{\mathcal{Z}}_m \leftarrow \left\{ i : m = \underset{\ell}{\operatorname{argmin}} \|\mathbf{a}_\ell - y_i\| \right\}$$

$$\mathbf{a}'_m \leftarrow \sum_{i \in \hat{\mathcal{Z}}_m} \frac{y_i}{|\hat{\mathcal{Z}}_m|}$$

end for

return $\{\hat{\mathcal{Z}}_1, \ldots, \hat{\mathcal{Z}}_K\}$ and $\{\mathbf{a}'_1, \ldots, \mathbf{a}'_K\}$

Intuitively, condition (3.19) requires α and θ to be small enough so that the different cones from the (α, θ)-OCS do not overlap.

PROOF OF PROPOSITION 1. We provide a detailed proof for the case $K = 2$. By the definition of (θ, α)-OCS, there exist orthogonal vectors e_1, e_2 such that a fraction $1 - \alpha$ of the embedded samples with latent label m lie within an angle θ of e_m, $m = 1, 2$. Let us say that the initialization is *unfortunate* if some $\mathbf{a}_j$ falls within angle $\frac{\theta}{2}$ of the angular bisector of e_1 and e_2, an event which occurs with probability $\frac{4\theta}{2\pi}$.

Suppose without loss of generality that $\mathbf{a}_1$ is closer to e_1, and let $\mathbf{a}'_1, \mathbf{a}'_2$ denote the updates

$$\mathbf{a}'_m = \sum_{i \in \hat{\mathcal{Z}}_m} \frac{v_i}{|\hat{\mathcal{Z}}_m|}, \qquad m = 1, 2.$$

If the initialization is *not* unfortunate, then all points in the θ-cone around e_1 are closer to $\mathbf{a}_1$ than $\mathbf{a}_2$. In this case, the (θ, α)-OCS implies that the e_2-coordinate of $\mathbf{a}'_1$ is at most

$$\frac{\alpha n + (1 - \alpha)|\mathcal{Z}_1| \sin\theta}{(1 - \alpha)|\mathcal{Z}_1|} \le \sin\frac{\pi}{8},$$

and the e_1-coordinate of $\mathbf{a}'_1$ is at least

$$\frac{(1 - \alpha)|\mathcal{Z}_1| \cos\theta - \alpha n}{|\mathcal{Z}_1| + \alpha n} \ge \frac{1}{2}.$$

We conclude that all points in the θ-cone about e_1 are closer to $\mathbf{a}'_1$ than $\mathbf{a}'_2$. Consequently, we find that after a single update step of K-means, all but a fraction α of the samples are correctly labeled. Moreover, this holds for all subsequent K-means updates. This completes the proof for $K = 2$.

The proof for general K follows the same steps. The probability that any $\mathbf{a}_m$ falls within angle $\frac{\theta}{2}$ of the angular bisector of any pair e_j, e_ℓ is still proportional to θ, with a constant of proportionality c_K that depends on K. $\square$

4. Proofs. We now turn to the proofs of our main results, beginning with the population level result stated in Theorem 1. We then provide the proof of Theorem 2 and Proposition 1.

4.1. *Proof of Theorem* 1. Our proof leverages an operator perturbation theorem due to Stewart [16] to show that $\mathcal{Q}$ is an approximate invariant subspace of the normalized Laplacian operator $\bar{\mathbf{T}}$ from equation (2.5). Recalling that $\Pi_\mathcal{Q}$ denotes the projection onto subspace $\mathcal{Q}$ (with $\Pi_{\mathcal{Q}^\perp}$ defined analogously), consider the following three operators:

$$\mathbf{A} := \Pi_\mathcal{Q} \bar{\mathbf{T}} \Pi_\mathcal{Q}^*, \qquad \mathbf{B} := \Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}} \Pi_{\mathcal{Q}^\perp}^* \quad \text{and} \quad \mathbf{G} := \Pi_{\mathcal{Q}^\perp} \bar{\mathbf{T}} \Pi_\mathcal{Q}^*.$$

By definition, a subspace $\mathcal{Q}$ is invariant under $\bar{\mathbf{T}}$ if and only if $\mathbf{G} = 0$. In our setting, this ideal situation occurs when there is no overlap between mixture components. More generally, operator perturbation theory can be used to guarantee that a space is approximately invariant as long as the Hilbert–Schmidt norm $\|\|\mathbf{G}\|\|_{\mathrm{HS}}$ is not too large relative to the spectral separation between $\mathbf{A}$ and $\mathbf{B}$. In particular, define the quantities

$$\gamma := \|\|\mathbf{G}\|\|_{\mathrm{HS}} \quad \text{and} \quad \mathrm{sep}(\mathbf{A}, \mathbf{B}) := \inf\{|a - b| | a \in \sigma(\mathbf{A}), b \in \sigma(\mathbf{B})\}.$$

In application to our problem, Theorem 3.6 of Stewart [16] guarantees that as long as $\frac{\gamma}{\mathrm{sep}(\mathbf{A},\mathbf{B})} < \frac{1}{2}$, then there is an operator $\mathbf{S} : \mathcal{Q} \to \mathcal{Q}^{\perp}$ such that

$$(4.1) \qquad\qquad\qquad \|\|\mathbf{S}\|\|_{\mathrm{HS}} \leq \frac{2\gamma}{\mathrm{sep}(\mathbf{A}, \mathbf{B})}$$

such that $\mathrm{Range}(\Pi_{\mathcal{Q}}^{*} + \Pi_{\mathcal{Q}^{\perp}}^{*}\mathbf{S})$ is an invariant subspace of $\bar{\mathbf{T}}$.

Accordingly, in order to apply this result, we first need to control the quantities $\|\|\mathbf{G}\|\|_{\mathrm{HS}}$ and $\mathrm{sep}(\mathbf{A}, \mathbf{B})$. The bulk of our technical effort is devoted to proving the following two lemmas:

LEMMA 1 (Hilbert–Schmidt bound).   *We have*

$$(4.2) \qquad\qquad \|\|\mathbf{G}\|\|_{\mathrm{HS}} \leq \frac{\sqrt{K(12 + b_{\max})}}{w_{\min}} \sqrt{\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})},$$

where $b_{\max} := \max_{m=1,\ldots,K} \| \int k_m(x, y) \, d\mathbb{P}_m(y)\|_{\infty}^2$.

LEMMA 2 (Spectral separation bound).   *Under the hypothesis of the theorem, we have*

$$\sigma_{\min}(\mathbf{A}) \geq 1 - 13K[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}$$

and

$$\sigma_{\max}(\mathbf{B}) \leq 1 - \frac{\Gamma^2}{8} + \frac{3[\mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})]^{1/2}}{w_{\min}}.$$

Consequently, the spectral separation is lower bounded as

$$(4.4) \qquad\qquad\qquad \mathrm{sep}(\mathbf{A}, \mathbf{B}) \geq \frac{\Gamma^2}{16}.$$

See Appendices A.1 and A.2 (supplementary material [13]), respectively, for the proof of these lemmas.

Combined with our earlier bound (4.1), these two lemmas guarantee that

$$(4.5) \qquad\qquad\qquad \|\|\mathbf{S}\|\|_{\mathrm{HS}} \leq 16\sqrt{12 + b_{\max}}\varphi(\bar{\mathbb{P}}, k).$$

Moreover, we find that $\mathrm{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^* \mathbf{S})$ is equal to $\mathcal{R}$, the principal eigenspace of $\bar{\mathbf{T}}$. Indeed, by Stewart's theorem, the spectrum of $\bar{\mathbf{T}}$ is the disjoint union $\sigma(\bar{\mathbf{T}}) = \sigma(\mathbf{A} + \mathbf{G}^*\mathbf{S}) \cup \sigma(\mathbf{B} - \mathbf{SG}^*)$. After some calculation using the upper bound on $\varphi(\bar{\mathbb{P}}, k)$ in the theorem hypothesis, we find that the spectrum of $\bar{\mathbf{T}}$ satisfies $\sigma_{\min}(\mathbf{A} + \mathbf{G}^*\mathbf{S}) > \sigma_{\max}(\mathbf{B} - \mathbf{SG}^*)$, and any element $x \in \mathrm{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^* \mathbf{S})$ must satisfy

$$\sup_{q \in \mathcal{Q}} \left\{ \frac{x^* \bar{\mathbf{T}} x}{x^* x} \,\middle|\, x = (\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^* \mathbf{S}) q \right\} > \sigma_{\max}(\mathbf{B} - \mathbf{SG}^*).$$

Therefore $\mathrm{Range}(\Pi_{\mathcal{Q}}^* + \Pi_{\mathcal{Q}^\perp}^* \mathbf{S}) = \mathcal{R}$.

The only remaining step is to translate bound (4.5) into a bound on the norm $\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}\|_{\mathrm{HS}}$. Observe that the difference of projection operators can be written as

$$\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}} = (\Pi_{\mathcal{R}} + \Pi_{\mathcal{R}^\perp})(\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}) = \Pi_{\mathcal{R}} \Pi_{\mathcal{Q}^\perp} - \Pi_{\mathcal{R}^\perp} \Pi_{\mathcal{Q}}.$$

Now Lemma 3.2 of Stewart [16] gives the explicit representations

$$\Pi_{\mathcal{R}} = (\mathbf{I} + \mathbf{S}^*\mathbf{S})^{-1/2} (\Pi_{\mathcal{Q}} + \mathbf{S}^* \Pi_{\mathcal{Q}^\perp})$$

and

$$\Pi_{\mathcal{R}^\perp} = (\mathbf{I} + \mathbf{SS}^*)^{-1/2} (\Pi_{\mathcal{Q}^\perp} + \Pi_{\mathcal{Q}} \mathbf{S}).$$

Consequently, we have

$$\|\Pi_{\mathcal{R}^\perp} \Pi_{\mathcal{Q}}\|_{\mathrm{HS}} \le \left\| (\mathbf{I} + \mathbf{SS}^*)^{-1/2} \mathbf{S} \right\|_{\mathrm{HS}}$$

and

$$\|\Pi_{\mathcal{R}} \Pi_{\mathcal{Q}^\perp}\|_{\mathrm{HS}} \le \left\| (\mathbf{I} + \mathbf{S}^*\mathbf{S})^{-1/2} \mathbf{S}^* \right\|_{\mathrm{HS}}.$$

By the continuous functional calculus (see Section VII.1 of Reed and Simon [11]), we have the expansion

$$(\mathbf{I} + \mathbf{SS}^*)^{-1/2} = \sum_{n=1}^{\infty} \binom{2n}{n} \frac{(\mathbf{SS}^*)^{n-1}}{2^{2n}}.$$

Putting together the pieces, in terms of the shorthand $\varepsilon = \|\mathbf{S}\|_{\mathrm{HS}}$, we have

$$\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{Q}}\|_{\mathrm{HS}} \le \frac{\varepsilon}{2} \sum_{n=1}^{\infty} \binom{2n}{n} \left(\frac{\varepsilon}{2} \right)^{2(n-1)} = \frac{2}{\varepsilon} \left(\frac{1}{\sqrt{1 - \varepsilon^2}} - 1 \right) \le \varepsilon,$$

which completes the proof.

4.2. *Proof of Theorem* 2. We say that a K-element subset (or K-tuple) of $\{X_1, \ldots, X_n\}$ is *diverse* if the latent labels of all points in the subset are distinct. Given some $\theta \in (0, \frac{\pi}{4})$, a K-tuple is θ-orthogonal if all its distinct pairs, when embedded, are orthogonal up to angle $\frac{\theta}{2}$. In order to establish (α, θ)-angular structure in the normalized Laplacian embedding of $\{X_1, \ldots, X_n\}$, we must show that there is a subset of $\{X_1, \ldots, X_n\}$ with at least $(1 - \alpha)n$ elements, and with the property that every diverse K-tuple from the subset is θ-orthogonal.

We break the proof into two steps. We first lower bound the total number of K-tuples that are diverse and θ-orthogonal. In the second step we construct the desired subset. We present the first step below and defer the second step to Appendix B.1 in the supplementary material [13].

Step 1. Consider a diverse K-tuple $(X^1, \ldots, X^K)$ constructed randomly by selecting X^m uniformly at random from the set $\{X_i | Z_i = m\}$ for $m = 1, \ldots, K$. Form the $K \times K$ random matrix

$$
V = \begin{bmatrix} | & & | \\ \Phi_{\mathcal{V}}(X^1) & \cdots & \Phi_{\mathcal{V}}(X^K) \\ | & & | \end{bmatrix},
$$

where $\Phi_{\mathcal{V}}$ denotes the normalized Laplacian embedding from equation (2.3). Let $\tilde{V}$ denote an independent copy of V. Let $Q \in \mathbb{R}^{K \times K}$ denote the diagonal matrix with entries $Q_{mm} = \frac{q_m(X^m)}{\|q_m\|_{\bar{\mathbb{P}}_n}}$, where $\bar{\mathbb{P}}_n$ is the empirical distribution over the samples $X_1, \ldots, X_n$, and define $Q_{\max} := \max_m \frac{\|q_m\|_{\infty}}{\|q_m\|_{\bar{\mathbb{P}}}}$.

At the core of our proof lies the following claim involving a constant c_3. For at least a fraction $1 - \frac{2Kc_3\varphi_n(\delta)}{\sqrt{w_{\min}}}$ of the diverse K-tuples, we have the inequality

$$
(4.6) \qquad \|V^T \tilde{V} - Q^2\|_{\mathrm{HS}} \le \frac{32\sqrt{3}}{w_{\min}^3} Q_{\max}^2 \sqrt{c_3} \sqrt{\varphi_n(\delta)},
$$

holding on a high probability set $\mathcal{A}$. For the moment, we take this claim as given, before returning to define $\mathcal{A}$ explicitly and prove the claim.

When inequality (4.6) is satisfied, we obtain the following upper bound on the off-diagonal elements of $V^T \tilde{V}$:

$$
(V^T \tilde{V})_{m\ell} \le \frac{32\sqrt{3} Q_{\max}^2}{w_{\min}^3} \sqrt{c_3} \sqrt{\varphi_n(\delta)} \qquad \text{for } m \ne \ell.
$$

This is useful because

$$
\cos \mathrm{angle}(\Phi_{\mathcal{V}}(X^m), \Phi_{\mathcal{V}}(X^\ell)) = \frac{(V^T \tilde{V})_{m\ell}}{\sqrt{(V^T V)_{mm}(V^T \tilde{V})_{\ell\ell}}}.
$$

However, we must also lower bound $\min_m (V^T \tilde{V})_{mm}$. To this end, by union bound, we obtain

$$\Pr\left\{\min_m Q_{mm}^2 \le t\right\} = \Pr\left\{\min_m \frac{q_m^2(X^m)}{\|q_m\|_{\mathbb{P}_n}^2} \le t\right\}$$

$$(4.7) \qquad\qquad \le \sum_{m=1}^{K} \Pr\left\{\frac{q_m^2(X^m)}{\|q_m\|_{\mathbb{P}_n}^2} \le t\right\}$$

$$:= \psi_n(t).$$

On the set $\mathcal{A}_\psi := \{\sup_t |\psi_n(t) - \psi(t)| \le \delta\} \subset \mathcal{A}$, we may combine equations (4.6) and (4.7) to obtain

$$\min_m (V^T \tilde{V})_{mm} \ge t - \frac{32\sqrt{3} Q_{\max}^2}{w_{\min}^3} \sqrt{c_3} \sqrt{\varphi_n(\delta)},$$

with probability at least $1 - \psi(2t)$. Therefore, there is a θ satisfying

$$|\cos\theta| \le \frac{32\sqrt{3} Q_{\max}^2 \sqrt{c_3} \sqrt{\varphi_n(\delta)}}{w_{\min}^3 t - 32\sqrt{3} Q_{\max}^2 \sqrt{c_3} \sqrt{\varphi_n(\delta)}}$$

such that at least a fraction $1 - \frac{2K c_3 \varphi_n(\delta)}{\sqrt{w_{\min}}} - \psi(2t)$ of the diverse K-tuples are θ-orthogonal on the set $\mathcal{A}$. This establishes the finite sample bound (3.18) with $c_0 := 2c_3$ and $c_1 := 32\sqrt{3} Q_{\max}^2 \sqrt{c_3}$.

It remains to prove the intermediate claim (4.6). Define the matrix

$$(4.8) \qquad A := \begin{bmatrix} \left\langle \dfrac{q_1}{\|q_1\|_{\bar{\mathbb{P}}_n}}, v_1 \right\rangle_{\bar{\mathbb{P}}_n} & \cdots & \left\langle \dfrac{q_1}{\|q_1\|_{\bar{\mathbb{P}}_n}}, v_K \right\rangle_{\bar{\mathbb{P}}_n} \\ \vdots & \ddots & \vdots \\ \left\langle \dfrac{q_K}{\|q_K\|_{\bar{\mathbb{P}}_n}}, v_1 \right\rangle_{\bar{\mathbb{P}}_n} & \cdots & \left\langle \dfrac{q_K}{\|q_K\|_{\bar{\mathbb{P}}_n}}, v_K \right\rangle_{\bar{\mathbb{P}}_n} \end{bmatrix}.$$

Note that the entries of AA^T are

$$(AA^T)_{m\ell} = \frac{\langle \Pi_\mathcal{V} q_m, \Pi_\mathcal{V} q_\ell \rangle_{\bar{\mathbb{P}}_n}}{\|q_m\|_{\bar{\mathbb{P}}_n} \|q_\ell\|_{\bar{\mathbb{P}}_n}}.$$

The off-diagonal elements satisfy

$$(AA^T)_{m\ell} \le 3(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}}) \qquad \text{for } m \ne \ell,$$

where $\hat{\varphi} = \max_m \frac{\|q_m - \Pi_\mathcal{V} q_m\|_{\bar{\mathbb{P}}_n}}{\|q_m\|_{\bar{\mathbb{P}}_n}}$, and $\hat{\mathcal{S}}_{\max} = \max_{m \ne \ell} \frac{\|q_\ell\|_{\mathbb{P}_m^n}^2}{\|q_m\|_{\bar{\mathbb{P}}_n}^2}$ (and $\mathbb{P}_m^n$ denotes the empirical distribution for the samples with latent label m). Similarly, the diagonal elements satisfy $|(AA^T)_{mm} - 1| \le 3\hat{\varphi}$. Putting together the pieces yields

$\|AA^T - I\|_{\mathrm{HS}}^2 \le 3K^2(\hat{\varphi} + \hat{\mathcal{S}}_{\max})$, which in turn implies

$$\|(AA^T)^{-1} - I\|_{\mathrm{HS}}^2 \le \frac{3K^2(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}})}{1 - 3K^2(\hat{\varphi} + \sqrt{\hat{\mathcal{S}}_{\max}})}.$$

We now transform this inequality into one involving $V^T \tilde{V}$. Write $B = AV$ and $\tilde{B} = A\tilde{V}$, and note that $V^T \tilde{V} = B^T (AA^T)^{-1} \tilde{B}$. Therefore, we find that

$$\|V^T \tilde{V} - Q^2\|_{\mathrm{HS}} \le \|B^T \tilde{B} - Q^2\|_{\mathrm{HS}} + \|B^T[(AA^T)^{-1} - I]\tilde{B}\|_{\mathrm{HS}}$$

$$\le 3\|Q\|_{\mathrm{HS}}\|B - Q\|_{\mathrm{HS}} + \|B\|_{\mathrm{HS}}^2\|(AA^T)^{-1} - I\|_{\mathrm{HS}},$$

where the last inequality used $\|B\|_{\mathrm{HS}} \le 2\|Q\|_{\mathrm{HS}}$. Now note that the entries of B are $B_{m\ell} = \frac{\Pi_{\mathcal{V}} q_m(X^\ell)}{\|q_m\|_{\tilde{\mathbb{P}}_n}}$. Therefore the difference $B - Q$ satisfies

$$(4.9) \quad \mathbb{E}[\|B - Q\|_{\mathrm{HS}}^2 | X_1, \ldots, X_n] \le K^2 \left(\frac{\hat{\varphi}}{\sqrt{\hat{w}_{\min}}} + \sqrt{\hat{\mathcal{S}}_{\max}} \right)^2 + K \frac{\hat{\varphi}^2}{\hat{w}_{\min}},$$

where $\hat{w}_{\min} = \min_m \frac{n_m}{n}$, and the expectation above is over the selection of the random K-tuple $(X^1, \ldots, X^K)$.[7]

Both $\hat{\mathcal{S}}_{\max}$ and $\hat{\varphi}$ are small with high probability. Indeed, Bernstein's inequality guarantees that

$$(4.10) \qquad\qquad \sqrt{\hat{\mathcal{S}}_{\max}} \le \sqrt{\mathcal{S}_{\max}} + \delta$$

with probability at least $1 - 2K^2 \exp \frac{-n(\mathcal{S}_{\max}+\delta^2)^2}{8Q_{\max}^2(2\mathcal{S}_{\max}+\delta^2)}$. We control $\hat{\varphi}$ with a finite sample version of Theorem 1, which we state as Proposition 2 below.

Let $\mathcal{V} = \mathrm{span}\{v_1, \ldots, v_K\}$ denote the principal eigenspace of the normalized Laplacian matrix.

PROPOSITION 2. *There are constants c_2', c_3 such that for any $\delta \in (0, \frac{\|k\|_{\tilde{\mathbb{P}}}}{b\sqrt{2\pi}})$ satisfying condition* (3.17), *we have*

$$(4.11) \qquad\qquad \hat{\varphi} \le c_3 \varphi_n(\delta)$$

with probability at least $1 - 10K \exp(\frac{-nc_2'\delta^4}{\delta^2 + \mathcal{S}_{\max}(\bar{\mathbb{P}}) + \mathcal{C}(\bar{\mathbb{P}})})$.

See Section 4.3 for the proof of this auxiliary result.

---

[7] Note that there are two different types of randomness at play in the construction of V and hence B; there is randomness in the generation of the i.i.d. samples $X_1, \ldots, X_n$ from $\bar{\mathbb{P}}$, and there is randomness in the selection of the diverse K-tuple $(X^1, \ldots, X^K)$.

On the set $\{\hat{\varphi} \leq c_3\varphi_n(\delta)\} \cap \{\hat{w}_{\min} \geq \frac{1}{2}w_{\min}\} := \mathcal{A}_\zeta \cap \mathcal{A}_w$, equation (4.9) simplifies to

$$\mathbb{E}\big[\|\!|B - Q\|\!|_{\mathrm{HS}}^2 | X_1, \ldots, X_n\big] \leq \frac{4K^2 c_3^2 \varphi_n^2(\delta)}{w_{\min}},$$

whenever $(\frac{\sqrt{2}c_3\varphi_n(\delta)}{\sqrt{w_{\min}}} + \mathcal{S}_{\max} + \delta)^2 \leq \frac{3c_3^2\varphi_n^2(\delta)}{w_{\min}}$, which is a consequence of condition (3.17). By Markov's inequality we obtain the following result: at least a fraction $1 - \frac{2Kc_3\varphi_n(\delta)}{\sqrt{w_{\min}}}$ of the diverse K-tuples satisfies

(4.12) $$\|\!|B - Q\|\!|_{\mathrm{HS}}{}^2 \leq \frac{2Kc_3\varphi_n}{\sqrt{w_{\min}}}.$$

For the diverse K-tuples that do satisfy inequality (4.12) we find that

$$\|\!|V^T\tilde{V} - Q^2\|\!|_{\mathrm{HS}} \leq \Big(\frac{6\sqrt{2}K^{3/2}Q_{\max}}{w_{\min}^{1/4}} + 32\sqrt{3}K^3 Q_{\max}^2\Big)\sqrt{c_3}\sqrt{\varphi_n},$$

valid on the set $\mathcal{A} = \mathcal{A}_w \cap \mathcal{A}_q \cap \mathcal{A}_\psi \cap \{\hat{\varphi} \leq c_3\varphi_n(\delta)\} \cap \{\sqrt{\hat{\mathcal{S}}_{\max}} \leq \sqrt{\mathcal{S}_{\max}} + \delta\}$, thereby establishing the bound (4.6).

To complete the first step of the proof of Theorem 2, it remains to control the probability of $\mathcal{A}$. By Hoeffding's inequality, we have $\mathbb{P}[\mathcal{A}_w] \geq 1 - Ke^{-nw_{\min}^2/2}$. Finally, an application of Bernstein's inequality controls the probability of $\mathcal{A}_q$, and an application of Glivenko–Cantelli controls the probability of $\mathcal{A}_\psi$. Putting together the pieces we find that $\mathcal{A}$ holds with probability at least $1 - 8K^2 \exp(\frac{-nc_2\delta^4}{\delta^2 + \mathcal{S}_{\max}(\mathbb{P}) + \mathcal{C}(\bar{\mathbb{P}})})$, where $c_2 := \min(c_2', \frac{1}{8Q_{\max}^2})$.

4.3. *Proof of Proposition* 2. Consider the operator $\hat{\mathbf{T}}: L^2(\bar{\mathbb{P}}_n) \to L^2(\bar{\mathbb{P}}_n)$ defined by

$$(\hat{\mathbf{T}}f)(x) = \int \frac{1}{\bar{q}_n(x)}k(x, y)\frac{f(y)}{\bar{q}_n(y)}\, d\bar{\mathbb{P}}_n(y),$$

where $\bar{q}_n(x) = \frac{1}{n}\sum_{i=1}^K k(X_i, x)$ is the square root kernelized density for the empirical distribution $\bar{\mathbb{P}}_n$ over the data $X_1, \ldots, X_n$. $\bar{k}^n(x, y) := \frac{k(x,y)}{\bar{q}_n(x)\bar{q}_n(y)}$ for the normalized kernel function. Note that for any $f \in L^2(\bar{\mathbb{P}}_n)$ and $v \in \mathbb{R}^n$ with coordinates $v_i = f(X_i)$, we have $(\hat{\mathbf{T}}f)(X_j) = (Lv)_j$, where L is the normalized Laplacian matrix (2.2). Consequently, the principal eigenspace $\mathcal{V}$ of L is isomorphic to the principal eigenspace of $\hat{\mathbf{T}}$ which we also denote by $\mathcal{V}$ for simplicity.

To prove the proposition, we must relate $\hat{\mathbf{T}}$ to the normalized Laplacian operator $\bar{\mathbf{T}}$. These operators differ in both their measures of integration—namely, $\bar{\mathbb{P}}_n$

versus $\bar{\mathbb{P}}$—and their kernels, namely $\frac{k(x,y)}{\bar{q}_n(x)\bar{q}_n(y)}$ versus $\frac{k(x,y)}{\bar{q}(x)\bar{q}(y)}$. To bridge the gap we introduce an intermediate operator $\tilde{\mathbf{T}} : L^2(\bar{\mathbb{P}}_n) \to L^2(\bar{\mathbb{P}}_n)$ defined by

$$(\tilde{\mathbf{T}}f)(x) = \int \frac{1}{\bar{q}(x)} k(x,y) \frac{f(y)}{\bar{q}(y)} \, d\bar{\mathbb{P}}_n(y).$$

Let $\tilde{\mathcal{V}}$ denote the principal eigenspace of $\tilde{\mathbf{T}}$. The following lemma bounds the ρ-distance between the principal eigenspaces of $\tilde{\mathbf{T}}$ and $\hat{\mathbf{T}}$.

LEMMA 3. *For any* $\delta \in [0, \frac{\|k\|_{\bar{\mathbb{P}}}}{b\sqrt{2\pi}}]$ *satisfying condition* (3.17), *we have*

$$(4.13) \qquad \rho(\mathcal{V}, \tilde{\mathcal{V}}) \le \frac{c_4}{\Gamma^2} \left(\frac{1}{\sqrt{n}} + \delta \right),$$

with probability at least $1 - 4e^{-n\pi\delta^2/4} - 2e^{-n\mathbf{E}\bar{k}\delta^2}$, *where* $c_4 = 1024\sqrt{2\pi K} \frac{\|k\|_{\bar{\mathbb{P}}}b}{r^4}$.

See Appendix B.3 (supplementary material [13]) for a proof of this lemma. We must upper bound $\|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n}$. By the triangle inequality,

$$\|q_m - \Pi_{\mathcal{V}} q_m\|_{\bar{\mathbb{P}}_n} \le \|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n} + \|\Pi_{\mathcal{R}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n}$$
$$+ \|\Pi_{\mathcal{V}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n}.$$

Note that $\|\Pi_{\mathcal{V}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n} \le \|q_m\|_{\bar{\mathbb{P}}_n} \rho(\tilde{\mathcal{V}}, \mathcal{V})$. We can control this term with the lemma. The term $\|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n}$ is the empirical version of a quantity controlled by Theorem 1. We handle the empirical fluctuations with a version of Bernstein's inequality. For $\delta_p \ge 0$ we have the inequality

$$(4.14) \qquad \|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}_n} \le \|q_m - \Pi_{\mathcal{R}} q_m\|_{\bar{\mathbb{P}}} + \delta_p$$

with probability at least $1 - 2\exp(-\frac{n\delta_p^4}{8(\delta_p^2 + c_{\mathrm{pop}}^2 \varphi^2)\tilde{Q}_{\max}^2})$, where

$$\tilde{Q}_{\max} = \max_m \frac{\|q_m - \Pi_{\mathcal{R}} q_m\|_\infty}{\|q_m\|_{\bar{\mathbb{P}}}} \quad \text{and} \quad c_{\mathrm{pop}} := 16\sqrt{12 + \frac{b_{\max}}{K}}.$$

It remains to control $\|\Pi_{\mathcal{R}} q_m - \Pi_{\tilde{\mathcal{V}}} q_m\|_{\bar{\mathbb{P}}_n}$. Let $\bar{\mathcal{H}}$ denote the reproducing kernel Hilbert space (RKHS)[8] for the kernel $\bar{k}$. Now we define two integral operators on $\bar{\mathcal{H}}$. Let $\bar{\mathbf{H}}$ denote the operator defined by

$$(\bar{\mathbf{H}}h)(x) = \int \bar{k}(x,y)h(y) \, d\bar{\mathbb{P}}(y),$$

---

[8]We give a brief introduction to the theory of reproducing kernel Hilbert spaces and provide some references for further reading on the subject in Appendix C.2 (supplementary material [13]).

and similarly let $\tilde{\mathbf{H}} : \bar{\mathcal{H}} \to \bar{\mathcal{H}}$ denote the operator defined by

$$(\tilde{\mathbf{H}}h)(x) = \int \bar{k}(x, y)h(y) \, d\bar{\mathbb{P}}_n(y).$$

Both $\bar{\mathbf{H}}$ and $\tilde{\mathbf{H}}$ are self-adjoint, compact operators on $\bar{\mathcal{H}}$ and have real, discrete spectra. Let $\mathcal{G}$ denote the principal K-dimensional eigenspace of $\bar{\mathbf{H}}$, and let $\tilde{\mathcal{G}}$ denote the principal K-dimensional principal eigenspace of $\tilde{\mathbf{H}}$. The following lemma bounds the ρ-distance between these subspaces of $\bar{\mathcal{H}}$.

LEMMA 4.   *For any $\delta > 0$ satisfying condition* (3.17), *we have*

$$\rho(\mathcal{G}, \tilde{\mathcal{G}}) \le \frac{c_5}{\Gamma^2}\left(\frac{1}{\sqrt{n}} + \delta\right)$$

with probability at least $1 - 2e^{-n\pi\mathbb{E}\bar{k}(\bar{X}, \bar{X})\delta^2}$, *where* $c_5 = 64\sqrt{2\pi K}\sqrt{\mathbb{E}\bar{k}(\bar{X}, \bar{X})}\frac{b}{r^2}$.

See Appendix B.2 (supplementary material [13]) for the proof of this lemma. By the triangle inequality, we have

$$\|\Pi_{\mathcal{R}}q_m - \Pi_{\tilde{\mathcal{V}}}q_m\|_{\bar{\mathbb{P}}_n} \le \|\Pi_{\mathcal{R}}q_m - \Pi_{\mathcal{G}}q_m\|_{\bar{\mathbb{P}}_n} + \|\Pi_{\mathcal{G}}q_m - \Pi_{\tilde{\mathcal{G}}}q_m\|_{\bar{\mathbb{P}}_n}$$
$$+ \|\Pi_{\tilde{\mathcal{G}}}q_m - \Pi_{\tilde{\mathcal{V}}}q_m\|_{\bar{\mathbb{P}}_n}.$$

We claim that

(4.15) $\|\Pi_{\mathcal{R}}q_m - \Pi_{\mathcal{G}}q_m\|_{\bar{\mathbb{P}}_n} = 0$ and $\|\Pi_{\tilde{\mathcal{G}}}q_m - \Pi_{\tilde{\mathcal{V}}}q_m\|_{\bar{\mathbb{P}}_n} = 0.$

We take these identities as given for the moment, before returning to prove them at the end of this subsection.

Now the term $\|\Pi_{\mathcal{G}}q_m - \Pi_{\tilde{\mathcal{G}}}q_m\|_{\bar{\mathbb{P}}_n}$ can be controlled using the lemma in the following way. For any $h \in \bar{\mathcal{H}}$, note that

$$\|h\|_{\bar{\mathbb{P}}_n}^2 = \frac{1}{n}\sum_{i=1}^n \langle h, \bar{k}_{X_i}\rangle_{\bar{\mathcal{H}}}^2 \le \frac{1}{n}\sum_{i=1}^n \|h\|_{\bar{\mathcal{H}}}^2 \bar{k}(X_i, X_i)$$

by Cauchy–Schwarz for the RKHS inner product. Using this logic with $h = \Pi_{\mathcal{G}}q_m - \Pi_{\tilde{\mathcal{G}}}q_m$, we find

(4.16) $\|\Pi_{\mathcal{G}}q_m - \Pi_{\tilde{\mathcal{G}}}q_m\|_{\bar{\mathbb{P}}_n} \le \|q_m\|_{\bar{\mathcal{H}}}\sqrt{\frac{1}{n}\sum_{i=1}^n \bar{k}(X_i, X_i)}\rho(\mathcal{G}, \tilde{\mathcal{G}}).$

Collecting our results and applying Lemmas 3 and 4 yields

$$\|q_m - \Pi_{\mathcal{V}}q_m\|_{\bar{\mathbb{P}}_n} \le (c_{\text{pop}}\varphi + \delta_p)\|q_m\|_{\bar{\mathbb{P}}} + \frac{c_n\|q_m\|_{\bar{\mathbb{P}}_n}}{\Gamma^2}\left(\frac{1}{\sqrt{n}} + \delta\right),$$

where

$$(4.17) \qquad c_n := \frac{256\sqrt{2\pi K}b}{r^2}\left[\frac{\|q_m\|_{\bar{\mathcal{H}}}}{\|q_m\|_{\bar{\mathbb{P}}}}\mathbb{E}\bar{k}(\bar{X},\bar{X}) + \frac{2b}{r^2}\right].$$

By an application of Bernstein's inequality, we have

$$\|q_m\|_{\bar{\mathbb{P}}_n} \leq \sqrt{2}\|q_m\|_{\bar{\mathbb{P}}}$$

with probability at least $1 - 2e^{-n/(16Q_{\max}^2)}$. For $\delta \in (0, \frac{1}{2\sqrt{2\pi}Q_{\max}})$, we have

$$2e^{(-nc_{\mathrm{pop}}^2\delta^4)/(8\Gamma^4\tilde{Q}_{\max}^2(\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}})))} + 6e^{-n\pi\delta^2/2} + 2e^{-n/(16Q_{\max}^2)}$$

$$\leq 10e^{-(nc_2'\delta^4)/(\delta^2+\mathcal{S}_{\max}(\bar{\mathbb{P}})+\mathcal{C}(\bar{\mathbb{P}}))},$$

where $\delta_p = \frac{c_{\mathrm{pop}}\delta}{\Gamma^2}$, and $c_2' = \min(\frac{c_{\mathrm{pop}}^2}{8\Gamma^4\tilde{Q}_{\max}^2}, \frac{\pi}{2})$. Modulo the claim, this proves the proposition with $c_3 = 2\max(c_{\mathrm{pop}}, c_n)$.

We now return to prove claim (4.15). Note the following relation between the eigenfunctions of $\bar{\mathbf{T}}$ and those of $\bar{\mathbf{H}}$: if r_i is an eigenfunction of $\bar{\mathbf{T}}$ with eigenvalue λ_i and $\|r_i\|_{\bar{\mathbb{P}}} = 1$, then $g_i := \sqrt{\lambda_i}r_i$ has unit norm in $\bar{\mathcal{H}}$ and is an eigenfunction of $\bar{\mathbf{H}}$ with eigenvalue λ_i. Note that the eigenfunctions r_i of $\bar{\mathbf{T}}$ form an orthonormal basis of $L^2(\bar{\mathbb{P}})$, and therefore $q_m = \sum_{i=1}^\infty a_i r_i$, where a_i are the coefficients $\langle q_m, r_i\rangle_{\bar{\mathbb{P}}}$. By the observation above, we have the equivalent representation $q_m = \sum_{i=1}^\infty \frac{a_i}{\sqrt{\lambda_i}}g_i$. Therefore the $L^2(\bar{\mathbb{P}})$ projection onto $\mathcal{R} = \mathrm{span}\{r_1, \ldots, r_K\}$ is $\Pi_{\mathcal{R}}q_m = \sum_{i=1}^K a_i r_i$, and the $\bar{\mathcal{H}}$ projection onto $\mathcal{G} = \mathrm{span}\{g_1, \ldots, g_K\}$ is $\Pi_{\mathcal{G}}q_m = \sum_{i=1}^K \frac{a_i}{\sqrt{\lambda_i}}g_i$. Therefore the relation $g_i = \sqrt{\lambda_i}r_i$ implies $\|\Pi_{\mathcal{R}} - \Pi_{\mathcal{G}}\|_{\bar{\mathbb{P}}_n} = 0$. Similar reasoning yields $\|\Pi_{\tilde{\mathcal{G}}}q_\ell - \Pi_{\tilde{\mathcal{V}}}q_\ell\|_{\bar{\mathbb{P}}_n} = 0$.

5. Discussion. In this paper, we have analyzed the performance of spectral clustering in the context of nonparametric finite mixture models. Our first main contribution is an upper bound on the distance between the population level normalized Laplacian embedding and the square root kernelized density embedding. This bound depends on the maximal similarity index, the coupling parameter, and the indivisibility parameter. These parameters all depend on the kernel function, and we present our analysis for a fixed but arbitrary kernel.

Although this dependence on the kernel function might seem undesirable, it is actually necessary to guarantee identifiability of the mixture components in the following sense. A mixture with fully nonparametric components is a very rich model class: without any restrictions on the mixture components, any distribution can be written as a K-component mixture in uncountably many ways. Conversely, when the clustering difficulty function is zero, the representation of a distribution as a mixture is unique. In principle, one could optimize over the convex cone of symmetric positive definite kernel functions so to minimize our clustering difficulty

parameter. In our preliminary numerical experiments, we have found promising results in using this strategy to choose the bandwidth in a family of kernels.

Building on our population-level result, we have also provided a result that characterizes the normalized Laplacian embedding when applied to a finite collection of n i.i.d. samples. We find that when the clustering difficulty is small, the embedded samples take on approximate orthogonal structure: samples from different components are almost orthogonal with high probability. The emergence of this form of angular structure allows an angular version of K-means to correctly label most of the samples.

Perhaps surprising is the fact that the optimal bandwidth (minimizing our upper bound) is nonzero. Although we only provide an upper bound, we believe this is fundamental to spectral clustering, not an artifact of our analysis. Again, the principal K-dimensional eigenspace of the Laplacian operator is not a well-defined mathematical object when the bandwidth is zero. Indeed, as the bandwidth shrinks to zero, the eigengap distinguishing this eigenspace from the remaining eigenfucntion vanishes. This eigenspace, however, is the population-level version of the subspace onto which spectral clustering projects. For this reason, we caution against shrinking the bandwidth indefinitely to zero, and we conjecture that there is an optimal population level bandwidth for spectral clustering. However, we should mention that we cannot provably rule out the optimality of an appropriately slowly shrinking bandwidth, and we leave this to future work. Further investigation of kernel bandwidth selection for spectral clustering is an interesting avenue for future work.

SUPPLEMENTARY MATERIAL

Appendix: Remaining proofs and background material (DOI: 10.1214/14-AOS1283SUPP; .pdf). Due to space constraints, we relegate technical details of the remaining proofs to the supplement [13]. This supplementary appendix also gives an overview of some useful background material, and it includes a reference list for the symbols used in this paper.

REFERENCES

[1] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

[2] BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

 [3] CAO, Y. and CHEN, D.-R. (2011). Consistency of regularized spectral clustering. *Appl. Comput*. *Harmon*. *Anal*. **30** 319–336. MR2784567

 [4] DONATH, W. E. and HOFFMAN, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM J*. *Res*. *Develop*. **17** 420–425. MR0329965

 [5] FIEDLER, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math*. *J*. **23** 298–305. MR0318007

 [6] GINÉ, E. and KOLTCHINSKII, V. (2006). Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results. In *High Dimensional Probability*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **51** 238–259. IMS, Beachwood, OH. MR2387773

 [7] JENSSEN, R. (2010). Kernel entropy component analysis. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **32** 847–860.

 [8] LAWLER, G. F. and SOKAL, A. D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality. *Trans*. *Amer*. *Math*. *Soc*. **309** 557–580. MR0930082

 [9] MEILA, M. and SHI, J. (2001). A random walks view of spectral segmentation. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, Key West, FL.

[10] NG, A. Y., JORDAN, M. I. and WEISS, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (*NIPS*), Vancouver, BC, Canada.

[11] REED, M. and SIMON, B. (1980). *Methods of Modern Mathematical Physics*. *I*: *Functional Analysis*, 2nd ed. Academic Press, New York. MR0751959

[12] ROSASCO, L., BELKIN, M. and DE VITO, E. (2010). On learning with integral operators. *J*. *Mach*. *Learn*. *Res*. **11** 905–934. MR2600634

[13] SCHIEBINGER, G., WAINWRIGHT, M. J. and YU, B. (2015). Supplement to "The geometry of kernelized spectral clustering." DOI:10.1214/14-AOS1283SUPP.

[14] SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **22** 888–905.

[15] SHI, T., BELKIN, M. and YU, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann*. *Statist*. **37** 3960–3984. MR2572449

[16] STEWART, G. W. (1971). Error bounds for approximate invariant subspaces of closed linear operators. *SIAM J*. *Numer*. *Anal*. **8** 796–808. MR0293831

[17] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann*. *Statist*. **36** 555–586. MR2396807

[18] YAN, D., CHEN, A. and JORDAN, M. I. (2013). Cluster forests. *Comput*. *Statist*. *Data Anal*. **66** 178–192. MR3064033

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: gschiebinger@gmail.com
            wainwrig@stat.berkeley.edu
            binyu@stat.berkeley.edu