**Orthogonal Cone Structure of Dimensionality Reduction Embeddings**

By

RUI HU

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

---

Wolfgang Polonik, Chair

---

Krishnakumar Balasubramanian

---

James Sharpnack

Committee in Charge

2023

# Contents


Abstract iv

Acknowledgments v

Chapter 1. Introduction and motivation 1

Chapter 2. Setup and Main Results 11

2.1. Basic setting 11

2.2. OCS and parameters for OCS control 21

2.3. Main results 24

Chapter 3. Discussion 37

3.1. Weighted overlapping, coupling and indivisibility parameters 37

3.2. Overlapping parameters and eigen-tail paramater 39

3.3. Angles, coverage and radius 39

3.4. Performance of $k$-means clustering under OCS 41

3.5. Examples 44

Chapter 4. Proof of main results 50

4.1. OCS of spectral embedding: The population setting 51

4.2. OCS of spectral embedding: The sample setting 68

4.3. OCS of kernel PCA embedding: The population setting 90

4.4. OCS of kernel PCA embedding: The sample setting 97

4.5. Strong Version of OCS 102

Chapter 5. Simulations 108

5.1. Numerical experiments illustrating the theoretical results 108

5.2. The inverse problem: Does the OCS contain information about the separateness of the mixture model? 113

Chapter 6. Conclusions and future work 117

Appendix A. Behaviors of incorrect coverage ratio with respect to some selected parameters 119

Bibliography 123

Orthogonal Cone Structure of Dimensionality Reduction Embeddings

**Abstract**

We analyze geometric aspects of clustering procedures based on low-dimensional embeddings. In particular, we are interested in understanding the occurrence of the so-called orthogonal cone structure (OCS) that can be observed empirically in various low-dimensional embeddings, including kernel PCA, spectral clustering, Isomap, and clustering based on the Hodge Laplacian. Inspired by recent work on the OCS based on graph Laplacians, we study OCS in the context of weighted Laplacian and kernel PCA. This involves the development of a notion of a well-separated mixture model and other characteristics of the methodology. These characteristics are then used to quantify the OCS. We illustrate this for weighted Laplacian and kernel PCA in both the population setting and the sample setting.

**Acknowledgments**

This dissertation would not have been possible without the support of my advisor, committee members, colleagues, and friends. I am deeply grateful for their guidance and assistance throughout my graduate study, particularly during my time at the University of California, Davis (UC Davis).

I would like to express my utmost appreciation to my advisor, Prof. Wolfgang Polonik, who deserves my sincerest gratitude. Our initial meeting in the Mathematical Statistics course left a lasting impression on me, as his strong statistical expertise, patient teaching, and enthusiasm inspired me to embark on my research journey. He introduced me to the fascinating field of Topological Data Analysis (TDA), high-dimensional statistics, manifold learning, and related dimensionality reduction methods. His unwavering support and valuable insights have been instrumental in shaping my research ideas. Whenever I faced challenges and technical difficulties in my projects, Wolfgang's encouragement and unwavering belief in my abilities kept me motivated. He generously shared his experiences, which proved invaluable during my job search. It is thanks to his guidance that I secured my first academic position immediately after completing my Ph.D., becoming an assistant professor in my area of expertise.

I am also grateful to Prof. Krishnakumar Balasubramanian, who contributed to my Ph.D. research for three years. His intelligence, open-mindedness, and impressive research contributions have been instrumental in overcoming various technical obstacles. Even during Wolfgang's absence from campus, Prof. Balasubramanian provided invaluable support, ensuring the continuity of my Ph.D. work.

Furthermore, I extend my sincere thanks to the members of my defense committee: Prof. James Sharpnack, Prof. Naoki Saito, and Prof. Bala Rajaratnam. I am fortunate to have taken courses from each of them, acquiring a wealth of statistical knowledge and insights into cutting-edge research questions. Their presence at my Ph.D. defense, along with their valuable comments and suggestions, greatly enriched my work.

Numerous other professors have played significant roles during my Ph.D. journey. I am indebted to Prof. Alexander Aue, Prof. Xiaodong Li, and Prof. Xiucai Ding for their invaluable assistance during my job search. It has been an honor to work as a teaching assistant for their courses on multiple occasions. Prof. Debashis Paul and Prof. Miles Lopes enlightened me in my first quarter

at UC Davis, instilling in me a deep understanding of statistics and its magic. Prof. Hans-Georg Müller, my master's advisor, provided invaluable guidance when I first arrived at UC Davis, while Prof. Jane-Ling Wang broadened my perspectives on statistical consulting and allowed me to contribute statistical insights to diverse research domains, including food science, plant science, and psychology. Prof. Jie Peng introduced me to statistical methods in research and effective teaching techniques. Their contributions to my teaching, research, and academic life have been indispensable. I also had the privilege of completing two summer internships in data science and machine learning, and I am grateful to my mentor, Dr. Jin Cao, for providing me with valuable industry research experience. These internships allowed me to bridge the gap between academic statistics and real-life problems.

I would also like to express my gratitude to my friends and colleagues. Dr. Jeremy Halim, who invited me to collaborate on his food science research project, enabled my first publication in statistics. My roommate, Yue Kang, who will also be graduating next year, has been a constant source of engaging discussions on cutting-edge statistical research and technical details within our own work.

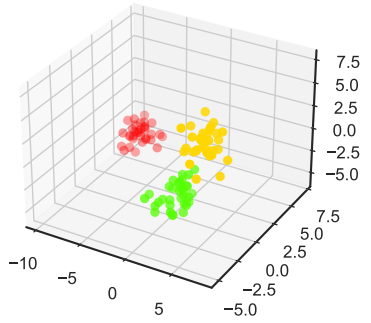
Lastly, but certainly not least, I would like to extend my deepest appreciation to my parents, Xiusen Hu and Yuexian Duan. Their unwavering support and assistance in various aspects of my life have been invaluable, and I am grateful for the opportunity to share my achievements with them.

CHAPTER 1

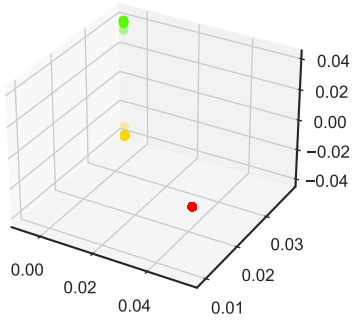
# Introduction and motivation

In this thesis, we explore geometric aspects of low-dimensional embeddings and their impact on clustering procedures. In particular, we study kernel PCA and weighted Laplacian embeddings where the latter can be considered as a generalization of spectral graph clustering. Our analyses are based on studying the so-called "orthogonal cone structure" (OCS). This describes a striking geometric feature that can be observed in low-dimensional embeddings (see Figure 1.1, Figure 1.2 and Figure 1.3). Many existing popular clustering algorithms use dimension reduction techniques as preprocessing step, which motivates the exploration of the conditions giving rise to such geometric structures to appear and the exploration of the information contained in this structures. We will see that if data are sampled from a well separated mixture model, i.e. different components have no or little overlap (a more formal definition will be given below), then the embedded data is more probable to exhibit an OCS, meaning that, with high probability, a large proportion of embedded data from different components will fall into different orthogonal cones. Again, this heuristic statement will be made precise below. Several parameters describing properties of the underlying mixture model will be defined that then also describe and quantify the existence of an OCS, and this quantification will prove that a strong OCS guarantees a successful clustering performance of the $k$-means algorithm with uniformly orthonormal vectors as random initialization. An inverse problem also occurs: When observing the OCS from one embedded dataset, can one conclude that the data is sampled from a well-separated mixture model? This problem is investigated through simulations studies, which indicate that a strong OCS is an indication of a well-separated mixture model.

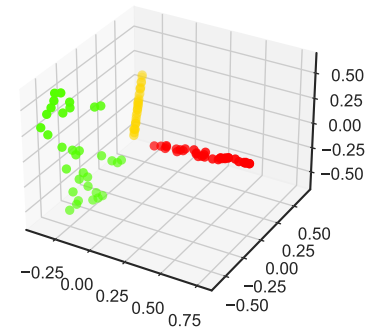
The OCS can be observed in many clustering algorithms. Figure 1.1, Figure 1.2 and Figure 1.3 show some OCS examples for different kernel embeddings, including graph Laplacian, kernel PCA, and Isomap. The only difference among the three data sets, all simulated from a mixture of 3-dimensional normals with covariance $\sigma^2$ times the identity matrix, is the standard deviation $\sigma$.



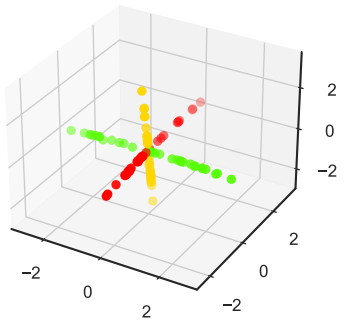

(a) Sample Data




(b) Graph Laplacian embedding (10 nearest neighbors)




(c) Kernel PCA embedding $\left(\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{3}\right)\right)$

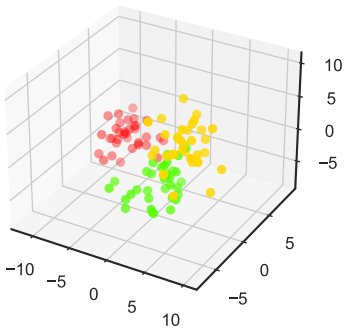



(d) Isomap embedding (5 nearest neighbors)

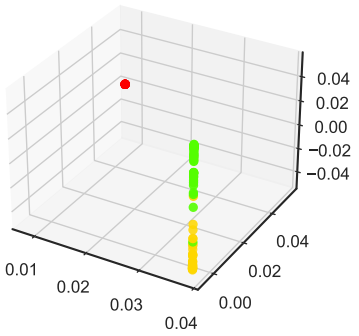
**Figure 1.1.** Examples of Orthogonal Cone Structure for different embeddings, where the standard deviation of each cluster is 1.

We plot the eigenvectors corresponding to the first three smallest eigenvalues of graph Laplacian embedding, and the eigenvectors corresponding to the first three largest eigenvalues corresponding to Kernel PCA and Isomap, respectively, for all the three data sets. The kernels and tuning parameters are kept the same in the three examples. Details can be found in the captions of those figures. It is important to understand whether this geometric structure is induced by the method itself artificially or by the underlying structure of the model. This may have effects on the performance of clustering and the interpretation of the embeddings. Thus, insights into the OCS can be helpful for improving the data analysis.

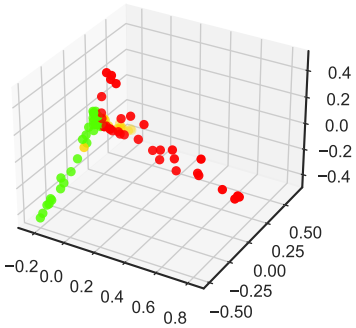
In order to heuristically explain the OCS phenomenon, consider a graph Laplacian $L$ of an undirected, unweighted graph. $L$ has the form $L = D - A$ with $D$ the degree matrix and $A$ the binary (0-1



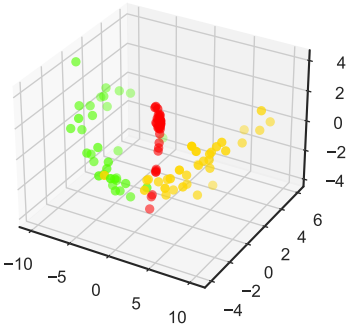
(a) Sample Data



(b) Graph Laplacian embedding (10 nearest neighbors)



(c) Kernel PCA embedding $\left(\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{3}\right)\right)$



(d) Isomap embedding (5 nearest neighbors)

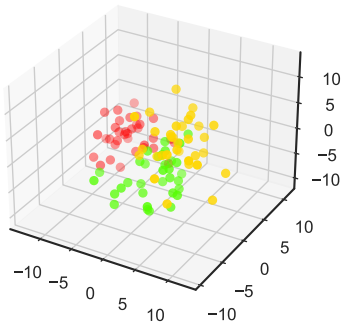
**Figure 1.2.** Examples of Orthogonal Cone Structure for different embeddings, where the standard deviation of each cluster is 2.

valued) adjacency matrix. The graph Laplacian embedding is then defined by $F : \mathbb{R}^d \to \mathbb{R}^N (N \leq d)$

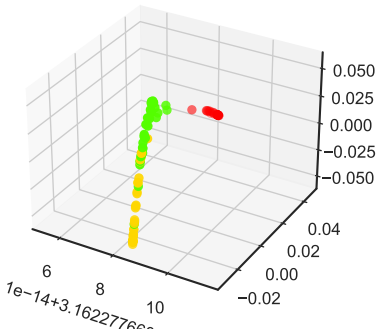
$$F : \mathbf{x}_i \longmapsto \begin{pmatrix} u_{1i} \\ \vdots \\ u_{Ni} \end{pmatrix} \text{ for } i = 1, \cdots, n,$$

where $d$ is the dimension of the original data, $u_1, \ldots, u_N$ are the $N$ eigenvectors corresponding to the $N$ smallest eigenvalues of $L$ and $u_{ki}$ denotes the $i$th element of vector $u_k$.

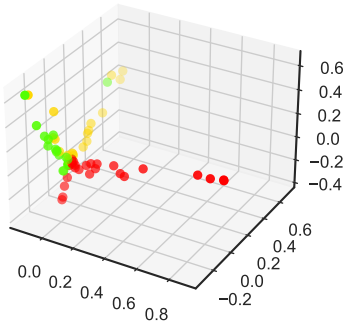
Now consider a simple scenario where the $m$-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$ has $N$ connected components. Suppose that $n$ data points $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$ are uniformly distributed on $\mathcal{M}$, then one can construct the proximity graph consisting of $N$ connected components. The $N$ eigenvectors corresponding to the first smallest $N$ eigenvalues of the corresponding graph Laplacian (considered as functions on the n data points) coincide with rescaled versions of the indicator



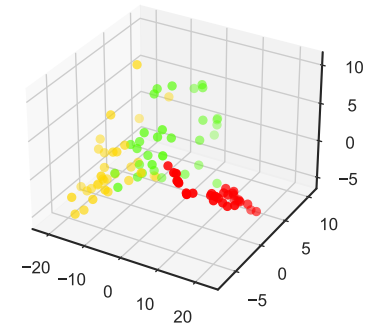
(a) Sample Data



(b) Graph Laplacian embedding (10 nearest neighbors)



(c) Kernel PCA embedding $\left(\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{3}\right)\right)$



(d) Isomap embedding (5 nearest neighbors)

**Figure 1.3.** Examples of Orthogonal Cone Structure for different embeddings, where the standard deviation of each cluster is 3.

functions of the $N$ connected components of $\mathcal{M}$, so that the resulting graph Laplacian embedding map sends the original data set into a set of $N$ orthogonal vectors on $\mathbb{R}^N$. This is an extreme case of an OCS: The embedding consists of N orthogonal cones with common cone tips at the origin and with opening angle 0. The question then is, how exactly can the embedding be described when the underlying data are drawn from a mixture distribution which are not entirely separated. This is where the OCS with non-degenerate cones comes into play. A rigorous definition of an OCS is as follows.

DEFINITION 1. *(Orthogonal Cone Structure) Let* $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4), \delta \in [0, 1)$, *and* $r > 0$. *A probability measure* $\mu \in \mathcal{P}\left(\mathbb{R}^N\right)$ *has an orthogonal cone structure with parameters* $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$

*if there exists an orthonormal basis for* $\mathbb{R}^N, e_1, \ldots, e_N$, *such that*

$$\mu\left(\bigcup_{j=1}^{N} C\left(e_j, \sigma_j, r\right)\right) \geq 1 - \delta,$$

*where* $C\left(e_j, \sigma_j, r\right)$ *is the set*

$$C\left(e_j, \sigma_j, r\right) \coloneqq \left\{ z \in \mathbb{R}^k : \frac{z \cdot e_j}{|z|} > \cos(\sigma_j), \quad |z| > r \right\}.$$

Notice that each set $C$ is a spherical cone with cone tip in the origin, intersected with the complement of a ball of radius $r$. The OCS has first been defined in Schiebinger et al. ( [**75**]) in the context of the simple graph Laplacian and it was further studied in Garcia-Trillos et al. ( [**89**]).

An OCS is not a rare structure but appears in any probability measure: One can always choose $\sigma_j$'s large enough for the union of the cones to cover the entire $\mathbb{R}^N$, resulting in a trivial OCS with $\delta = 0$. More generally, an OCS with $\sigma_j$'s close to $\pi/4$ and $\delta$ close to 1, is a weak OCS and a clustering algorithm applied to the embeddings will in general not perform well. The goal of this thesis is to explore in the setting of a finite mixture model, how the separateness of the individual model components effects the geometric parameters of the OCS. One can conclude that given a well-separated mixture model (formally defined later), one only needs small angles to achieve a large coverage by the cones. Informally but intuitively, an OCS with small angles and large coverage is called a 'good' or a 'strong' OCS. The sequence of Figure 1.1, Figure 1.2 and Figure 1.3 show a change from 'good' OCS to 'bad' OCS when the clusters move closer to each other, i.e. when they become less separated.

REMARK 1. *The above definition of an OCS guarantees coverage proportions of the union of the orthogonal cones, but the cones are not required to cover one specific component. A modified (strong) version of OCS for mixture models addresses this. Let* $\nu = \sum_{j=1}^{N} w_j \nu_j$ *be a mixture measure with positive weights* $w_j$, $j = 1, 2, \cdots, N$, *and let* $F_\sharp$ *denote the push-forward operator through some given embedding* $F$. *Then we say that a mixture measure* $\mu \coloneqq F_\sharp \nu \in \mathcal{P}\left(\mathbb{R}^N\right)$ *(*$\mu = \sum_{j=1}^{N} w_j \mu_j$, *where* $\mu_j = F_\sharp \nu_j$*) has a strong orthogonal cone structure with parameters* $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ *if*

*there exists an orthonormal basis* $e_1, e_2, \cdots, e_N$ *of* $\mathbb{R}^N$*, such that*

$$\sum_{j=1}^{N} w_j \mu_j \left( C\left(e_j, \sigma_j, r\right)\right) \geq 1 - \delta,$$

*where* $C\left(e_j, \sigma_j, r\right)$ *is the same set defined above.*

Here we study the OCS for *weighted* graph Laplacians and Kernel PCA. This choice constitutes two basic by different types of examples, where the weighted Laplacian operator corresponds to a differential operator while the Kernel PCA operator corresponds to an integral operator.

**Related Work.** To the best of our knowledge, the first work analyzing the OCS is Schiebinger, Wainwright and Yu (2015)( [**75**]), where kernelized spectral clustering is discussed. Inspired by this work, Garcia Trillos, Hoffman and Hosseini (2019)( [**89**]) also study the OCS for kernelized spectral clustering, but they are using a slightly different and more general approach than in Schiebinger et al., also allowing for data sampled from a manifold.

Spectral properties of Graph Laplacians play crucial roles in unsupervised and semi-supervised learning algorithms. Hoffmann et al. (2022)( [**47**]) studied the large data limit of scaled graph Laplacians, which approach limiting continuum operators. García Trillos et al. (2020)( [**37**]) showed convergence of eigenvalues and eigenvectors of graph Laplacian to the eigenvalues and eigenfunctions of the weighted Laplace-Beltrami operator. Hein et al. (2007)( [**45**]) determined the pointwise limit of graph Laplacians as the sample size increases and the neighborhood size approaches zero. Burago et al. (2015)( [**21**]) also showed such convergence in a different scenario where a proximity graph on an epsilon-net is considered. Giné et al. (2006)( [**40**]) proved a.s. and distributional convergence of graph Laplacians to Laplace-Beltrami operator. Koltchinskii et al. (2000)( [**57**]) gave a general result of random matrix approximation for spectra of integral operators. Bühler et al. (2009)( [**20**]) used graph $p$-Laplacian, a nonlinear generalization of the standard graph Laplacian to generalize the standard spectral clustering.

Principal component analysis (PCA) (Jolliffe, 1986)( [**54**]) is a well-used dimensionality reduction method with the idea of linearly projecting high-dimensional data to a lower dimensional subspace by retaining variability in the data. Schölkopf et al. (1998)( [**77**]) used kernel trick to extend the idea of PCA to reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950)( [**5**]) resulting in a

nonlinear dimension reduction method, which is kernel PCA. Kernel PCA is widely used in many applications. Sriperumbudur and Sterge (2018)( [**83**]) approximate kernel PCA by using random features and consider the computational and statistical trade-off. Blanchard, Bousquet and Zwald (2007)( [**18**]) considered the statistical properties of kernel PCA and prove concentration bounds for the reconstruction error, and also obtained convergence bounds for the partial sums of the biggest or smallest eigenvalues of the kernel Gram matrix towards eigenvalues of the corresponding kernel operator. Reiss and Wahl (2019)( [**96**]) analyzed the reconstruction error of PCA, and Cai and Zhang (2020)( [**23**]) gave different optimal rates for singular spaces under perturbation, which are applicable to a wide range of dimension reduction methods. Koltchinskii et al. (2020)( [**58**]) established the asymptotic normality and asymptotic properties for the risk of the estimators of linear functionals for eigenvectors of the covariance operator, and also proved matching minimax lower bounds of the estimators. Koltchinskii and Lounici (2016, 2017)( [**59**], [**60**], [**61**]) derived sharp concentration bounds for bilinear forms of empirical spectral projection in terms of sample size and effective dimension, and they also derived concentration inequalities and expectation bounds for the operator norm of the difference between covariance operator and its empirical version. Bengio et al. (2004)( [**13**]) showed relation between spectral embedding methods and kernel PCA and gave the level of error to the effect of small perturbations of the training set on the embedding. Albeverio et al. (2008)( [**2**]) considered all types of self-adjoint perturbations of a semi-bounded operator under the framework of additive perturbation theory (Kato, 2013)( [**56**]). Jirak and Wahl (2018, 2020)( [**52**], [**53**]) gave perturbation bounds for eigenspace of covariance operators and their empirical version under a relative gap condition. Wahl (2019)( [**96**]) also proves the analogue of standard perturbation result under a weighted condition, which leads to significant improvements in random perturbations. Fukumizu et al. (2007)( [**36**])proved statistical convergence of kernel CCA, which can be seen as an bivariate generalization of kernel PCA.

Identifiability of finite mixtures in a mixture model is essential for the exploration of OCS, which will be quantified by one important parameter called indivisibility parameter (formally defined later) related to OCS. Teicher (1963)( [**86**]) gave some results on identifiability of finite mixtures of some well-used distributions. Aragam et al. (2020)( [**4**]) also introduced a novel framework involving clustering overfitted parametric mixture models to establish general conditions of identifiability of nonparametric mixture models.

Non-linear dimension reduction methods are also closely related to manifold learning problems ( [**51**]), since these methodologies involve capturing the local and global structure of the underlying manifold in order to be able to obtain useful estimates of geodesic distances, which subsequently result in good clustering performance.

We also noticed that the top eigenvectors of operator don't always show good clustering result. Shi, Belkin and Yu (2009)( [**81**]) devised a clustering algorithm that selects only those eigenvectors which have clustering information not represented by the other eigenvectors already selected when the top eigenvectors is inadequate and redundant at the same time, which could appear when the clusters are not balanced and/or have different shapes.

We study the OCS in both the population setting and sample setting for kernel PCA case and weighted Laplacian case. The structures of the proofs is motivated by the proofs in Garcia Trillos, Hoffmann and Hosseini (2019)( [**89**]). Our main contribution includes the following:

- Prove the OCS in both the population setting and the sample setting of kernel PCA case and weighted Laplacian case under similar assumptions by using ideas laid out by Garcia Trillos, Hoffmann and Hosseini (2019)( [**89**]).
- Generalize the definition of OCS to allow different angle for different cone.
- Explore the role of bandwidth of the kernel function and the power parameter of weighted Laplacian in the context of the OCS.
- Derive the sufficient condition that $k$-means algorithm with uniformly orthonormal vectors as random initialization clusters most proportion of the data points correctly under the context of OCS.
- Conduct numerical explorations of the inverse problem: Can we infer any properties of the mixture model if the OCS of embedded data is observed?

Following ideas laid out in Garcia Trillos et al.( [**89**]), our basic approach is as follows:

(i) First, we show the OCS for the limit operator (i.e. the population case).

(ii) Then, we analyze the limit behavior of the corresponding top (or bottom) empirical eigenvectors of the matrix that is used in low-dimensional embeddings. This involves to determine the

corresponding limit operator, and to show that the eigenvectors converge in an appropriate sense to the eigenfunction of the limit operator.

(iii) The convergence of the eigenvectors (to the true eigenfunctions) will imply the OCS for the eigenvectors of our empirical matrices, which leads to the OCS of these low-dimensional embeddings (i.e. the sample case).

In the weighted Laplacian setting, we use a variant of the graph discretization method (Burago, Ivanov, and Kurylev, 2015)( [**21**]) to connect the weighted Laplacian operator and its empirical version. Discretization of eigenfunctions and interpolation of eigenvectors are used here to define the convergence of eigenvectors to eigenfunctions and they are inverse operators of each other. The error estimates method of spectral convergence (García Trillos, Gerlach, Hein and Slepčev, 2020)( [**37**]) is also modified to be applied on our weighted Laplacian setting.

In the Kernel PCA setting, Davis-Kahan theorem (Davis and Kahan, 1970)( [**31**]) and its variant (Yu, Wang and Samworth, 2015)( [**101**]) are used to bound the distance between subspaces spanned by population eigenfunctions and their empirical versions, which require eigenvalue separation condition of the corresponding operators. In order to prove OCS in the sample setting, the behavior of eigenfunctions of kernel PCA embedding operator and its empirical version also plays essential role, and also a Davis-Kahan type theorem is needed in the proofs.

The basic ideas underlying the derivation of the OCS in these two cases are similar, but details are quite different. In the former case, we use a $q$-th power of the mixture density scaled to integrate to one as an auxiliary step. We start from showing OCS of such rescaled measure and bound the distance of it and our desired measure to prove the population setting. Some useful error bounds can be established and leads to OCS of original measure under suitable conditions for the sample setting. For the latter case, in the population setting, similar steps apply but density assumption is no longer necessary and we may use kernelized density as the auxiliary step. Also, error bounds in Kernel PCA case are also constructed by using different ideas where we directly consider the closeness of spectrum between the covariance operator and its empirical version. The work just mentioned deals with spectral convergence, i.e. the convergence of eigenvalues and eigenfunctions (in a sense to be specified,) of matrices to limit objects that consist of eigenvalues and eigenfunction of operators.

This involves devising a methodology that allows to formulate convergence of eigenvectors to a limit function. This also plays an integral role in the proofs of the main results of this thesis.

In a word, the overarching goal of this thesis is to understand commonalities and differences between low-dimensional embeddings used by clustering procedures through the lens of the limiting operators and the OCS, and we consider Kernel PCA and weighted Laplacian as specific instances. Both population setting and sample setting are considered and the OCS under these settings are completely proved under some suitable conditions, which will be given before we introduce the main theorems. Notice that weighted Laplacian operator is one of the differential operator while kernel PCA operator is one of the integral operator. One of our future work is to further explore the commonality and difference between these two cases and to generalize our results in order to also cover other dimension reduction embeddings.

CHAPTER 2

# Setup and Main Results

In this section, we introduce the precise settings for both the weighted Laplacian and the Kernel PCA cases in both the population and the sample settings. In the Kernel PCA case, the existence of densities is not assumed. Our interest is to quantify the OCS in terms of well-separation properties of the assumed underlying mixture distribution. These separation properties are described by various parameters. In the weighted Laplacian case, we use parameters named weighted overlapping parameter, coupling parameter and indivisibility parameter, while in the Kernel PCA case, two different overlapping parameters are being used, and we also need to control the eigen-decay of the covariance operator. Precise definitions will be given below, where we also discuss all these parameters along with their interpretation in detail. Our main results then specify how the OCS can be quantitatively described by these parameters. If the model is well separated, then the OCS will be strong, and what we precisely mean by that will be made clear below. Intuitively, an OCS consists of the existence of spherically symmetric orthogonal cones in the embedding space that carry a high mass concentration. In the sample setting a strong OCS means that with high probability these orthogonal cones cover a high proportion of embeddings, and different cones will cover embeddings corresponding to different clusters. A precise definition of the OCS property will be given below.

## 2.1. Basic setting

**2.1.1. Weighted Laplacian and related differential operator.** We first introduce the definition of weighted Laplacian operator and weighted graph Laplacian, where the weighted Laplacian operator arises as a limit (in an appropriate sense) of the weighted graph Laplacian. These convergences play an integral role in the proofs of the main theorems.

Let $\mathcal{Z} \subset \mathbb{R}^d$ be bounded, we consider differential operators of the form

$$\begin{cases} \mathcal{L}u \coloneqq -\frac{1}{\rho^p} \operatorname{div}\left(\rho^q \nabla\left(u\right)\right), & \text{in } \mathcal{Z} \\ \rho^q \frac{\partial u}{\partial n} = 0, & \text{on } \partial\mathcal{Z}, \end{cases} \tag{2.1}$$

for parameters $p, q \in \mathbb{R}$ fixed, where $\nabla$ denotes gradient, div denotes divergence, and $\partial$ denotes partial differential, and we assume here that all these quantities are well defined.

The previous differential operators arise as large data limits of graph Laplacian operators of the form

$$\Delta_n \coloneqq \begin{cases} D_n^{\frac{1-p}{q-1}} \left(D_n - W_n\right), & \text{if } q \neq 1 \\ D_n - W_n, & \text{if } q = 1, \end{cases} \tag{2.2}$$

where $W_n = W_n(q)$ is a symmetric weighted adjacency matrix and it is defined by a pre-specified kernel which contains the information of similarities between empirical data points, and $D_n = D_n(q)$ is a weighted degree matrix computed based on $W_n$. The formal definitions will be given below before presenting the main results for the OCS of these graph Laplacians. We say that $\mathcal{L}$ is a large data limit of $\Delta_n$ in the sense that the eigenvalues and eigenvectors of $\Delta_n$ converge to the eigenvalues and eigenfunctions of $\mathcal{L}$ as the sample size tends to infinity, in a sense that will be described below in Lemma 7 and Lemma 4, which indicate that the difference of the discretization of eigenfunctions and eigenvectors, as well as the difference of the interpolation of eigenvectors and eigenfunctions are bounded, respectively. Also, these bounds can be arbitrarily small as long as the sample size $n$ is large enough. Rigorous proofs and related details are given after Lemma 4.

REMARK 2. *Hoffmann et al.( [47]) considered spectral analysis of a more general three parameter class of differential operators of the form*

$$\begin{cases} \mathcal{L}u \coloneqq -\frac{1}{\rho^p} \operatorname{div}\left(\rho^q \nabla\left(\frac{u}{\rho^s}\right)\right), & \textit{in } \mathcal{Z} \\ \rho^q \frac{\partial}{\partial n}\left(\frac{u}{\rho^s}\right) = 0, & \textit{on } \partial\mathcal{Z}, \end{cases} \tag{2.3}$$

| $p$ | $q$ | $s$ | $\mathcal{L}u$ | $\Delta_n$ | Unnormalized or Normalized |
|---|---|---|---|---|---|
| 1 | 2 | 0 | $-\frac{1}{\rho}\operatorname{div}\left(\rho^2\nabla(u)\right)$ | $D_n - W_n$ | Unnormalized graph Laplacian |
| 3/2 | 2 | 1/2 | $-\frac{1}{\rho^{3/2}}\operatorname{div}\left(\rho^2\nabla(u\rho^{-1/2})\right)$ | $D_n^{-1/2}(D_n - W_n)D_n^{-1/2}$ | Normalized graph Laplacian |
| 2 | 2 | 0 | $-\frac{1}{\rho^2}\operatorname{div}\left(\rho^2\nabla(u)\right)$ | $D_n^{-1}(D_n - W_n)$ | Normalized graph Laplacian |

**Table 2.1.** Examples of different choices of $(p, q, s)$ and corresponding differential operators and graph Laplacians.

*for parameters $p, q, s \in \mathbb{R}$ fixed, which also arise as large data limits of graph Laplacian operators of the form*

$$\Delta_n := \begin{cases} D_n^{\frac{1-p}{q-1}} (D_n - W_n) D_n^{-\frac{s}{q-1}}, & \text{if } q \neq 1 \\ D_n - W_n, & \text{if } q = 1. \end{cases} \tag{2.4}$$

The three-parameter family of differential operators and the corresponding three-parameter family of graph Laplacians are introduced here to unify various expressions of different normalizations of the graph Laplacian. Some special cases and related convergence properties are popular and well-studied. Several examples are listed in Table 2.1. Below we will consider the case where $\mathcal{Z}$ is a manifold embedded in $\mathbb{R}^d$.

The above definition of the graph Laplacian assume a given graph. In the literature, various constructions of such graphs are being considered. Two popular choices are the $\varepsilon$-graph and the $k$-NN graph (see below for details). Under appropriate assumptions, eigenvalues and eigenvectors of the corresponding graph Laplacians converge (in a sense to be specified) to the corresponding eigenvalues and eigenvectors of a continuous limit operator which belong to the class of operators introduced above ( [**24**]).

More precisely, assume we have an i.i.d. sample $X = \{x_1, \cdots, x_n\} \in \mathbb{R}^d$. Given a measure $\mu$, denote its associated empirical measure by $\mu_n$. In the following part, we will use the notation $L^2(\mu)$ to denote the space of $L^2$-functions with respect to the measure $\mu$, and by $L^2(\mu_n)$ the space of functions $u : X \to \mathbb{R}$.

Given an $\varepsilon > 0$, the weighted $\varepsilon$-graph $G^\varepsilon = (X, w^\varepsilon)$ is constructed by the following steps: An edge is first put between $x_i$ and $x_j$ if $|x_i - x_j| \leq \varepsilon$, where $|x_i - x_j|$ is the Euclidean distance between points $x_i$ and $x_j$. A non-increasing Lipschitz continuous function $\eta : [0, \infty) \to [0, \infty)$ is introduced

to endow weights to edges and $\eta$ is supported on $[0,1]$, i.e. $\eta(x)=0$ for $x>1$. The weight is then defined by $w_{xy}^{\varepsilon}=\eta\left(\frac{|x-y|}{\varepsilon}\right)$. Note that if the points $x_i$, $x_j$ are not connected then the weight is 0. Then the associated graph Laplacian is given by

$$\mathcal{L}^{\varepsilon}u(x)=\frac{1}{n\varepsilon^{m+2}}\sum_{j=1}^{n}w_{x_jx}^{\varepsilon}\left(u(x)-u\left(x_j\right)\right),$$

where $u(\cdot)$ is a function in $L^2\left(\mu_n\right)$.

The corresponding limit differential operator $\Delta_\rho$ for a smooth function $f$ turns out to be

$$\Delta_\rho f:=-\frac{1}{2\rho}\operatorname{div}\left(\rho^2\nabla f\right).$$

A different graph construction on $X$ proceeds not by fixing a length-scale $\varepsilon$ but rather by specifying for each point in $X$ a set of nearest neighbors, which is called $k$-NN graph. The (undirected) $k$-NN graph is constructed by connecting a pair of points if one is in the $k$ nearest neighbors of another one. More precisely, we first let $N_\varepsilon(x)=\sum_{j=1}^{n}\mathbf{1}_{0<|x_j-x|\leq\varepsilon}$ be the number of random samples in a Euclidean $\varepsilon$-neighborhood of $x$. Given $1\leq k\leq n-1$, define $\varepsilon_k(x):=\min\left\{\varepsilon>0:N_\varepsilon(x)\geq k\right\}$, which is the Euclidean distance from $x$ to the $k$-th nearest neighbor of $x$ from the samples $x_1,x_2,\cdots,x_n$. Finally, we define $r_k(x,y):=\max\left\{\varepsilon_k(x),\varepsilon_k(y)\right\}$. Then $x_i$ and $x_j$ are connected if and only if $|x_i-x_j|\leq r_k(x_i,x_j)$. (The mutual $k$-NN graph can be constructed by setting $r_k(x,y)=\min\left\{\varepsilon_k(x),\varepsilon_k(y)\right\}$.) Then the undirected $k$-NN graph Laplacian of $u\in L^2\left(\mu_n\right)$ is defined as

$$\mathcal{L}^{k}u(x)=\frac{1}{n}\left(\frac{n\alpha_m}{k}\right)^{1+2/m}\sum_{j=1}^{n}w_{x_jx}^{r_k(x_j,x)}\left(u(x)-u\left(x_j\right)\right),$$

where $\alpha_m$ is the volume of the $m$-dimensional Euclidean unit ball, and the weights have the same form as above.

The corresponding limit differential operator $\Delta_\rho^{NN}$ for a smooth function $f$ turns out to be

$$\Delta_\rho^{NN}f:=-\frac{1}{2\rho}\operatorname{div}\left(\rho^{1-2/m}\nabla f\right).$$

Given a weight matrix $W$ and degree matrix $D$, the *unnormalized graph Laplacian* is simply defined ( [**94**]) as

$$L=D-W,$$

and there are two matrices, called *normalized graph Laplacians*, are defined as

$$L_{\text{sym}} := D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2},$$

and

$$L_{\text{rw}} := D^{-1}L = I - D^{-1}W.$$

The normalization here is based on the degree matrix which itself depends on the weight matrix $W$. However, also the normalization of weights are being considered. This then gives rise to two normalizations at different levels: First the weights are normalized and then the Laplacian might be normalized by using the degree matrix based on the normalized weights. One thus has to carefully distinguish between normalized weights and normalized Laplacians.

The corresponding limit differential operator of these three matrices turn out to be ( [**92**]):

$$\mathcal{L} : u \mapsto -\frac{1}{\rho}\operatorname{div}\left(\rho^2\nabla u\right),$$

$$\mathcal{L}^{sym} : u \mapsto -\frac{1}{\rho^{3/2}}\operatorname{div}\left(\rho^2\nabla\left(\frac{u}{\sqrt{\rho}}\right)\right),$$

and

$$\mathcal{L}^{rw}(u) = -\frac{1}{\rho^2}\operatorname{div}\left(\rho^2\nabla u\right).$$

They are just special cases of our weighted Laplacian for the parameters $(p,q,s) = (1,2,0)$, $(p,q,s) = (\frac{3}{2},2,\frac{1}{2})$, and $(p,q,s) = (2,2,0)$, respectively. The limit behaviors and convergence properties are explored based on spectral convergence ( [**88**]), which will also be applied in the major proof part of our main theorems.

There is yet another notion of a Laplacian operator used in the literature, which is called the $p$-Laplacian. Recall that the standard graph Laplacian $\Delta_2$ can be defined as the operator which induces the quadratic form for a smooth function $f$ as

$$\langle f, \Delta_2 f\rangle = \frac{1}{2}\sum_{i,j=1}^{n} w_{x_ix_j}\left(f(x_i) - f(x_j)\right)^2.$$

The $p$-Laplacian is defined similarly as the operator which induces the quadratic form for a smooth function $f$ as

$$\langle f, \Delta_p f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{x_i x_j} \left| f(x_i) - f(x_j) \right|^p .$$

The $p$-Laplacian regularization can be added in a family of regression problems in a semi-supervised setting ( [**82**]). By using the same notations as above, in the constrained model in this problem, the estimator is constructed by minimizing

$$\mathcal{E}_n^{(p)}(f) = \frac{1}{\varepsilon_n^p} \frac{1}{n^2} \sum_{i,j=1}^{n} W_{ij} \left| f\left(x_i\right) - f\left(x_j\right) \right|^p$$

among $\{f : \Omega_n \to \mathbb{R}\}$ which satisfy the constraint $f\left(x_i\right) = y_i$ for all $i = 1, \cdots, N$, where $|\cdot|$ denotes the Euclidean distance in the ambient space $\mathbb{R}^d$. Then for $q > 0$, the penalization term is defined by

$$R^{(q)}(f) = \sum_{i=1}^{N} \left| y_i - f\left(x_i\right) \right|^q .$$

Then the penalized estimator is constructed by minimizing

$$\mathcal{S}_n^{(p)}(f) = \mathcal{E}_n^{(p)}(f) + \lambda R^{(q)}(f)$$

over all functions $f : \Omega_n \to \mathbb{R}$, where $\lambda > 0$ is a tuning parameter. The limit behaviors of objective function $\mathcal{E}_n^{(p)}(f)$ and the penalized estimator $\mathcal{S}_n^{(p)}(f)$ are explored in this problem when $n \to \infty$. The following continuum functionals describe the limiting problems as $n \to \infty$:

$$\mathcal{E}_\infty^{(p)}(f) = \begin{cases} \sigma_\eta \int_\Omega |\nabla f(x)|^p \rho^2(x) \mathrm{d}x & \text{if } f \in W^{1,p}(\Omega) \\ \infty & \text{else,} \end{cases}$$

where $W^{1,p}(\Omega)$ denotes a Sobolev space. Also, to describe the limit of the penalized model in the large data limit, the functional

$$\mathcal{S}_\infty^{(p)}(f) = \mathcal{E}_\infty^{(p)}(f) + \lambda R^{(q)}(f)$$

is introduced and is well defined whenever $p > d$.

In this thesis, we just consider the special case of $s = 0$ in the differential operators (2.3), which is exactly the differential operators (2.1). This special version of weighted Laplacian will be explored

and main theorems are based on this special case. More assumptions and techniques are needed to construct similar results for the generalized version where $s$ is not necessarily zero.

As already given above, the relevant differential operator for the case $s = 0$ is differential operator $\Delta_\rho(u)$ for smooth functions $u$ defined by

$$\Delta_\rho(u) = -\frac{1}{\rho^p}\operatorname{div}(\rho^q \nabla u) = -\rho^{q-p}\Delta u - q\rho^{q-p-1}\nabla\rho\nabla u.$$

Given a density $\rho$ on $\mathcal{M}$ we define the weighted function spaces

$$L^2(\mathcal{M}, \rho^q) := \{u : \mathcal{M} \mapsto \mathbb{R} \mid \langle u, u\rangle_{\rho^q} < +\infty\},$$

equipped with the inner product

$$\langle u, v\rangle_{\rho^q} := \int_{\mathcal{M}} u(x)v(x)\rho^q(x)dx.$$

We also use $L^2(\rho^q)$ instead of $L^2(\mathcal{M}, \rho^q)$ when the domain of the function space is clear with no ambiguity. With $d\nu(x) = \frac{\rho^q(x)dx}{\int \rho^q(t)dt}$ we write $L^2(\nu)$ and $\langle\cdot,\cdot\rangle_{L^2(\nu)}$ to denote $L^2(\mathcal{M}, \rho^q)$ and the corresponding dot product $\langle\cdot,\cdot\rangle_{\rho^q}$.

We will investigate the dependence of the OCS on the parameters $p$ and $q$, which are two non-negative real values . The number of components $N$ in the mixture model is also fixed throughout this dissertation.

The Algorithm of weighted Laplacian clustering has the same structure with Graph Laplacian clustering. In this algorithm, we first construct the weighted similarity matrix and its corresponding degree matrix and then compute the weighted Laplacian matrix. By doing clustering on the embedded data constructed by elements in the eigenvectors corresponding to the first several smallest eigenvalues, we get the cluster label of the original data. Detailed steps are listed in the following algorithm:

**Algorithm 1:** weighted Laplacian clustering

**Input:** $(n \times d)$ data matrix $X = [x_1, x_2, \cdots, x_n]^T$ ($n$ observations of $d$ dimensional data), number $N$ of clusters to construct, a non-increasing Lipschitz function $\eta$, a bandwidth $\varepsilon$;

Compute the weighted similarity matrix $\tilde{W}_n$ with entries $\left(\tilde{W}_n\right)_{ij} := \eta_\varepsilon\left(|x_i - x_j|\right)\mathbf{1}(i \neq j)$ and its re-weighted version $W_n$ with entries $(W_n)_{ij} = \frac{(\tilde{W}_n)_{ij}}{\tilde{d}_i^{1-q/2}\tilde{d}_j^{1-q/2}}$;

Compute the corresponding degree matrix $D_n = \mathrm{diag}\,(d_i)$ with $d_i := \sum_{j=1}^N W_{ij}$;

Compute the weighted Laplacian $\Delta_n$ as defined above;

Compute the $N$ eigenvectors $u_1, \cdots, u_N$ corresponding to the smallest $N$ eigenvalues of $\Delta_n$;

Let $U \in \mathbb{R}^{n\times d}$ be the matrix containing the vectors $u_1, \cdots, u_N$ as columns;

For $i = 1, 2, \cdots, n$, let $y_i \in \mathbb{R}^d$ be the vector corresponding to the $i$-th row of $U$;

Cluster the points $(y_i)_{i=1,2,\ldots,n}$ in $\mathbf{R}^N$ with the $k$-means algorithm into clusters $C_1, C_2, \cdots, C_N$;

**Output:** Clusters $A_1, A_2 \cdots, A_N$ with $A_i = \{j | y_j \in C_i\}$.

**2.1.2. Reproducing kernel Hilbert space and kernel PCA embedding.** We study the geometry of a continuum analogue Kernel PCA embedding. Kernel PCA just extends the idea of classical PCA in a Reproducing Kernel Hilbert Space (RKHS). The definition of a RKHS is as follows: Let $k$ be a symmetric positive definite kernel function

$$k : \Omega \times \Omega \longrightarrow \mathbb{R}$$

$$(s,t) \longmapsto k(s,t)$$

defined on $\Omega \times \Omega$ ($\Omega$ is a non empty abstract set). We call $k$ a reproducing kernel of the Hilbert space $\mathcal{H}$ if and only if

- $\forall t \in \Omega, \quad k(\cdot, t) \in \mathcal{H}$,
- $\forall t \in \Omega, \quad \forall \varphi \in \mathcal{H} \quad \langle \varphi, k(., t)\rangle = \varphi(t)$ ("the reproducing property").

A Hilbert space $\mathcal{H}$ possessing a reproducing kernel is called an RKHS (e.g., see [**17**] and [**68**]).

From the previous two properties, we can derive that

$$\forall (s,t) \in \Omega \times \Omega \quad k(s,t) = \langle k(\cdot, t), k(\cdot, s)\rangle.$$

In the following part, let $\nu$ denote a measure on $\Omega$. Define the kernelized density of $\nu$ as the function $q \in \mathcal{H} \subset L^2(\Omega, \nu)$ given by

$$q(x) = \int_\Omega k(x, y)\nu(dy), \tag{2.5}$$

where $k(x, y)$ is a reproducing kernel satisfying $|k(x, y)| \leq M$ for some $M > 0$. Notice that $q(\cdot) = \int_\Omega k(\cdot, x)\nu(dx) = \mathbb{E}_\nu k(\cdot, X)$, where $X$ is a random variable with measure $\mu$. So this kernelized density is the unique mean element of the measure $\nu$.

Similarly, we denote the kernelized density of $\nu_k$ by $q_k$, i.e.

$$q_k(x) = \int_\Omega k(x, y)\nu_k(dy). \tag{2.6}$$

We use the standard notation $L_2(\Omega, \nu)$ to denote the class of real-valued functions on $\Omega$ with finite $L^2$-norm with respect to $\nu$, i.e.

$$L^2(\Omega, \nu) \coloneqq \{u : \Omega \mapsto \mathbb{R} \mid \langle u, u\rangle_\nu < +\infty\},$$

where

$$\langle u, v\rangle_\nu \coloneqq \int_\Omega u(x)v(x)\nu(dx).$$

Define the centered kernel

$$\begin{aligned} \bar{k}(x, y) &= k(x, y) - \mathbb{E}_\nu k(x, Y) - \mathbb{E}_\nu k(X, y) + \mathbb{E}_{\nu\otimes\nu} k(X, Y) \\ &= k(x, y) - q(x) - q(y) + \int_\Omega \int_\Omega k(x, y)\nu(dx)\nu(dy) \\ &= k(x, y) - q(x) - q(y) + \mathbb{E}_\nu q(X), \end{aligned} \tag{2.7}$$

and the centered covariance operator

$$\Sigma_\nu f(\cdot) = \int_\Omega \bar{k}(\cdot, y)f(y)\nu(dy), \quad f \in \mathcal{H}(k). \tag{2.8}$$

For $x, y \in \mathcal{H}$, denote $x \otimes_\mathcal{H} y$ as an element of the tensor product space $\mathcal{H} \otimes \mathcal{H}$ which can also be seen as an operator from $\mathcal{H}$ to $\mathcal{H}$ as $(x \otimes_\mathcal{H} y)z = x\langle y, z\rangle_\mathcal{H}$ for any $z \in \mathcal{H}$.

It is obvious that $\Sigma_\nu$ is self-adjoint and under the assumption that the kernel is bounded, $\Sigma_\nu$ is also a trace-class operator and therefore Hilbert-Schmidt and compact. So by the spectral theorem, $\Sigma_\nu$

can be written as

$$\Sigma_\nu = \sum_{i \in I} \lambda_i u_i \otimes_{\mathcal{H}} u_i,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots$ denote the eigenvalues of $\Sigma_\nu$ in non-increasing order, and the associated orthonormal eigenfunctions are denoted as $u_1(\cdot), u_2(\cdot), \cdots$.

Then we may compute the kernelized desity $q(x)$ again by using the centered kernel. For convenience, we continue using the notation of $k(x, y)$ and $q(x)$ to denote these quantities. Then, $q_k$ has the representation

$$q_k = \sum_{l=1}^{\infty} a_{lk} u_l$$

with coefficients $a_{lk}, l = 1, 2, \cdots$.

The Algorithm of Kernel PCA clustering is well-used in many applications. In this algorithm, we first choose a suitable kernel and compute the inner-product (Gram) matrix based on the data, and then get the double-centered version of this Gram matrix. By using the eigenvectors corresponding to the first several leading eigenvalues, we obtain the embeddings that then are being clustered. Detailed steps are listed in the following algorithm:

---

**Algorithm 2:** Kernel PCA clustering

---

**Input:** $(n \times d)$ data matrix $X = [x_1, x_2, \cdots, x_n]^T$ ($n$ observations of $d$ dimensional data), number $N$ of clusters to construct, kernel $K(x_i, x_j)$;

Compute the inner-product (Gram) matrix $\mathbf{G}_{(n*n)}$ of the data set $V$, where $\mathbf{G}_{ij} = K(x_i, x_j)$;

Compute the double-centered version of the Gram matrix $\mathbf{K} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{G}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$;

Compute the $N$ eigenvectors $u_1, \cdots, u_N$ corresponding to the largest $N$ eigenvalues of $\mathbf{K}$;

**for** $i = 1 : n$ **do**

  Let $y_i = (y_{1i}, \cdots, y_{Ni})$ with $y_{ti} = \sum_{j=1}^{n} u_{tj} K(x_i, x_j)$ for $t = 1, \cdots, N$.

**end**

Cluster the points $(y_i)_{i=1,2,\ldots,n}$ in $\mathbf{R}^N$ with the $k$-means algorithm into clusters $C_1, C_2, \cdots, C_N$;

**Output:** Clusters $A_1, A_2 \cdots, A_N$ with $A_i = \{j | y_j \in C_i\}$.

---

### 2.2. OCS and parameters for OCS control

We are now defining the various parameters describing the separateness of the mixture components. As indicated above, they are somewhat different in the two cases considered here, namely spectral clustering using the weighted Laplacian and kernel PCA.

**2.2.1. Weighted Laplacian case.** Consider a smooth (infinitely differentiable), connected, orientable, $m$-dimensional compact Riemannian manifold $\mathcal{M}$ in $\mathbb{R}^d$. For a fixed positive integer $N$, let $\rho_1, \rho_2, \cdots, \rho_N$ denote probability density function on $\mathcal{M}$ with respect to the volume form on $\mathcal{M}$. Furthermore, let $w_1, w_2, \cdots, w_N$ be strictly positive weights which satisfy that $\sum_{k=1}^N w_k = 1$. Then the corresponding mixture density is given by

$$\rho(x) \coloneqq \sum_{k=1}^{N} w_k \rho_k(x), \quad x \in \mathcal{M}.$$

Assume that $\rho^q$ is integrable on $\mathcal{M}$. Let $\nu$ be the probability measure on $\mathcal{M}$ with density $\tilde{\rho}(x) \coloneqq \frac{\rho^q(x)}{\int \rho^q(t)dt}$. Note that $\tilde{\rho}$ depends on the parameter $q$. We also write

$$d\nu(x) = \frac{\rho^q(x)dx}{\int \rho^q(t)dt},$$

where $dx$ denotes integration with respect to $\mathcal{M}$'s volume form and $d\nu(x)$ is a complete Borel probability measures on $\mathcal{M}$.

The following assumption guarantees that the parameters related to OCS defined in this subsection are well-defined.

Assumption 1. *Assume that $\rho_1, \rho_2, \cdots, \rho_N \in C^1(\mathcal{M})$ with $\int_{\mathcal{M}} \rho_k(x)dx = 1$ for all $k$ are such that the operators $\Delta_{\rho_k}$ and $\Delta_\rho$ have discrete point spectrums with an associated orthonormal basis of eigenfunctions for $L^2(\rho_k^q)$ and $L^2(\rho^q)$, respectively.*

For $k = 1, \cdots, N$, define the functions

$$q_k \coloneqq \left(\sqrt{\frac{w_k \rho_k}{\rho}}\right)^q,$$

and with $I_j = \int \rho_j^q(x)dx$ and $I = \int \rho^q(x)dx$ let

$$I_{\min} \coloneqq \min_{i=1,2,\cdots,N} I_i \quad \text{and} \quad I_{\max} \coloneqq \max_{i=1,2,\cdots,N} I_i.$$

Moreover, for $i, j = 1, 2, \cdots, N$, let

$$\mathcal{S}_{ij} = \left\langle \left( \frac{q_i}{\sqrt{(w_i)^q}} \right)^2, \left( \frac{q_j}{\sqrt{(w_j)^q}} \right)^2 \right\rangle_{\rho^q} = \left\langle \left( \frac{\rho_i}{\rho} \right)^q, \left( \frac{\rho_j}{\rho} \right)^q \right\rangle_{\rho^q} = \int_{\mathcal{M}} \left( \frac{\rho_i(x)\rho_j(x)}{\rho(x)} \right)^q dx,$$

and define

$$\mathcal{S}_{\mathrm{b}} \coloneqq \max_{i,j=1,2,\cdots,N, i\neq j} \mathcal{S}_{ij} \quad \text{and} \quad \mathcal{S}_{\mathrm{w}} \coloneqq \max_{i=1,2,\cdots,N} \mathcal{S}_{ii},$$

and

$$\mathcal{S}_{\mathrm{adj}} \coloneqq \max_{j=1,2,\cdots,N} \left( \max(N^{q-1}, 1)\mathcal{S}_{\mathrm{b}} \sum_{k\neq j} w_k^q + \left( \max(N^{q-1}, 1) - 1 \right) w_j^q \mathcal{S}_{\mathrm{w}} \right),$$

where the subscript 'b' means 'between', 'w' means 'within', and 'adj' means 'adjust'.

With this notation we now define the following parameters:

**Overlapping parameter (similarity parameter)**

$$\overline{\mathcal{S}_i} \coloneqq \sum_{j=1,\cdots,N, j\neq i} w_j^q \mathcal{S}_{ij}.$$

**Coupling parameter**

$$\mathcal{C} \coloneqq \max_{k=1,\cdots,N} \mathcal{C}_k,$$

where

$$\mathcal{C}_k \coloneqq \frac{q^2}{4} \int \left| \frac{\nabla \rho_k}{\rho_k} - \frac{\nabla \rho}{\rho} \right|^2 \rho_k^q dx, \quad k = 1, \cdots, N.$$

**Indivisibility parameter**

$$\Theta \coloneqq \min_{k=1,\cdots,N} \Theta_k,$$

where

$$\Theta_k \coloneqq \inf_{u \perp \mathbf{1}} \frac{\int |\nabla u|^2 \rho_k^q dx}{\langle u, u \rangle_{\rho_k^q}},$$

where $\mathbf{1}$ denotes constant function and $\perp$ denotes orthogonality with respect to the inner product $\langle \cdot, \cdot \rangle_{\rho_k^q}$, and the infimum is taken over $C^1(\mathcal{M})$. Thus $\Theta_k$ is indeed the first non-trivial eigenvalue

of operator $\Delta_{\rho_k^q}$, which can be proved easily by using the min-max theorem of operators and proposition 2.

**2.2.2. Kernel PCA case.** Let $\nu_1, \nu_2, \cdots, \nu_N$ be probability distributions on $\Omega$ (e.g. $\mathbb{R}^k$). Similar to the above, we consider a mixture model of the form

$$\nu \coloneqq \sum_{k=1}^{N} w_k \nu_k,$$

where $w_1, w_2, \cdots, w_N$ be strictly positive weights which satisfy that $\sum_{k=1}^{N} w_k = 1$.

Recall the definition of $q$ and $q_k$ as kernelized densities given in equation (2.5) and equation(2.6) above. For $i, j = 1, 2, \cdots, N$, let

$$\mathcal{S}_{ij} = \left\langle \left(\frac{q_i}{q}\right), \left(\frac{q_j}{q}\right) \right\rangle_\nu = \int_\Omega \frac{q_i(x)}{q(x)} \frac{q_j(x)}{q(x)} \nu(dx),$$

and

$$\mathcal{S}^*_{ij} = \langle q_i, q_j \rangle_\nu = \int_\Omega q_i(x) q_j(x) \nu(dx).$$

Notice that if we replace $q_i$ and $q$ by $\rho_i$ and $\rho$, respectively, and if measure $\nu$ is a probability measure with density $\rho$, then $\mathcal{S}_{ij}$ in the kernel PCA case has exactly the same form as the $\mathcal{S}_{ij}$ in the weighted Laplacian case with power $q = 1$. These two cases have many commonalities and are also different in some important details illustrated later.

Then similar with weighted Laplacian case, we also define three sets of parameters to quantify the extent of separateness.

**Overlapping parameters (similarity parameters)** of $q_i(x)$

$$\mathcal{S}^*_{\mathrm{w}} \coloneqq \min_{i=1,\cdots,N} \langle q_i, q_i \rangle_\nu = \min_{i=1,\cdots,N} \int_\Omega q_i(x) q_i(x) \nu(dx) \coloneqq \min_{i=1,\cdots,N} \mathcal{S}^*_{ii},$$

$$\mathcal{S}^*_{\mathrm{w,up}} \coloneqq \max_{i=1,\cdots,N} \langle q_i, q_i \rangle_\nu = \max_{i=1,\cdots,N} \int_\Omega q_i(x) q_i(x) \nu(dx) \coloneqq \max_{i=1,\cdots,N} \mathcal{S}^*_{ii},$$

$$\mathcal{S}^*_{\mathrm{b}} \coloneqq \max_{i=1,\cdots,N, i\neq j} \langle q_i, q_j \rangle_\nu = \max_{i=1,\cdots,N, i\neq j} \int_\Omega q_i(x) q_j(x) \nu(dx) \coloneqq \max_{i=1,\cdots,N, i\neq j} \mathcal{S}^*_{ij},$$

where $\mathcal{S}^*_{ij} \coloneqq \int_\Omega q_i(x) q_j(x) \nu(dx)$.

**Overlapping parameter (similarity parameter)** of $\frac{q_i(x)}{q(x)}$

$$\overline{\mathcal{S}_i} \coloneqq \sum_{j=1,\cdots,N,j\neq i} w_j \mathcal{S}_{ij}.$$

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ denote the sorted eigenvalues of the centered covariance operator defined in (2.8) above. The **Eigen-tail parameter** is defined as

$$\Lambda \coloneqq \sum_{l=N+1}^{\infty} \lambda_l^2.$$

**2.2.3. Well-separated mixture model.** Based on the parameters defined in previous part, we have an informal definition of a well-separated mixture model. We say that a mixture model $\rho = \sum_{k=1}^{N} w_k \rho_k$ in the weighted Laplacian case tends to be well-separated if the overlapping parameter $S^*_{\text{between}}$ is small enough, and the ratio of the coupling parameter $\mathcal{C}$ is small enough in comparison to the indivisibility parameter $\Theta$. Similarly, we say a mixture model $\nu = \sum_{k=1}^{N} w_k \nu_k$ in the kernel PCA case tends to be well-separated if the ratio of two overlapping parameters $\frac{S^*_{\text{between}}}{S^*_{\text{within}}}$ is small enough, and the eigen-tail parameter $\Lambda$ is small enough in comparison to the similarity parameter $S^*_{\text{within}}$. These quantities play important role in the parameters in our main theorem, and specific examples will be discussed in the following sections.

Our main theorems in both the population setting and the sample setting are based on a mixture model, where the number of components $N$ is fixed for all future analysis. Among those theorems, we may see that there is a trade-off between the size of the angles $\sigma_i$ and the coverage proportion $\delta$, where smaller angles naturally tend to lead to smaller coverage, i.e. to larger value of $\delta$. So the choice of $\sigma_k, k = 1, 2, \cdots, N$ is arbitrary, and we may choose them according to our purpose. For example, we can keep $\delta = 0.05$ and choose appropriate $\sigma_k$'s to get 95% coverage proportion.

## 2.3. Main results

**2.3.1. OCS of spectral embedding: The population setting.** Let $u_1, \cdots, u_N$ be the $N$ orthonormal (with respect to $\langle \cdot, \cdot \rangle_{L^2(\nu)}$) eigenfunctions of $\Delta_\rho$ corresponding to its $N$ smallest

eigenvalues. Define the population version of the weighted spectral embedding as

$$F : x \in \mathcal{M} \longmapsto \begin{pmatrix} u_1(x) \\ \vdots \\ u_N(x) \end{pmatrix} \in \mathbb{R}^N.$$

Further let $\mu \coloneqq F_\sharp \nu$ be the push-forward of $\nu$ by $F$. Note that $\mu$ is a measure on $\mathbb{R}^N$. This measure is used to describe the distribution of points originally in $\mathcal{M}$ after transformation by the weighted Laplacian map $F$.

In order to prove that the measure $F_\sharp \nu$ has an OCS under the assumption of a well-separated mixture model, the related measure $F^Q_\sharp \nu$ is first explored to obtain our result, where $F^Q$ is the map

$$F^Q : x \in \mathcal{M} \longmapsto \begin{pmatrix} \frac{q_1}{\sqrt{I_1 w_1^q}} \\ \vdots \\ \frac{q_N}{\sqrt{I_N w_1^q}} \end{pmatrix} \in \mathbb{R}^N.$$

We will first show that $F^Q_\sharp \nu$ has an OCS under the assumption of a well-separated mixture model, which requires that the overlapping parameter and coupling parameter are small enough and the indivisibility parameter is large enough.

Assumption 2. *Assume that $\mathcal{M}$ is a smooth, connected, orientable, $m$-dimensional compact Riemannian manifold in $\mathbb{R}^d$. Let $\rho$ be the mixture density $\rho(x) = \sum_{i=1}^N w_i \rho_i(x)$ on $\mathcal{M}$. Then the assumption is that there exists a constant $\alpha \geq 1$ such that*

$$\frac{1}{\alpha} \leq \rho(x) \leq \alpha \text{ for all } x \in \mathcal{M},$$

*and that $\rho$ is $C_\rho$-Lipschitz.*

Theorem 1. *Given a mixture model with density $\rho(x) = \sum_{i=1}^N w_i \rho_i(x)$ satisfying Assumption 1 and Assumption 2, and let $\nu$ denote the probability measure with density $\tilde{\rho}(x) \coloneqq \frac{\rho^q(x)}{\int \rho^q(t)dt}$ depending on the parameter $q$. For $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4)$, define*

$$\delta^* \coloneqq \frac{N I_{max}}{I I_{min}} \left( \frac{w_{\max}}{w_{\min}} \right)^q \sum_{k=1}^N w_k^q \frac{\cos^2(\sigma_k)}{1 - \cos^2(\sigma_k)} \overline{\mathcal{S}_k},$$

*where* $w_{\max} \coloneqq \max_{i=1,\dots,N} w_i$, $w_{\min} \coloneqq \min_{i=1,\dots,N} w_i$. *Suppose that*

$$\mathcal{S}_{adj} < 1,$$

*and*

$$\mathcal{S}_b < I_{min} - \max(N^q, N)\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2. \tag{2.9}$$

*Further suppose that*

$$\sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\overline{\mathcal{S}_k} \le \frac{II_{min}}{NI_{max}}\left(\frac{w_{\min}}{w_{\max}}\right)^q.$$

*Also define*

$$\tau \coloneqq \frac{4\alpha^{\frac{|q-p|}{2}}}{\sqrt{I_{min}}}\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}\mathcal{C}}\left(\frac{1}{\max(N^q,N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min}-\mathcal{S}_b}\right)}\right.$$
$$\left. - \frac{\sqrt{N}}{I_{min}-\mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^{-1} + \frac{\sqrt{I\mathcal{S}_b}}{I_{min}}.$$

*Then suppose that* $\tau - \frac{\sqrt{I\mathcal{S}_b}}{I_{min}} > 0, \quad \tau N < 1$ *and* $s, t > 0$ *satisfy*

$$\frac{t^2\sin^2(s)}{N^q w_{\max}^q} \ge N\left(\frac{\tau - \frac{\sqrt{I\mathcal{S}_b}}{I_{min}}}{2}\right)^2 + 4N^{3/2}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right), \quad \sigma_i + s < \frac{\pi}{4}, i = 1, \cdots, N. \tag{2.10}$$

*Then, the probability measure* $\mu = F_\sharp \nu$ *has an orthogonal cone structure with parameters*

$$\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1-\sin(s)}{\sqrt{\max(N^{q-1},1) w_{\max}^q I_{max}}}\right) \text{ for any } \delta \in [\delta^*, 1).$$

REMARK 3. *Note that if* $0 < q \le 1$, *the expressions in Theorem 1 simplify significantly, making them easier to interpret. For instance, we have*

$$\mathcal{S}_{adj} = \mathcal{S}_b \max_{j=1,2,\cdots,N}\left(\sum_{k\neq j} w_k^q\right),$$

*and thus we have* $\mathcal{S}_{adj} \le (N-1)\mathcal{S}_b$. *So if the individual mixture components are close to being orthogonal (so that* $\mathcal{S}_{ij} \approx 0$ *for* $i \neq j$*), then* $\mathcal{S}_b$ *is close to zero, and so is* $\mathcal{S}_{adj}$. *As a result, the condition* $\mathcal{S}_{adj} < 1$ *and condition (2.13) will hold in this case, because (for* $0 < q \le 1$*) the latter*

*one simplifies to*

$$\mathcal{S}_b < \frac{I_{min}}{1+N}. \tag{2.11}$$

*If we formally set $\mathcal{S}_b$ to zero, then it is straightforward to see that $\tau$ becomes small if the ratio $\frac{\mathcal{C}}{\Theta}$ is small. And a small $\tau$, in turn, means that $s$ and $t$ can be small as well. So, for $0 < q \leq 1$, the OCS is "strong" if $\mathcal{S}_b$ is small and $\frac{\mathcal{C}}{\Theta}$ is small. This is consistent with Garcia-Trillos et al. (2019)( [**89**]).*

**2.3.2. OCS of spectral embedding: The sample setting.** Suppose we have i.i.d. samples $\mathcal{M}_n \coloneqq \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ from the density $\rho$ that is supported on a manifold $\mathcal{M}$ embedded in some Euclidean space $\mathbb{R}^d$, then the empirical measure associated to the samples is denoted as

$$\nu_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i},$$

where $\delta_{\mathbf{x}_i}$ denotes Dirac measure in the points $\mathbf{x}_i$.

We denote by $L^2(\nu_n)$ the space of functions $u : \mathcal{M}_n \to \mathbb{R}$ and identify $u \in L^2(\nu_n)$ with a column vector $(u(\mathbf{x}_1), \cdots, u(\mathbf{x}_n))^T$ in $\mathbb{R}^n$.

Let $\eta : [0, \infty) \to [0, \infty)$ be a non-increasing Lipschitz function with support $[0, 1]$ such that

$$\int_{\mathbb{R}^m} \eta(|x|)dx = 1.$$

For a given bandwidth $\varepsilon > 0$, let

$$\eta_\varepsilon(r) := \frac{1}{\varepsilon^m}\eta\left(\frac{r}{\varepsilon}\right),$$

$$\hat{d}_\varepsilon(y) := \sum_{j=1}^{n} \eta_\varepsilon\left(|y - \mathbf{x}_j|\right) \quad y \in \mathcal{M},$$

and let $\tilde{W}_n$ denote the similarity matrix with entries $(\tilde{W}_n)_{ij} \coloneqq \eta_\varepsilon(|x_i - x_j|)\mathbf{1}(i \neq j)$. Denote $\tilde{d}_i \coloneqq \sum_{j=1}^{n}(\tilde{W}_n)_{ij}$, then re-weighted similarity matrix $W_n$ is defined as

$$(W_n)_{ij} = \frac{(\tilde{W}_n)_{ij}}{\tilde{d}_i^{1-q/2}\tilde{d}_j^{1-q/2}},$$

and the corresponding degree matrix is $D_n = \mathrm{diag}(d_i)$ with $d_i \coloneqq \sum_{j=1}^{n} W_{ij}$.

We identify the Laplacian matrix with the following empirical weighted Laplacian operator $\Delta_n : L^2(\nu_n) \to L^2(\nu_n)$, which is defined through the following matrix for $(p, q) \in \mathbb{R}^2$:

$$\Delta_n \coloneqq \begin{cases} D_N^{\frac{1-p}{q-1}} \left(D_n - W_n\right), & \text{if } q \neq 1, \\ D_n - W_n, & \text{if } q = 1. \end{cases}$$

The spectrum of $\Delta_n$ induces a weighted spectral embedding

$$F_n : \mathbf{x}_i \in \Omega \longmapsto \begin{pmatrix} u_{n,1}(\mathbf{x}_i) \\ \vdots \\ u_{n,N}(\mathbf{x}_i) \end{pmatrix} \in \mathbb{R}^N,$$

where $u_{n,1}, \cdots, u_{n,N}$ are the eigenvectors of $\Delta_n$ associated to the $N$ smallest eigenvalues.

Assumption 3. *Assume that $n$ is large enough and $\varepsilon$ is small enough such that*

$$\varepsilon \left(1 + \sqrt{\lambda_N}\right) + \frac{\log(n)^{p_m}}{n^{1/m}\varepsilon} \leq \min\left\{C, \frac{1}{2}\left(\lambda_{N+1} - \lambda_N\right)\right\},$$

*where $p_m = \frac{3}{4}$ if $m = 2$ and $p_m = \frac{1}{m}$ if $m \geq 3$, $C > 0$ is a finite constant that depends on $\mathcal{M}$.*

Theorem 2. *Let $N \geq 2$ and $\beta > 1$. Suppose that $\mathcal{M}$ and $\rho$ satisfy Assumptions 1,2 and 3 for some $\varepsilon > 0$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. samples from the measure $\nu$ with associated empirical measure $\nu_n$, and let $F_n$ be the Laplacian embedding defined as*

$$F_n : \mathbf{x}_i \longmapsto \begin{pmatrix} u_{n,1}\left(\mathbf{x}_i\right) \\ \vdots \\ u_{n,N}\left(\mathbf{x}_i\right) \end{pmatrix}.$$

*Then there exists a constant $C_\beta > 0$ depending only on $\beta$ such that with probability at least $1 - C_\beta n^{-\beta}$, the probability measure $\mu_n \coloneqq F_{n\sharp}\nu_n$ has an orthogonal cone structure with parameters*

$$\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1 - \sin(s)}{\sqrt{\max(N^{q-1}, 1) w_{\max}^q I_{max}}}\right)$$

*for any $\delta \in [\delta^*, 1)$ (where $\delta^*$ is defined in Theorem 1) and $s, t > 0$ satisfying*

$$\sigma_i + s < \frac{\pi}{4}, i = 1, \cdots, N,$$

*and*

$$\frac{t\sin(s)}{\sqrt{N^q w_{\max}^q}} \geq \sqrt{N\left(\frac{\tau - \frac{\sqrt{I\mathcal{S}_b}}{I_{min}}}{2}\right)^2 + 4N^{3/2}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right)} + \sqrt{N\phi}, \tag{2.12}$$

*where*

$$\phi = \phi(\mathcal{S}_b, \mathcal{C}, \Theta, I_{min}, \mathcal{I}^*, N, \varepsilon, n, m) \coloneqq c_{\mathcal{M}}\left(\left(\frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\left(\varepsilon + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}}\right)\psi^{-1} + \left(\frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\varepsilon^{m+2}\left(\varepsilon + \varepsilon^2 + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} + \frac{\log(n)^{p_m}}{n^{1/m}}\right)\right),$$

*and*

$$\psi = \psi(\mathcal{S}_b, \mathcal{S}_{adj}, \mathcal{C}, \Theta, I_{min}, \mathcal{I}^*, N, \varepsilon, n, m) \coloneqq \left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{\max(N^q, N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min} - \mathcal{S}_b}\right)} - \frac{\sqrt{N\mathcal{C}}}{I_{min} - \mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^2 - \frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}},$$

*and $c_{\mathcal{M}}$ is a constant depending on $N$, $\beta$, manifold $\mathcal{M}$, density bound $\alpha$ (defined in Assumption 2), Lipschitz function $\eta$ and Lipschitz constant $C_\rho$. Here we require $\mathcal{S}_b$ and $\frac{\mathcal{C}}{\Theta}$ small enough such that all the terms above are well-defined.*

REMARK 4. *Similar with Theorem 1, when $0 < q \leq 1$, we have a simplified version of parameter $\psi$ as*

$$\psi = \psi(\mathcal{S}_b, \mathcal{S}_{adj}, \mathcal{C}, \Theta, I_{min}, \mathcal{I}^*, N, \varepsilon, n, m) \coloneqq \left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{N} - \frac{\mathcal{S}_b}{I_{min} - \mathcal{S}_b}\right)} - \frac{\sqrt{N\mathcal{C}\mathcal{S}_b}}{I_{min} - \mathcal{S}_b}\right)^2 - \frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}.$$

REMARK 5. *If we further assume that the manifold $\mathcal{M}$ satisfies a specific property, then the constant $c_{\mathcal{M}}$ in Theorem 2 can be written in a more detailed form, which will be discussed in the proof part. Notice that the parameters $\phi$ and $\psi$ rely on many parameters. Among these parameters, $\mathcal{S}_b$, $\mathcal{C}$ and $\Theta$ are most useful ones that can quantify how well a mixture model is separated, given $N$ and $q$ fixed. If $\mathcal{S}_b$ is small enough, then $\phi$ is also small, which leads to smaller value on the right hand side of the inequality (2.15). Then the choice of $t$ and $s$ can also be as small as possible, which leads to more concentrated OCS. Similarly, smaller ratio $\frac{\mathcal{C}}{\Theta}$ can also leads to more concentrated OCS. More properties of these parameters are discussed in the next chapter. Notice that in contrast to*

*the population case, not only the ratio $\frac{\mathcal{C}}{\Theta}$ matters in the sample case, but the sample size $n$ also needs to be large enough to guarantee a small $\phi$.*

**2.3.3. OCS of kernel PCA embedding: The population setting.** Let $u_1, \cdots, u_N$ be orthonormal (with respect to $\langle \cdot, \cdot \rangle_\nu$) eigenfunctions of $\Sigma_\nu$ corresponding to its $N$ largest eigenvalues. Define the Kernel PCA embedding

$$F : x \in \Omega \longmapsto \begin{pmatrix} u_1(x) \\ \vdots \\ u_N(x) \end{pmatrix} \in \mathbb{R}^N.$$

Further let $\mu \coloneqq F_\sharp \nu$ be the push-forward of the measure $\nu$ by $F$. In this thesis, we show the orthogonal cone structure of a well-seperated mixture model. The concept of OCS appears from point clouds and is generalized to arbitrary probability distributions. Another related measure $F^Q_\sharp \nu$ is first explored to obtain our result, where $F^Q$ is the map

$$F^Q : x \in \Omega \longmapsto \begin{pmatrix} \frac{q_1}{\|q_1\|_\nu} \\ \vdots \\ \frac{q_N}{\|q_N\|_\nu} \end{pmatrix} \in \mathbb{R}^N.$$

THEOREM 3. *Let $\nu$ be a mixture model $\nu(x) = \sum_{i=1}^N w_i \nu_i(x)$. For $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4)$, define*

$$\delta^* \coloneqq \frac{N w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{w,up}}{\mathcal{S}^*_w} \sum_{k=1}^N w_k \frac{\cos(\sigma_k)}{\sqrt{1 - \cos^2(\sigma_k)}} \overline{\mathcal{S}_k},$$

*where $w_{\max} \coloneqq \max_{i=1,\ldots,N} w_i$, $w_{\min} \coloneqq \min_{i=1,\ldots,N} w_i$.*

*Suppose that*

$$\sum_{k=1}^N w_k \frac{\cos(\sigma_k)}{\sqrt{1 - \cos^2(\sigma_k)}} \overline{\mathcal{S}_k} \leq \frac{w_{\min}}{N w_{\max}}.$$

*Also define*

$$\tau \coloneqq 4\sqrt{\frac{\Lambda}{\mathcal{S}^*_w w_{\min}}} + \frac{\mathcal{S}^*_b}{\mathcal{S}^*_w}.$$

*Then suppose that* $\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*} > 0, \quad \tau N < 1$ *and* $s, t > 0$ *satisfy*

$$\frac{t^2 \sin^2(s)}{N^2 w_{\max}^2} \geq N \left( \frac{\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*}}{2} \right)^2 + 4N^{3/2} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right), \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N.$$

*Then, the probability measure* $\mu = F_\sharp \nu$ *has an orthogonal cone structure with parameters*

$\left( \sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1 - \sin(s)}{\sqrt{N} w_{\max}} \right)$ *for any* $\delta \in [\delta^*, 1)$.

**2.3.4. OCS of kernel PCA embedding: The sample setting.** Let $X \coloneqq \{x_1, \cdots, x_n\}$ be i.i.d. samples from the probability measure $\nu$, and as above let $\nu_n$ denote the empirical measure based on $X$.

Recall that $\bar{k}$ denotes the centered kernel introduced above (see (2.7)). With this, we define the empirical covariance operator as

$$\Sigma_{\nu_n} f(\cdot) = \int_\Omega \bar{k}(\cdot, y) f(y) \nu_n(dy) = \frac{1}{n} \sum_{i=1}^n \bar{k}(\cdot, x_i) f(x_i), \quad f \in \mathcal{H}(\bar{k}).$$

This operator is closely related to the $n \times n$ kernel matrix $K_n$, where

$$(K_n)_{ij} = \frac{\bar{k}(x_i, x_j)}{n}.$$

Also define the empirical Kernel PCA embedding as

$$F_n : x_i \longmapsto \begin{pmatrix} u_{n,1}(x_i) \\ \vdots \\ u_{n,N}(x_i) \end{pmatrix} \in \mathbb{R}^N,$$

where $u_{n,1}, \cdots, u_{n,N}$ are the eigenvectors of $K_n$ associated to the $N$ largest eigenvalues, and the adjusted Kernel PCA embedding

$$\tilde{F}_n : x_i \longmapsto \begin{pmatrix} sign(\langle u_1, u_{n,1} \rangle_{\mathcal{H}}) u_{n,1}(x_i) \\ \vdots \\ sign(\langle u_N, u_{n,N} \rangle_{\mathcal{H}}) u_{n,N}(x_i) \end{pmatrix} \in \mathbb{R}^N,$$

where $sign(x) = \mathbf{1}_{(x\geq 0)} - \mathbf{1}_{(x<0)}$. Notice that $F_n = \tilde{O}\tilde{F}_n$, where $\tilde{O}$ defines an orthogonal transformation as

$$\tilde{O} = \mathrm{diag}(sign(\langle u_1, u_{n,1}\rangle_{\mathcal{H}}), sign(\langle u_2, u_{n,2}\rangle_{\mathcal{H}}), \cdots, sign(\langle u_N, u_{n,N}\rangle_{\mathcal{H}})).$$

Our next theorem is about the OCS in the sample setting of kernel PCA embedding, which needs the assumptions of sub-Gaussian and pre-Gaussian. The definitions of these two conceptes are given below.

Definition 2. *A centered random element $\Phi(x) := k(\cdot, x)$, with $x \sim \nu$, in $\mathcal{H}$ is called sub-Gaussian if for all $f \in \mathcal{H}$,*

$$\|\langle \Phi(x), f\rangle_{\mathcal{H}}\|_{\psi_2} \lesssim \|\langle \Phi(x), f\rangle_{\mathcal{H}}\|_{\nu},$$

*where*

$$\|\eta\|_{\psi_2} := \inf\left\{C > 0 : \mathbb{E}\psi_2\left(\frac{|\eta|}{C}\right) \leq 1\right\} \text{ with } \psi_2(x) \coloneqq e^{u^2} - 1, u \geq 0.$$

Definition 3. *A weakly square integrable centered random element $\Phi(x) := k(\cdot, x)$, with $x \sim \nu$, in $\mathcal{H}$ with covariance operator $\Sigma_\nu$ is called pre-Gaussian if there exists a centered Gaussian random element in $\mathcal{H}$ with the same covariance operator $\Sigma_\nu$.*

Theorem 4. *Assume that all assumptions in Theorem 3 hold, and that the random element $\Phi(x) := k(\cdot, x)$, with $x \sim \nu$ is sub-Gaussian and pre-Gaussian. Let $x_1, \cdots, x_n$ be i.i.d. samples from the measure $\nu$ with associated empirical measure $\nu_n$, and let $F_n$ be the empirical Kernel PCA embedding. Then, there exists a numerical constant $C$ such that with probability at least $1 - e^{-\beta}$, the probability measure $\mu_n \coloneqq F_{n\sharp}\nu_n$ has an orthogonal cone structure with parameters $\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1-\sin(s)}{\sqrt{N}w_{\max}}\right)$, for any $\delta \in [\delta^*, 1)$ (where $\delta^*$ is defined in Theorem 3) and $s, t > 0$ satisfying*

$$\frac{t\sin(s)}{Nw_{\max}} \geq \sqrt{\frac{128C^2N\|\Sigma_\nu\|_\infty^2 \left(r^*(\Sigma_\nu)\right)^2}{(\bar{g}_{\min})^2} + 8M^2\sum_{j=1}^{N}\frac{1}{\left(\lambda_j - Cr^*(\Sigma_\nu)\right)^2}}$$
$$+ \sqrt{N\left(\frac{\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*}}{2}\right)^2 + 4N^{\frac{3}{2}}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right)}, \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N,$$

*where* $\|\cdot\|_\infty$ *is the operator norm,* $r^*(\Sigma_\nu) \coloneqq \left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)$, $\bar{g}_{\min} = \min_{i=1,\dots,N}(\lambda_i - \lambda_{i+1})$ *is the minimum spectral gap, and* $r(\Sigma_\nu) \coloneqq \frac{\mathrm{tr}(\Sigma_\nu)}{\|\Sigma_\nu\|_\infty}$ *is the effective rank of the covariance operator* $\Sigma_\nu$*. Here we require that* $n$ *is large enough such that* $Cr^*(\Sigma_\nu) < \lambda_j$ *and thus all the inner terms are positive.*

REMARK 6. *All the four main theorems can be generalized to the strong version of OCS, i.e. each cone covers one specific component as stated in Remark 1. The stronger version of the four main theorems are given below. Detailed proofs are not given since they only require a small modification in proposition 6 and proposition 11 and their proofs, where we need to bound the measure of each cone respectively rather than only consider their union. This will lead to a loss of the tightness for the bound of the coverage ratio, thus both two cases have their advantages and disadvantages.*

THEOREM 5. *Given a mixture model with density* $\rho(x) = \sum_{i=1}^N w_i \rho_i(x)$ *satisfying Assumption 1 and Assumption 2, and let* $\nu$ *denote the probability measure with density* $\tilde{\rho}(x) \coloneqq \frac{\rho^q(x)}{\int \rho^q(t)dt}$ *depending on the parameter* $q$*. For* $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4)$*, define*

$$\delta^* = \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \frac{I_{max} N^{2q+1}}{I_{min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{l=1}^N w_l^q \overline{\mathcal{S}_l},$$

*where* $w_{\max} \coloneqq \max_{i=1,\dots,N} w_i$*,* $w_{\min} \coloneqq \min_{i=1,\dots,N} w_i$*. Suppose that*

$$\mathcal{S}_{adj} < 1,$$

*and*

$$\mathcal{S}_b < I_{min} - \max(N^q, N)\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2. \tag{2.13}$$

*Further suppose that*

$$\sum_{k=1}^N w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \overline{\mathcal{S}_k} \le \frac{II_{min}}{NI_{max}} \left(\frac{w_{\min}}{w_{\max}}\right)^q.$$

*Also define*

$$\tau := \frac{4\alpha^{\frac{|q-p|}{2}}}{\sqrt{I_{min}}}\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}\mathcal{C}}\left(\frac{1}{\max(N^q,N)}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min}-\mathcal{S}_b}\right)}\right.$$
$$\left.-\frac{\sqrt{N}}{I_{min}-\mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^{-1}+\frac{\sqrt{I\mathcal{S}_b}}{I_{min}}.$$

*Then suppose that* $\tau-\frac{\sqrt{I\mathcal{S}_b}}{I_{min}}>0, \quad \tau N<1$ *and* $s,t>0$ *satisfy*

$$\frac{t^2\sin^2(s)}{N^q w_{\max}^q}\geq N\left(\frac{\tau-\frac{\sqrt{I\mathcal{S}_b}}{I_{min}}}{2}\right)^2+4N^{3/2}\left(\frac{1}{\sqrt{1-N\tau}}-1\right), \quad \sigma_i+s<\frac{\pi}{4}, i=1,\cdots,N. \tag{2.14}$$

*Then, the probability measure* $\mu=F_\sharp\nu$ *has a strong orthogonal cone structure with parameters*

$$\left(\sigma_1+s,\sigma_2+s,\cdots,\sigma_N+s,\delta+t^2,\frac{1-\sin(s)}{\sqrt{\max(N^{q-1},1)w_{\max}^q I_{max}}}\right) \text{ for any } \delta\in[\delta^*,1).$$

THEOREM 6. *Let* $N\geq 2$ *and* $\beta>1$. *Suppose that* $\mathcal{M}$ *and* $\rho$ *satisfy the Assumption 2 and Assumption 3 for some* $\varepsilon>0$, *and all assumptions in Theorem 5 are also satisfied. Let* $\mathbf{x}_1,\ldots,\mathbf{x}_n$ *be i.i.d. samples from the measure* $\nu$ *with associated empirical measure* $\nu_n$, *and let* $F_n$ *be the Laplacian embedding defined as*

$$F_n:\mathbf{x}_i\longmapsto\begin{pmatrix}u_{n,1}(\mathbf{x}_i)\\ \vdots\\ u_{n,N}(\mathbf{x}_i)\end{pmatrix}.$$

*Then there exists a constant* $C_\beta>0$ *depending only on* $\beta$ *such that with probability at least* $1-C_\beta n^{-\beta}$, *the probability measure* $\mu_n:=F_{n\sharp}\nu_n$ *has a strong orthogonal cone structure with parameters*

$$\left(\sigma_1+s,\sigma_2+s,\cdots,\sigma_N+s,\delta+t^2,\frac{1-\sin(s)}{\sqrt{\max(N^{q-1},1)w_{\max}^q I_{max}}}\right)$$

*for any* $\delta\in[\delta^*,1)$ *(where* $\delta^*$ *is defined in Theorem 5) and* $s,t>0$ *satisfying*

$$\sigma_i+s<\frac{\pi}{4}, i=1,\cdots,N,$$

*and*

$$\frac{t\sin(s)}{\sqrt{N^q w_{\max}^q}} \geq \sqrt{N\left(\frac{\tau - \frac{\sqrt{I\mathcal{S}_b}}{I_{min}}}{2}\right)^2 + 4N^{3/2}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right)} + \sqrt{N\phi}, \tag{2.15}$$

*where*

$$\phi = \phi(\mathcal{S}_b, \mathcal{C}, \Theta, I_{min}, \mathcal{I}^*, N, \varepsilon, n, m) \coloneqq c_{\mathcal{M}}\left(\left(\frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\left(\varepsilon + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}}\right)\psi^{-1} + \left(\frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\varepsilon^{m+2}\left(\varepsilon + \varepsilon^2 + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} + \frac{\log(n)^{p_m}}{n^{1/m}}\right)\right),$$

*and*

$$\psi = \psi(\mathcal{S}_b, \mathcal{S}_{adj}, \mathcal{C}, \Theta, I_{min}, \mathcal{I}^*, N, \varepsilon, n, m) \coloneqq \left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{\max(N^q, N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min} - \mathcal{S}_b}\right)} - \frac{\sqrt{N\mathcal{C}}}{I_{min} - \mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^2 - \frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}},$$

*and $c_{\mathcal{M}}$ is a constant depending on $\mathcal{M}, \alpha, C_\rho, \eta, \beta$ and $N$. Here we require $\mathcal{S}_b$ and $\frac{\mathcal{C}}{\Theta}$ small enough such that all the terms above are well-defined.*

THEOREM 7. *Let $\nu$ be a mixture model $\nu(x) = \sum_{i=1}^N w_i\nu_i(x)$. For $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4)$, define*

$$\delta^* \coloneqq \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}}\frac{N^3 w_{\max}}{w_{\min}}\frac{\mathcal{S}_{w,up}^*}{\mathcal{S}_w^*}\sum_l w_l\overline{\mathcal{S}_l},$$

*where $w_{\max} \coloneqq \max_{i=1,\ldots,N} w_i$, $w_{\min} \coloneqq \min_{i=1,\ldots,N} w_i$.*

*Suppose that*

$$\sum_{k=1}^N w_k\frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}}\overline{\mathcal{S}_k} \leq \frac{w_{\min}}{Nw_{\max}}.$$

*Also define*

$$\tau \coloneqq 4\sqrt{\frac{\Lambda}{\mathcal{S}_w^* w_{\min}}} + \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*}.$$

*Then suppose that $\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*} > 0, \quad \tau N < 1$ and $s, t > 0$ satisfy*

$$\frac{t^2\sin^2(s)}{N^2 w_{\max}^2} \geq N\left(\frac{\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*}}{2}\right)^2 + 4N^{3/2}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right), \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N.$$

*Then, the probability measure $\mu = F_\sharp \nu$ has a strong orthogonal cone structure with parameters* $\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1-\sin(s)}{\sqrt{N} w_{\max}}\right)$ *for any* $\delta \in [\delta^*, 1)$.

THEOREM 8. *Assume that all the assumptions in Theorem 3 hold, and that the random element* $\Phi(x) := k(\cdot, x)$ *with* $x \sim \nu$, *and all the random elements* $\Phi(x_i) := k(\cdot, x_i)$, *with* $x_i \sim \nu_i$ *are sub-Gaussian and pre-Gaussian for all* $i = 1, \ldots, N$. *Let* $x_1, \cdots, x_n$ *be i.i.d. samples from* $\nu$ *with associated empirical measure* $\nu_n$, *and let* $F_n$ *be the empirical Kernel PCA embedding. Then, there exists a numerical constant* $C$ *such that with probability at least* $1 - e^{-\beta}$, *the probability measure* $\mu_n := F_{n\sharp}\nu_n$ *has a strong orthogonal cone structure with parameters* $\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1-\sin(s)}{\sqrt{N} w_{\max}}\right)$, *for any* $\delta \in [\delta^*, 1)$ *(where* $\delta^*$ *is defined in Theorem 7) and* $s, t > 0$ *satisfying*

$$\frac{t \sin(s)}{N w_{\max}} \geq \sqrt{\frac{128 C^2 N \|\Sigma_\nu\|_\infty^2 \left(r^*(\Sigma_\nu)\right)^2}{(\bar{g}_{\min})^2} + 8M^2 \sum_{j=1}^{N} \frac{1}{\left(\lambda_j - C r^*(\Sigma_\nu)\right)^2}}$$
$$+ \sqrt{N \left(\frac{\tau - \frac{\mathcal{S}_b^*}{\mathcal{S}_w^*}}{2}\right)^2 + 4N^{\frac{3}{2}} \left(\frac{1}{\sqrt{1 - N\tau}} - 1\right)}, \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N,$$

*where* $\|\cdot\|_\infty$ *is the operator norm,* $r^*(\Sigma_\nu) := \left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)$, $\bar{g}_{\min} = \min_{i=1,\ldots,N} (\lambda_i - \lambda_{i+1})$, *and* $r(\Sigma_\nu) := \frac{\mathrm{tr}(\Sigma_\nu)}{\|\Sigma_\nu\|_\infty}$ *is the effective rank of the covariance operator* $\Sigma_\nu$. *Here we require that* $n$ *is large enough such that* $C r^*(\Sigma_\nu) < \lambda_j$ *and thus all the inner terms are positive.*

CHAPTER 3

# Discussion

Comparing the obtained bounds for the OCS of spectral embeddings in the sample settings and the population settings, one of the main differences is the additive term $\sqrt{N\phi}$ in Theorem 2 (see inequality (2.15)), which quantifies the closeness of the true measure and the empirical measure. When the sample size increases, this term tends to be small. (Notice that this is not a rigorous monotonic relationship and is just a general trend.) For a suitably chosen bandwidth $\varepsilon$, when $n$ tends to infinity, this term also tends to zero and the theorem in the sample setting degenerates to the one in the population setting. This is consistent with large sample behavior as the samples with infinite sample size should approximate the population well enough.

When $p = q = 1$, our results degenerate to the case that are mentioned in [**89**]. For general choice of $p$ and $q$, the properties of the mixture model encoded in the parameters and the properties of corresponding OCS are explained in detail below.

## 3.1. Weighted overlapping, coupling and indivisibility parameters

When the component $\rho_i$ and $\rho_j$ have well-separated mass, $\mathcal{S}_{ij}$ should be small. Intuitively, for most points $x \in \mathcal{M}$, the two components $\rho_i(x)$ and $\rho_j(x)$ cannot be large simultaneously. Thus if the **weighted overlapping parameter (similarity parameter)** of $\left(\frac{\rho_i(x)}{\rho(x)}\right)^q$ is small, all the pairwise similarities $\mathcal{S}_{ij}$ are small for those components $j$ having non-negligible weights. In this sense, at $x \in \mathcal{M}$, at most one non-negligible component $\rho_k(x)$ dominates the density.

The **coupling parameter** $\mathcal{C}$ is required to be small since we need this metastability condition on the relative entropy of measures $\rho_k^q dx$ with respect to the measure $\rho^q dx$. To understand this, we first define the relative entropy of a probability measure $\varrho^q dx$ with respect to $\rho^q dx$ by

$$H(\varrho^q \mid \rho^q) := \int_{\mathcal{M}} \left(\frac{\varrho^q}{\rho^q}\right) \log\left(\frac{\varrho^q}{\rho^q}\right) \rho^q dx = q \int_{\mathcal{M}} \varrho^q (\log \varrho - \log \rho) dx.$$

Given one probability density $\rho$, define $\varrho(t,x)$ as a function of $t$ and $x$ that satisfies the following evolution equation

$$\partial_t \varrho^q = q\varrho^{q-1}\Delta\varrho - q\,\mathrm{div}(\varrho^q \nabla \log \rho) = q\,\mathrm{div}(\varrho^q \nabla(\log\varrho - \log\rho))$$

with initial condition $\varrho(0,x) = \rho_k(x)$ for a fixed $k \in \{1,\cdots,N\}$. By using Boltzmann's H-theorem ( [**19**]), the relative Fisher Information $I(\varrho^q \mid \rho^q)$ is defined as the entropy dissipation along the solutions of above equation as follows

$$\frac{d}{dt}H(\varrho^q(t) \mid \rho^q) = \int_{\mathcal{M}} \partial_t \varrho^q(\log\varrho - \log\rho)dx = -q\int_{\mathcal{M}} \varrho^q |\nabla(\log\varrho - \log\rho)|^2 dx =: -I(\varrho^q(t) \mid \rho^q).$$

Thus we can rewrite the coupling parameter as

$$\mathcal{C}_k = \frac{q^2}{4} I(\varrho^q(0) \mid \rho^q) = \frac{q^2}{4}\int_{\mathcal{M}} \left| \frac{\nabla\rho_k}{\rho_k} - \frac{\nabla\rho}{\rho} \right|^2 \rho_k^q dx$$

with the approximated entropy for small initial time $t > 0$,

$$H(\varrho^q(t) \mid \rho^q) = H(\rho_k \mid \rho) - \frac{4\mathcal{C}_k}{q^2}t + O\left(t^2\right).$$

From the above equation, we can see that when the coupling parameter $\mathcal{C}_k$ is small, $H(\varrho^q(t) \mid \rho^q)$ varies slowly in the neighborhood of $t = 0$ and thus $H(\varrho^q(t) \mid \rho^q)$ is in a metastable state. Suppose we have a well-separated mixture model, then the weighted overlapping parameter $\overline{\mathcal{S}_i}(i = 1,\cdots,N)$ is small and the initial entropy cannot to be quite small, and we also require $\mathcal{C}$ to be small such that the Fisher Information is small initially when starting the evolution process at each component $\rho_k$.

The **indivisibility parameter** $\Theta$ is used to quantify whether the mixture model gives rise to 'reasonable' clusters or not. If $\Theta$ is small, at least one component $\Theta_k$ is small and the corresponding $\rho_k$ is self-separable in the sense that it can be decomposed into at least two components that have small overlap. This is indicated by the Courant-Fisher max-min theorem. So we require $\Theta$ to be large enough such that the second eigenvalue is bounded away from zero, and thus the Fiedler vector contains useful information about two-way clustering.

In conclusion, for a well-separated model, we require the weighted overlapping parameter $\overline{\mathcal{S}_i}(i = 1,\cdots,N)$ and the coupling parameter $\mathcal{C}$ to be small enough and the indivisibility parameter $\Theta$ to

be large enough. Roughly speaking, the stronger the above statements hold, the easier to observe the Orthogonal Cone Structure.

### 3.2. Overlapping parameters and eigen-tail paramater

Noticed that if the kernel is well-chosen, the kernelized density $q(\cdot)$ has a support similar to the original measure $\nu(\cdot)$. The **overlapping parameters** $\overline{\mathcal{S}_i}$ of $\frac{q_i(x)}{q(x)}$ are small if the components $\nu_i(dx)$ are well-separated. More specifically, a small overlapping parameter means that for most points $x \in \Omega$, at most one of the components $\nu_i$ has large value $\nu_i(dx)$ and other components have small values. Thus $\nu(dx) \approx w_i\nu_i(dx)$ and $q(x) \approx w_i q_i(x)$. This also means that $\mathcal{S}^*_{\text{pair}}$ tends to be smaller while $\mathcal{S}^*_{\text{self}}$ tends to be larger since the former considers integrals over the product of two well separated parts and the latter considers integrals over the square of a single component.

The **eigen-tail parameter** $\Lambda$ is also of vital importance. If we assume the eigenvalues to have exponential decay rate, i.e. $\lambda_l \leq \lambda_1 e^{-r(l-1)}$ for some $r > 0$, then

$$\Lambda = \sum_{l=N+1}^{\infty} \lambda_l^2 \leq \frac{e^{-2rN}}{1 - e^{-2r}} \lambda_1^2.$$

If we assume the eigenvalues to have polynomial decay rate, i.e. $\lambda_l \leq \lambda_1 l^{-r}$ for some $r > 0$, then

$$\Lambda = \sum_{l=N+1}^{\infty} \lambda_l^2 \leq \lambda_1^2 \sum_{l=N+1}^{\infty} l^{-r}.$$

So a fast decay of eigenvalues, implies a small value of $\Lambda$, and $\tau$ in the major theorem then also tends to be small. Thus we may have a 'better' OCS in the sense that the cones could have smaller angle and more coverage proportion.

### 3.3. Angles, coverage and radius

There are several important parameters in the definition of the orthogonal cone structure including $\delta$, $r$, and $\sigma_j$'s for $j = 1, 2, \cdots, N$. The parameters $\sigma_1, \sigma_2, \cdots, \sigma_N$ denotes the opening angles of every spherically symmetric cone with direction determined by the orthogonal basis $e_1, e_2, \cdots, e_N$. The proportion not covered by these cones is $\delta$, and $r$ denotes the radius of the hyper-ball centered at the origin that need to be removed in our geometric structure.

Based on our main theorems, we can choose any set of angles $\sigma_1, \sigma_2, \cdots, \sigma_N$ and obtain the corresponding coverage proportions $\delta$ of these cones. Larger angles tends to have larger coverage proportions, meaning that they result in a smaller value of $\delta$. For a fixed measure $\mu$, our definition of OCS works for many pairs of angles and proportions as long as their relationships satisfy the condition (2.14) given in the theorem. We say that a pair $(t, s)$ is a 'good' choice if the equal sign in (2.14) is achieved.

There is also a trade-off between $t$ and $s$ in the major theorem. In the proof section, we first construct an ancillary measure that has an OCS with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ for some $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \frac{\pi}{4})$, $\delta^* \leq \delta < 1$ and $r = \frac{1}{\sqrt{\max(N^{q-1},1)w_{\max}^q}}$. We then use the fact that if the Wasserstein distance between our target measure and the ancillary measure can be controlled, then this together with the OCS of the ancillary measure allows the quantification of the OCS of the target measure. The OCS of the target measure will be weaker, and how much weaker it is will depend in the distance between the two measures. Indeed, there is a trade-off between the angles and the coverage proportion. Moreover, for a perfectly separated mixture model, the similarity parameters and coupling parameters in our model are zero and thus the values of $t$ and $s$ (that are being used to modify the OCS parameters of the target measures as compared to the ancillary measure) can just be zero. In such case, the parameters of our desired probability measure got their optimal value in the sense that these cones have smallest angles and largest coverage proportion.

These $N$ angles are not required to be the same, but of course they could be chosen as such. When we use the same value for all the cones and consider the case $p = q = 1$, our theorem degenerates to the results in [**89**]. Allowing those angles to take different values are useful especially when we consider some manifolds with measures that have anisotropy, which means some directions may weight more than others.

The hyper-ball with radius $r$ centered at origin is removed to construct our geometric structure. Consider the case if one underestimate the true number of clusters, e.g. only choose the first $M$ dimensions while the true number of cluster is $N$ ($M < N$), then only a $M$-dimensional subspace of the embedded data is considered and the missing $N - M$ dimensions are projected onto this subspace. If the OCS in the theoretical $N$-dimensional case is 'good' enough in the sense that the cones are concentrated with large coverage, and $N$ clusters are mostly lying around the $N$ axes,

respectively, then most of the points from the last $N - M$ clusters are concentrated around the origin in the $M$-dimensional subspace. After removing the hyper-ball, one can still observe 'good' OCS even when the number of clusters are mis-specified.

### 3.4. Performance of $k$-means clustering under OCS

The $K$-means algorithm is one of the most popular clustering algorithms. Here, we consider the $K$-means algorithm applied to unit vectors obtained by normalizing the embeddings. In practice, if an embedded data set has a good OCS in the sense that the corresponding angle parameters are small enough with large coverage proportion, then $k$-means algorithm with uniformly random orthonormal vectors as random initialization works well for clustering. In order to describe the performance quantitatively, we need to first say what we understand by the OCS in the sample setting. This is nothing but the OCS based on the empirical measure. More specifically:

DEFINITION 4. *(Orthogonal Cone Structure of a finite set) Given an data set $x_1, x_2, \cdots, x_n \in \mathbb{R}^N$, and parameters $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \pi/4), \delta \in [0, 1)$, and $r > 0$. The data set has an orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ if there is an orthogonal basis $\{e_1, \cdots, e_N\}$ of $\mathbb{R}^N$ such that*

$$\left| \bigcup_{j=1,2,\cdots,N} \left\{ i \in [n] \middle| \frac{x_i \cdot e_j}{|x_i|} > \cos(\sigma_j), \quad |x_i| > r \right\} \right| \geq (1 - \delta)n.$$

*Here, $|A|$ denotes the cardinality of the set $A \subset \mathbb{R}^d$.*

In words, this definition states that the union of the $N$ spherically symmetric cones with axis of symmetries given by the $e_j, j = 1, \cdots, N$ (minus a ball at the origin of radius $r$) covers at least a portion of $(1 - \delta) \times 100\%$ of the data.

The following is a stronger notion of OCS of an embedding, which has subsequent classification in mind:

REMARK 7. *Suppose that the data $x_1, x_2, \cdots, x_n$ have latent lables $z_1, z_2, \cdots, z_n$ attached to them. We say the embedded (latently labeled) data set $\{x_1, z_1\}, \{x_2, z_2\}, \cdots, \{x_n, z_n\}$ has an orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ if there is an orthogonal basis $\{e_1, \cdots, e_N\}$ of*

$\mathbb{R}^N$ *such that*

$$\sum_{j=1}^{N} \left| \left\{ i \in [n] \middle| \frac{x_i \cdot e_j}{|x_i|} > \cos(\sigma_j), \quad |x_i| > r, \quad z_i = j \right\} \right| \geq (1-\delta)n$$

*holds for all* $j = 1, 2, \cdots, N$.

The following proposition gives a relation between the OCS of an embedding $x_1, \cdots, x_n$ and the performance of a subsequently $k$-means algorithm applied on a sphere based on the unit vectors $\frac{x_i}{|x_i|}, i = 1, 2, \cdots, n$. This algorithm is only applied on selected embedded points that falls outside the ball centered at origin with radius $r$. To make notations simpler, we still use $n$ to denote the total number of embedded points that satisfying this condition in the following proposition.

PROPOSITION 1. *Suppose that the embedded data* $x_1, x_2, \cdots, x_n$ *with latent labels* $z_1, z_2, \cdots, z_n$ *has an orthogonal cone structure with parameters* $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$. *Then if* $\sigma_1, \sigma_2, \cdots, \sigma_N$ *and* $\delta$ *are small enough such that*

$$\frac{(1-\delta)|\{i \in [n] | z_i = j\}| \cos\sigma_j - \delta n}{|\{i \in [n] | z_i = j\}| + \delta n} \geq \frac{1}{2}$$

*and*

$$\frac{\delta n + (1-\delta)|\{i \in [n] | z_i = j\}| \sin\sigma_j}{(1-\delta)|\{i \in [n] | z_i = j\}|} \leq \sin\frac{\pi}{8}$$

*hold for all* $j = 1, 2, \cdots, N$, *then there exists a constant* $c_N$ *such that with probability at least* $1 - \frac{2c_N(\sum_{i=1}^{N}\sigma_i)}{\pi}$ *over the random initialization* $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N$, *where* $\mathbf{a}_j$*'s are uniformly random orthonormal vectors for* $j = 1, 2, \cdots, N$, *the* $k$*-means algorithm on a sphere based on the unit vectors* $\frac{x_i}{|x_i|}, i = 1, 2, \cdots, n$ *clusters at least* $(1-\delta)$ *proportion of the data points correctly. For example,* $c_N = 1$ *for* $N = 2$.

**Proof of Proposition 1**. First consider the case with $N = 2$. From the definition, we know that there exists orthonormal vectors $e_1, e_2$ such that a fraction $1 - \delta$ of the embedded labeled sample lie within an angle $\sigma_i$ of $e_i$ for $i = 1$ or 2. In the $k$-means algorithm, we have the initialization of the mean vector $\mathbf{a}_1$ and $\mathbf{a}_2$, and they are updated after each step of clustering. We consider a non-symmetric cone centered at the origin and the angular bisector of $e_1$ and $e_2$, with an angle $\sigma_1 + \sigma_2$, among which $\sigma_1$ angle is close to $e_1$ and $\sigma_2$ angle is close to $e_2$. A random initialization falls in this angle with probability $\frac{2(\sigma_1+\sigma_2)}{\pi}$.

If the initialization is good enough in the sense that the mean vectors do not fall in the angles defined above, then all points in the $\sigma_1$-cone concentrated around $e_1$ are closer to one of the initialized mean vector than another. Without loss of generality we denote the closer mean vector as $\mathbf{a}_1$ and another mean vector as $\mathbf{a}_2$.

Now by the definition of OCS of a finite sample and the update rule for $k$-means algorithm, the updated $\mathbf{a}_1$ has $e_1$ coordinate at least

$$\frac{(1-\delta)|\{i \in [n] | z_i = 1\}| \cos\sigma_1 - \delta n}{|\{i \in [n] | z_i = 1\}| + \delta n}$$

and $e_2$ coordinate at most

$$\frac{\delta n + (1-\delta)|\{i \in [n] | z_i = 1\}| \sin\sigma_1}{(1-\delta)|\{i \in [n] | z_i = 1\}|}.$$

Based on the assumptions stated in the proposition, we know all points in the $\sigma_1$-cone concentrated around $e_1$ are closer to the updated $\mathbf{a}_1$ vector than the updated $\mathbf{a}_2$ vector. The same analysis also applies to the $\sigma_2$-cone concentrated around $e_2$. Thus the update rule of $k$-means algorithm always keeps at least a $1-\delta$ fraction of the sample correctly labeled. By induction, this holds for finite steps of updates and the proposition holds for $N = 2$.

When $N > 2$, we can generalize the proof by similar procedures since the probability of the event that the initialized mean vector lies in the non-symmetric cone centered at the origin and the angular bisector of $e_i$ and $e_j$, with an angle $\sigma_i + \sigma_j$, is proportional to $\sigma_i + \sigma_j$ with a constant $c'_N$ depends only on $N$. Combine this for all angular bisectors of $e_i$ and $e_j$ for $i, j = 1, 2, \cdots, N, i \neq j$, the proof is completed. ■

Combining the major theorems and the Proposition 1, we get useful practical applications. Heuristically, if the original data set is sampled from a well-separated mixture model (with latent labels), the modified major theorems guarantee a high probability that the embedded data has an orthogonal cone structure (with labels of components considered). Proposition 1 guarantees that with high probability (taken over the random initializations), the $k$-means algorithm applied to the normalized embeddings has its classification error bounded by $\delta$ (up to relabelling).

### 3.5. Examples

In order to gain a better understanding of the meaning of the parameters describing the OCS, we now discuss several examples. Our first example is mixture of two Gaussians, and we consider both weighted Laplacian and kernel PCA embeddings. The mixture of normals allows for some more explicit computations, which helps to provide some insights. As another example, we use a mixture of uniform distributions, even though they violate Assumptions 1 due to the discontinuities of the densities on the boundaries. Nevertheless, this example still provides useful insights.

**3.5.1. Weighted Laplacian case. Mixture of two Gaussians**. Consider a mixture of two standard Gaussian densities on $\mathbb{R}$ obtained by shifting the two densities. More precisely, let $\rho = \frac{1}{2}\rho_1 + \frac{1}{2}\rho_2$, where, for some $\gamma \in \mathbb{R}$,

$$\rho_1(x) \coloneqq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}, \quad \rho_2(x) \coloneqq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\gamma)^2\right\}.$$

In this example, we illustrate the importance of both weighted overlapping parameter and coupling parameter. In order to compute the latter one, notice that

$$\frac{\rho_1'(x)}{\rho_1(x)} = -x, \quad \frac{\rho_2'(x)}{\rho_2(x)} = -(x-\gamma),$$

and

$$\frac{\rho'(x)}{\rho(x)} = -x + \frac{\gamma}{2}\frac{\rho_2(x)}{\rho(x)} = -(x-\gamma) - \frac{\gamma}{2}\frac{\rho_1(x)}{\rho(x)}.$$

So by the definition of coupling parameter, we have the following relationship that

$$\mathcal{C} \leq \mathcal{C}_1 + \mathcal{C}_2 = \frac{\gamma^2 q^2}{16}\int_{\mathbb{R}} \frac{\rho_1^2\rho_2^q + \rho_1^q\rho_2^2}{\rho^2}dx \leq \frac{\gamma^2 q^2}{16} 2^{3-q}\mathcal{S}_{12} \propto \gamma^2 \mathcal{S}_{12}.$$

Also, it is worth computing the weighted overlapping parameter itself as

$$\begin{aligned}
\mathcal{S}_{12} &= \int \left(\frac{\rho_1(x)\rho_2(x)}{\rho(x)}\right)^q dx \\
&= \int_{-\infty}^{\infty} \left(\frac{\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}x^2\}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\gamma)^2\}}{\frac{1}{2}\left(\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}x^2\} + \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\gamma)^2\}\right)}\right)^q dx \\
&= \left(\frac{2}{\pi}\right)^{\frac{q}{2}} \int_{-\infty}^{\infty} \left(\frac{\exp\{-\frac{1}{2}x^2 - \frac{1}{2}(x-\gamma)^2\}}{\exp\{-\frac{1}{2}x^2\} + \exp\{-\frac{1}{2}(x-\gamma)^2\}}\right)^q dx.
\end{aligned}$$

Heuristically, larger $\gamma$ corresponds to stronger separation. The above computations are consistent with this intuition. When $\gamma$ is large, $\mathcal{S}_{12}$ decays exponentially fast and the coupling parameter $\mathcal{C}$ decays as well, and thus intuitively the mixture model is well-separated. Also, when $\gamma$ is close to zero, $\mathcal{S}_{12}$ is close to $I = \int \rho^q(x)dx$ but $\mathcal{C}$ is small of order $\gamma^2$. This example shows that both weighted overlapping parameter and coupling parameter are essential to evaluate whether a model is well-separated or not. In the following example, the importance of indivisibility parameter $\Theta$ is also illustrated.

**Mixture of two Uniforms**. Consider a mixture of two uniform densities on $\mathbb{R}^k$ ($k \in \mathbf{Z}$) obtained by shifting the two densities. More precisely, let $\rho = \frac{1}{2}\rho_1 + \frac{1}{2}\rho_2$, where

$$\rho_1(x) \coloneqq \mathbf{1}_{[0,1]^k}(x), \quad \rho_2(x) \coloneqq \frac{1}{(b-a)^k}\mathbf{1}_{[a,b]^k}(x), \quad \text{for} \quad x \in \mathbb{R}^k.$$

We assume $0 < a < 1 < b$ since it is the best order to depict the separateness, and under this assumption, the two components have overlapping parts and neither one covers another one. After tedious but straightforward computations, we get

$$I_1 = 1, \quad I_2 = (b-a)^{k-kq}, \quad I_{\min} = \min\{1, (b-a)^{k-kq}\}, \quad I_{\max} = \min\{1, (b-a)^{k-kq}\},$$

$$I = \frac{1}{2^q}\left[1 - (1-a)^k + (b-a)^{k-kq} - (1-a)^k(b-a)^{-kq} + (1-a)^k(1+(b-a)^{-k})^q\right],$$

$$\mathcal{S}_{12} = 2^q(1-a)^k(1+(b-a)^k)^{-q}, \quad \overline{\mathcal{S}_1} = \overline{\mathcal{S}_2} = (1-a)^k(1+(b-a)^k)^{-q}.$$

But $\mathcal{C}$ and $\Theta$ are not well-defined since assumption 1 is not satisfied. We can just infer the extent of well-separation based on $\overline{\mathcal{S}_1}$. There are three parameters in $\overline{\mathcal{S}_1}$. For fixed $a$ and $b$, $\overline{\mathcal{S}_1}$ decreases as $q$ increases. For fixed $q$ and $a$, $\overline{\mathcal{S}_1}$ decreases as $b$ increases. For fixed difference $b-a$, $\overline{\mathcal{S}_1}$ decreases as $a$ increases (and, simultaneously, $b$ increases). We can understand these behaviors heuristically. For fixed $a$ and $b$, an increase of $q$ leads to bigger change of the second density, which can be an increase of the value of $\rho_2^q$ on its support when $b-a < 1$ or a decrease of the value of $\rho_2^q$ on its support when $b-a > 1$. For both the sharp case ($b-a < 1$) and the flat case ($b-a > 1$), the power $q$ also affects the denominator in the original formula of weighted overlapping parameter and leads to consistent decreasing result. For fixed difference $b-a$, when $a$ increase to 1, the supports of two components have less overlapping and thus $\overline{\mathcal{S}_1}$ is smaller. For fixed $q$ and $a$, when $b$ increase, the second component is more concentrated, resulting in a higher value for $\overline{\mathcal{S}_1}$.

More generally, the relative concentration of every component, the closeness of all pairs of components and power parameter $q$ affect the value of three key parameters of mixture model. We will also explore this numerically in the simulation section. In order to see the importance of indivisibility parameter $\Theta$ for the definition of well-separated mixture models, we consider the simple special case of a mixture of two uniforms where $a = 1, b = 2$ and dimension $k = 1$, i.e.

$$\rho_1(x) \coloneqq \mathbf{1}_{[0,1]}(x), \quad \rho_2(x) \coloneqq \mathbf{1}_{[1,2]}(x), \quad \text{for} \quad x \in \mathbb{R}.$$

Heuristically, two components are right next to each other and they have no overlap. It is not surprising that data generated from these two components are indivisible when a clustering algorithm is applied. However, straightforward computation shows that $\overline{\mathcal{S}_1} = \overline{\mathcal{S}_2} = 0$, seems indicating a well-separated model. But undefined $\mathcal{C}$ and $\Theta$ led by lack of Assumption 1 forbid the inference of well-separateness based on our theorems. Small modification can be applied to rectify this example: Assume the modified densities (for a given small value $\varepsilon > 0$) are defined as

$$\rho_1(x) = \begin{cases} 0 & \text{if } x < -\varepsilon, \\ \frac{(x+\varepsilon)^2}{2\varepsilon^2} & \text{if } -\varepsilon \le x < 0, \\ -\frac{(x-\varepsilon)^2}{2\varepsilon^2} + 1 & \text{if } 0 \le x < \varepsilon, \\ 1 & \text{if } \varepsilon \le x < 1-\varepsilon, \\ -\frac{(x-(1-\varepsilon))^2}{2\varepsilon^2} + 1 & \text{if } 1-\varepsilon \le x < 1, \\ \frac{(x-(1+\varepsilon))^2}{2\varepsilon^2} & \text{if } 1 \le x < 1+\varepsilon, \\ 0 & \text{if } x \ge \varepsilon, \end{cases} \quad \rho_2(x) = \begin{cases} 0 & \text{if } x < 1-\varepsilon, \\ \frac{(x-(1-\varepsilon))^2}{2\varepsilon^2} & \text{if } 1-\varepsilon \le x < 1, \\ -\frac{(x-(1+\varepsilon))^2}{2\varepsilon^2} + 1 & \text{if } 1 \le x < 1+\varepsilon, \\ 1 & \text{if } 1+\varepsilon \le x < 2-\varepsilon, \\ -\frac{(x-(2-\varepsilon))^2}{2\varepsilon^2} + 1 & \text{if } 2-\varepsilon \le x < 2, \\ \frac{(x-(2+\varepsilon))^2}{2\varepsilon^2} & \text{if } 2 \le x < 2+\varepsilon, \\ 0 & \text{if } x \ge \varepsilon. \end{cases}$$

When $\varepsilon \to 0$, these two new densities converge almost surely to the two uniform distributions, respectively, and they are just two smoothed versions of uniform distributions. Fortunately, Assumption 1 is satisfied and the parameters can be computed. Also, when $\varepsilon$ increase from 0 to 0.5, the behaviors of $\rho_1$ and $\rho_2$ are more similar with normal distributions and the corresponding parameters also behave similar with the normal case. Thus our theorems support our intuition that these two components are not well-separated if $\varepsilon$ is too small, and the separateness increases as $\varepsilon$ increases.

**3.5.2. Kernel PCA case. Mixture of two Gaussians**. Consider the same mixture of two standard Gaussian densities as in previous case. To be consistent with our theorem, we use the notation about corresponding probability measure and let $\nu = \frac{1}{2}\nu_1 + \frac{1}{2}\nu_2$, where

$$\nu_1(dx) = \rho_1(x) \coloneqq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}, \quad \nu_2(dx) = \rho_2(x) \coloneqq \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\gamma)^2\right\}.$$

The Gaussian kernel $k(x,y) = \frac{1}{\sqrt{\pi h}} \exp\left\{-\frac{(x-y)^2}{h}\right\}$ is chosen to illustate this example. We can see how the parameters defined in previous part behave with the off-set $\gamma$ and the bandwidth $h$. (We assume $\gamma \geq 0$ without loss of generality.)

Firstly, we can compute the kernelized densities as

$$q_1(x) = \int_\Omega k(x,y)\nu_1(dy) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi h}} \exp\left\{-\frac{(x-y)^2}{h}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\} dy = \sqrt{\frac{1}{\pi(h+2)}} \exp\left\{-\frac{1}{h+2}x^2\right\},$$

$$q_2(x) = \int_\Omega k(x,y)\nu_2(dy) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi h}} \exp\left\{-\frac{(x-y)^2}{h}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\gamma)^2\right\} dy = \sqrt{\frac{1}{\pi(h+2)}} \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\},$$

$$q(x) = \frac{1}{2}\sqrt{\frac{1}{\pi(h+2)}} \left(\exp\left\{-\frac{1}{h+2}x^2\right\} + \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\}\right),$$

and

$$\mathbb{E}_\nu q(X) = \int_\Omega q(x)\nu(dx) = \exp\left\{-\frac{\gamma^2}{(h+4)\sqrt{\pi(h+4)}}\right\}.$$

Since we only have two components, the overlapping parameter of $\frac{q_i(x)}{q(x)}$ can be computed as follows:

$$\begin{aligned}
\mathcal{S}_{12} &= \int \frac{q_1(x)q_2(x)}{q^2(x)}\nu(dx) \\
&= 4\int_{-\infty}^{\infty} \frac{\frac{1}{\pi(h+2)} \exp\left\{-\frac{1}{h+2}x^2\right\} \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\}}{\frac{1}{\pi(h+2)}\left(\exp\left\{-\frac{1}{h+2}x^2\right\} + \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\}\right)^2} \frac{1}{2\sqrt{2\pi}} \left(\exp\left\{-\frac{1}{2}x^2\right\} + \exp\left\{-\frac{(x-\gamma)^2}{2}\right\}\right) dx \\
&= \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} \frac{\exp\left\{-\frac{1}{h+2}x^2 - \frac{1}{h+2}(x-\gamma)^2\right\} \left(\exp\{-\frac{1}{2}x^2\} + \exp\left\{-\frac{(x-\gamma)^2}{2}\right\}\right)}{\left(\exp\left\{-\frac{1}{h+2}x^2\right\} + \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\}\right)^2} dx,
\end{aligned}$$

and

$$\overline{\mathcal{S}_1} = \overline{\mathcal{S}_2} = \frac{1}{2}\mathcal{S}_{12}.$$

The behavior of $\mathcal{S}_{12}$ can be analyzed based on $\gamma$ and $h$, respectively. For fixed $h$, $\mathcal{S}_{12}$ decreases as $\gamma$ increases. When $\gamma = 0$, $\mathcal{S}_{12}$=1; when $\gamma$ tends to infinity, $\mathcal{S}_{12}$ tends to 0. For fixed $\gamma$, $\mathcal{S}_{12}$ increases

as $h$ increases. When $h$ tends to zero, $\mathcal{S}_{12}$ tends to a fixed value between 0 and 1; when $h$ tends to infinity, $\mathcal{S}_{12}$ tends to 1.

The overlapping parameter of $q_i(x)$ can also be computed similarly. We have that

$$\begin{aligned}
\mathcal{S}_{11}^* &= \int_\Omega q_1^2(x)\nu(dx) \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi(h+2)} \exp\left\{-\frac{2}{h+2}x^2\right\} \frac{1}{2\sqrt{2\pi}} \left(\exp\left\{-\frac{1}{2}x^2\right\} + \exp\left\{-\frac{(x-\gamma)^2}{2}\right\}\right) dx \\
&= \frac{1}{2\pi\sqrt{(h+2)(h+6)}} \left(1 + \exp\left\{-\frac{2}{h+6}\gamma^2\right\}\right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{S}_{22}^* &= \int_\Omega q_2^2(x)\nu(dx) \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi(h+2)} \exp\left\{-\frac{2}{h+2}(x-\gamma)^2\right\} \frac{1}{2\sqrt{2\pi}} \left(\exp\left\{-\frac{1}{2}x^2\right\} + \exp\left\{-\frac{(x-\gamma)^2}{2}\right\}\right) dx \\
&= \frac{1}{2\pi\sqrt{(h+2)(h+6)}} \left(1 + \exp\left\{-\frac{2}{h+6}\gamma^2\right\}\right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{S}_{12}^* &= \int_\Omega q_1(x)q_2(x)\nu(dx) \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi(h+2)} \exp\left\{-\frac{1}{h+2}x^2\right\} \exp\left\{-\frac{1}{h+2}(x-\gamma)^2\right\} \frac{1}{2\sqrt{2\pi}} \left(\exp\left\{-\frac{1}{2}x^2\right\} + \exp\left\{-\frac{(x-\gamma)^2}{2}\right\}\right) dx \\
&= \frac{1}{2\pi\sqrt{(h+2)(h+6)}} \exp\left\{-\frac{h+4}{(h+2)(h+6)}\gamma^2\right\}.
\end{aligned}$$

So we have

$$S_{\text{within}}^* = \min\{S_{11}^*, S_{22}^*\} = \frac{1}{2\pi\sqrt{(h+2)(h+6)}} \left(1 + \exp\left\{-\frac{2}{h+6}\gamma^2\right\}\right),$$

and

$$S_{\text{between}}^* = \mathcal{S}_{12}^* = \frac{1}{2\pi\sqrt{(h+2)(h+6)}} \exp\left\{-\frac{h+4}{(h+2)(h+6)}\gamma^2\right\}.$$

Thus $\frac{S_{\text{between}}^*}{S_{\text{within}}^*} = \frac{\exp\left\{-\frac{h+4}{(h+2)(h+6)}\gamma^2\right\}}{1+\exp\left\{-\frac{2}{h+6}\gamma^2\right\}}$. For a fixed $\gamma$, as $h$ decreases, this ratio also decreases to zero, which gives a good example of well-separation.

**Mixture of two Uniforms**. We use the same uniform densities as above. However, there is no closed form of the three important parameters since the computation requires convolutions between Gaussian densities and uniform densities with finite support. The good point is that we can still analyze the behaviors of these parameters with respect to the bandwidth and boundary values of

the uniform distribution ($a$ and $b$). Their behaviors are quite similar with previous example. More specifically, the behavior of $\mathcal{S}_{12}$ can be analyzed based on $a$, $b$ and $h$, respectively. For fixed $a$ and $b$, $\mathcal{S}_{12}$ increases as $h$ increases. For fixed $h$ and $a$, $\mathcal{S}_{12}$ decreases as $b$ increases. For fixed difference $b - a$, $\mathcal{S}_{12}$ decreases as $a$ increases (and, simultaneously, $b$ increases).

The choice of kernel and its bandwidth plays essential role in the well-separation behavior. This will also be explored in the simulation section.

CHAPTER 4

# Proof of main results

In order to show the existence of OCS based on Weighted Laplacian embedding, we could first show OCS in the scenarios of some simple embedding in the sense that the OCS under these scenarios are easier to be proved. Then intuitively the Weighted Laplacian embedding that is close to the embedding that has an OCS tends to be more probable to have an OCS. Wasserstein distance is used to indicate this closeness and we would show that an embedding tends to be more probable to have an OCS if another close embedding has one, here the closeness is quantified by Wasserstein distance. Specifically, we first show the closeness of functions $q_k$ and their projections onto the eigenspace spanned by the $N$ eigenfunctions corresponding to the smallest $N$ eigenvalues of $\Delta_\rho$. Some lower bound and upper bound of related norm of $q_k$ functions and eigenvalues are used to bound the differences among different measures. The OCS is first constructed for the ancillary measures geometrically, then necessary orthogonal transformations are applied without loss of generality. For discrete case, similar steps are applied to construct the relationship between desired measures and the ancillary measures. The closeness of empirical measure and the population level measure is also needed. To achieve this, we need to construct discretization mapping and interpolation mapping to connect them. Then the approximation errors are controlled by considering the convergence of eigenvalues, eigenvectors and corresponding eigenspaces. The convergence rate highly depends on sample size, the properties of the manifold and the densities.

These general ideas have been layed out in Garcia-Trillos et al.( [**89**]) for the unweighted Laplacian case, and they also apply the ideas to both settings considered here. However, the specifics are different for different cases. For example, the convergence of empirical covariance operator towards the (population) covariance operator is studied in different techniques. Similar steps can be easily applied on kernel CCA embedding case, and can also be generalized to other embeddings with some modifications in detailed steps.

### 4.1. OCS of spectral embedding: The population setting

We first introduce a basic formula for equivalence from direct computation, which is useful to transfer two inner products with respect to $\rho^p$ and $\rho^q$.

PROPOSITION 2.

$$\langle \Delta_\rho u, v\rangle_{\rho^p} = \langle \nabla u, \nabla v\rangle_{\rho^q} \coloneqq \langle \nabla u, \nabla v\rangle_{L^2(\nu)}.$$

PROOF.

$$\begin{aligned}
\langle \Delta_\rho u, v\rangle_{\rho^p} &= -\int \rho^{q-p}\Delta u v \rho^p dx - q\int \rho^{q-p-1}\nabla\rho\nabla u v\rho^p dx \\
&= -\int \rho^{q-p} v\rho^p d(\nabla u) - q\int \rho^{q-1}\nabla\rho\nabla u v dx \\
&= \int \nabla u(\rho^q dv + q v\rho^{q-1}d\rho) - q\int \rho^{q-1}\nabla\rho\nabla u v dx \\
&= \int \nabla u\nabla v\rho^q dx \\
&= \langle \nabla u, \nabla v\rangle_{\rho^q} \\
&\coloneqq \langle \nabla u, \nabla v\rangle_{L^2(\nu)}.
\end{aligned}$$

□

DEFINITION 5. *(Wasserstein distance) Let $\mu_1, \mu_2$ be two probability measures on $\mathbb{R}^k$ with finite second moments. We define their Wasserstein distance by*

$$\left(W_2\left(\mu_1, \mu_2\right)\right)^2 \coloneqq \min_{\pi\in\Gamma(\mu_1,\mu_2)} \int_{\mathbb{R}^k\times\mathbb{R}^k} |x-y|^2 d\pi(x,y),$$

*where $\Gamma\left(\mu_1, \mu_2\right)$ stands for the set of transportation plans between $\mu_1$ and $\mu_2$, that is, the set of probability measures defined on $\mathbb{R}^k \times \mathbb{R}^k$ with first and second marginals equal to $\mu_1$ and $\mu_2$ respectively.*

The following proposition shows that two embeddings that are close in the Wasserstein distance have close orthogonal cone structures in the sense that the difference of the parameters for the two structures are close.

PROPOSITION 3. *Let $\mu_1, \mu_2$ be two probability measures on $\mathbb{R}^k$ with finite second moments and suppose that $\mu_1$ has an orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$, where $\sigma_k <$*

$\pi/4$ *for* $k = 1, 2, \cdots, N$. *Let* $s, t > 0$ *be such that*

$$\frac{rt\sin(s)}{\sqrt{k}} \geq W_2\left(\mu_1, \mu_2\right),$$

*and such that* $\sigma_k + s < \pi/4$ *for* $k = 1, 2, \cdots, N$. *Then,* $\mu_2$ *has an orthogonal cone structure with parameters* $(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, r(1 - \sin(s)))$.

Proof. This proposition is proved by Garcia Trillos, Hoffman and Hosseini (2019)( [**89**]), where the mixture model that consists of $C^1(\mathcal{M})$ probability density functions are covered. But the proof of this proposition only needs the setting with probability measures, which is satisfied by our model. □

Based on previous proposition, the idea of our proof is to show the OCS of $F_\sharp^Q\nu$ and the closeness of $F_\sharp^Q\nu$ and $F_\sharp\nu$. To quantify the closeness of $F_\sharp^Q\nu$ and $F_\sharp\nu$, we start from considering the projection of $q_k$ to the subspace of eigenspace of $\Delta_\rho$. Intuitively, this closeness can be implied if the projection of $q_k$ is close to $q_k$ itself, which can be derived from the next proposition.

Proposition 4. *For every* $k = 1, \cdots, N$, *let* $q_k \coloneqq (\frac{w_k\rho_k}{\rho})^{\frac{q}{2}}$ *and* $\mathcal{C}_k = \frac{q^2}{4}\int \left|\frac{\nabla\rho_k}{\rho_k} - \frac{\nabla\rho}{\rho}\right|^2 \rho_k^q dx$. *Then*

$$\frac{1}{w_k^q}\|q_k - \pi_N(q_k)\|_{\rho^p}^2 \leq \frac{\alpha^{|q-p|}\mathcal{C}}{\lambda_{N+1}},$$

*where* $\Pi_N$ *stands for the projection onto* $U$, *the span of the* $N$ *eigenfunctions corresponding to the smallest* $N$ *eigenvalues of* $\Delta_\rho$.

Proof.

$$\begin{aligned}
\langle\Delta_\rho q_k, q_k\rangle_{\rho^p} &= \int |\nabla q_k|^2 \rho^q dx \\
&= \int \left|\frac{q}{2}\left(\frac{w_k\rho_k}{\rho}\right)^{\frac{q}{2}-1} w_k \frac{\rho\nabla\rho_k - \rho_k\nabla\rho}{\rho^2}\right|^2 \rho^q dx \\
&= \frac{q^2}{4} w_k^q \int \left|\frac{\nabla\rho_k}{\rho_k} - \frac{\nabla\rho}{\rho}\right|^2 \rho_k^q dx \\
&= w_k^q \mathcal{C}_k.
\end{aligned}$$

Also, we can write $q_k$ in the orthonormal basis of eigenfunctions $\{u_1, u_2, \ldots\}$ of $\Delta_\rho$ as

$$q_k = \sum_{l=1}^{\infty} a_{lk} u_l$$

for some coefficients $\{a_{lk}\}_{l\in\mathbb{N}}$. Thus we have

$$\langle \Delta_\rho q_k, q_k \rangle_{\rho^q} = \sum_{l=1}^{N} a_{lk}^2 \lambda_l + \sum_{l=N+1}^{\infty} a_{lk}^2 \lambda_l.$$

Notice that $\langle u, v\rangle_{\rho^q} \leq \alpha^{|q-p|}\langle u, v\rangle_{\rho^p}$, where $\alpha$ is the assumed upper bound for $\rho$ (see Assumption 2), and recall that the eigenvalues have increasing order, so we have

$$\alpha^{|q-p|} w_k^q \mathcal{C} \geq \alpha^{|q-p|} w_k^q \mathcal{C}_k \geq \lambda_{N+1} \sum_{l=N+1}^{\infty} a_{lk}^2 = \lambda_{N+1} \|q_k - \pi_N(q_k)\|_{\rho^q}^2,$$

i.e.

$$\frac{1}{w_k^q} \|q_k - \pi_N(q_k)\|_{\rho^q}^2 \leq \frac{\alpha^{|q-p|}\mathcal{C}}{\lambda_{N+1}}.$$

□

Proposition 4 will be used to derive a lower bound for $\lambda_{N+1}$ (see Proposition 5). For this aim, we will need the following two lemmas.

Lemma 1. *For every $j \in \{1, \cdots, N\}$,*

$$\left| \langle q_j, q_j \rangle_{\rho_j^q} - I_j \right| \leq \max(N^{q-1}, 1)\mathcal{S}_b \sum_{k\neq j} w_k^q + \left(\max(N^{q-1}, 1) - 1\right) w_j^q \mathcal{S}_w.$$

Proof.

$$\begin{aligned} \langle q_j, q_j \rangle_{\rho_j^q} &= \int (\frac{w_j \rho_j}{\rho})^q \rho_j^q dx \\ &= \int ((\frac{w_j \rho_j}{\rho})^q - 1)\rho_j^q dx + I_j, \end{aligned}$$

where

$$\left|\int\left(\left(\frac{w_j\rho_j}{\rho}\right)^q-1\right)\rho_j^q dx\right| \le \int\left|(w_j\rho_j)^q-\rho^q\right|\frac{\rho_j^q}{\rho^q}dx$$
$$\le \max(N^{q-1},1)\sum_{k\neq j}w_k^q\int\left(\frac{\rho_k\rho_j}{\rho}\right)^q dx+\left(\max(N^{q-1},1)-1\right)w_j^q\int\left(\frac{\rho_j^2}{\rho}\right)^q dx$$
$$= \max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j}w_k^q+\left(\max(N^{q-1},1)-1\right)w_j^q\mathcal{S}_{\mathrm{w}}.$$

Thus

$$\left|\langle q_j,q_j\rangle_{\rho_j^q}-I_j\right| \le \max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j}w_k^q+\left(\max(N^{q-1},1)-1\right)w_j^q\mathcal{S}_{\mathrm{w}}.$$

□

REMARK 8. *Notice that the bound on the right hand side can be simplified in the case of* $0<q\le 1$ *and* $q>1$*, respectively. When* $0<q\le 1$*, we have*

$$\left|\langle q_j,q_j\rangle_{\rho_j^q}-I_j\right| \le \mathcal{S}_b\sum_{k\neq j}w_k^q.$$

*And when* $q>1$*, we have*

$$\left|\langle q_j,q_j\rangle_{\rho_j^q}-I_j\right| \le N^{q-1}\mathcal{S}_b\sum_{k\neq j}w_k^q+\left(N^{q-1}-1\right)w_j^q\mathcal{S}_w.$$

*The former case is more useful in the sense that* $\mathcal{S}_b$

The case of $0<q\le 1$ is worth exploring as the bound can be arbitrarily small if the model are well-separated enough. In the extreme case that the supports of the mixture components have disjoint support, $\mathcal{S}_{\mathrm{b}}=0$ and $\langle q_j,q_j\rangle_{\rho_j^q}=I_j$. If the mixture components only have very small overlap and $\mathcal{S}_{\mathrm{b}}$ is quite small, then $\left|\langle q_j,q_j\rangle_{\rho_j^q}-I_j\right|$ is also small, which is useful to get better parameters of OCS in our main theorems. Similarly, some of the following results can also be simplified in the case of $0<q\le 1$ ans the simplified version is more useful in practice.

LEMMA 2. *For every* $j\in\{1,\dots,N\}$*,*

$$\inf_{\langle v,q_j\rangle_{\rho_j^p}=0}\frac{\int_{\mathcal{M}}|\nabla v|^2\rho_j^p dx}{\langle v,v\rangle_{\rho_j^q}} \ge \frac{\Theta}{\alpha^{|p-q|}}\left(1-\left(\max(N^{q-1},1)\mathcal{S}_b\sum_{k\neq j}w_k^q+\left(\max(N^{q-1},1)-1\right)w_j^q\mathcal{S}_w\right)\right).$$

PROOF. For fixed $j \in \{1, \cdots, N\}$, we pick a vector $v \in H_q^1\left(\mathcal{M}, \rho_j^q\right)$ such that $\langle v, q_j\rangle_{\rho_j^q} = 0$ and $\langle v, v\rangle_{\rho_j^q} = 1$, where

$$H_q^1\left(\mathcal{M}, \rho_j^q\right) := \left\{ u \in L^2\left(\mathcal{M}, \rho_j^q\right) \,\middle|\, \int_{\mathcal{M}} \left(|\nabla u|^2 + |u|^2\right) \rho_j^q dx < +\infty \right\}.$$

Notice that

$$\int |\nabla v|^2 \rho_j^q dx = \langle \Delta_\rho v, v\rangle_{\rho^p} \geq \frac{1}{\alpha^{|p-q|}} \langle \Delta_\rho v, v\rangle_{\rho^q} = \frac{1}{\alpha^{|p-q|}} \sum_{k=1}^{\infty} \langle v, e_{j,k}\rangle_{\rho_j^q}^2 \lambda_{j,k},$$

where $\{\lambda_{j,k}, e_{j,k}\}$ are the orthonormal (w.r.t. $\langle\cdot,\cdot\rangle_{\rho_j^q}$ ) eigenpairs of $\Delta_{\rho_j}$ with $\lambda_{j,1} = 0, e_{j,1} = \mathbf{1}$.

So

$$\begin{aligned}
\int |\nabla v|^2 \rho_j^q dx &\geq \frac{\lambda_{j,2}}{\alpha^{|p-q|}} \sum_{k=2}^{\infty} \langle v, e_{j,k}\rangle_{\rho_j^q}^2 \\
&= \frac{\lambda_{j,2}}{\alpha^{|p-q|}} \left( \langle v, v\rangle_{\rho_j^q}^2 - \langle v, e_{j,1}\rangle_{\rho_j^q}^2 \right) \\
&= \frac{\Theta_j}{\alpha^{|p-q|}} \left( 1 - \langle v, e_{j,1}\rangle_{\rho_j^q}^2 \right).
\end{aligned}$$

Then, to find an upper bound of $\langle v, e_{j,1}\rangle_{\rho_j^q}^2$ for the vector $v$ chosen above:

$$\langle v, e_{j,1}\rangle_{\rho_j^q} = \int v \rho_j^q dx = \int v(1 - q_j)\rho_j^q dx.$$

Using this formula and Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
\langle v, e_{j,1}\rangle_{\rho_j^p}^2 &\leq \|v\|_{\rho_j^q}^2 \int (1 - q_j)^2 \rho_j^q dx \\
&= \int (1 - q_j)^2 \rho_j^q dx \\
&= I_j + \int (q_j^2 - 2q_j)\rho_j^q dx \\
&= I_j - \int q_j^2 \rho_j^q dx \\
&= I_j - \langle q_j, q_j\rangle_{\rho_j^q} \\
&\leq \max(N^{q-1}, 1)\mathcal{S}_{\mathrm{b}} \sum_{k \neq j} w_k^q + \left(\max(N^{q-1}, 1) - 1\right) w_j^q \mathcal{S}_{\mathrm{w}},
\end{aligned}$$

where we have been using Lemma 1. Thus, we obtain

$$\int |\nabla v|^2 \rho_j^q dx \geq \frac{\Theta}{\alpha^{|p-q|}} \left(1 - \left(\max(N^{q-1},1)\mathcal{S}_{\rm b} \sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q \mathcal{S}_{\rm w}\right)\right),$$

which finishes the proof since the denominator is chosen to be one. □

REMARK 9. *In the case of* $0 < q \leq 1$*, simplified version of previous lemma is*

$$\inf_{\langle v, q_j\rangle_{\rho_j^p}=0} \frac{\int_{\mathcal{M}} |\nabla v|^2 \rho_j^p dx}{\langle v, v\rangle_{\rho_j^q}} \geq \frac{\Theta}{\alpha^{|p-q|}} \left(1 - \mathcal{S}_b \sum_{k\neq j} w_k^q\right).$$

PROPOSITION 5. ***(Lower bound for*** $\lambda_{N+1}$***):***

$$\lambda_{N+1} \geq \left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{N^q} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min}-\mathcal{S}_b}\right)} - \frac{\sqrt{N\mathcal{C}}}{I_{min}-\mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^2.$$

PROOF. For fixed $j \in \{1, \cdots, N\}$, we pick a vector $u$ such that $\langle u, q_j\rangle_{L^2(\nu)} = 0$ and $\langle u, u\rangle_{L^2(\nu)} = 1$. Then, by again using Lemma 1, we have

$$\begin{aligned}
\langle u, q_j\rangle_{\rho_j^q} &= \int u q_j \rho_j^q dx \\
&= \frac{1}{w_j^q} \int u q_j (w_j\rho_j)^q dx \\
&= \frac{1}{w_j^q} \int u q_j \left[(w_j\rho_j)^q - \rho^q\right] dx \\
&= -\frac{1}{w_j^q} \int u q_j \left[\rho^q - (w_j\rho_j)^q\right] dx.
\end{aligned}$$

So

$$\begin{aligned}
\left|\langle u, q_j\rangle_{\rho_j^q}\right| &\le \frac{1}{w_j^q}\int |u| q_j \left[\max(N^{q-1},1)\sum_{k\neq j}(w_k\rho_k)^q + (\max(N^{q-1},1)-1)(w_j\rho_j)^q\right] dx \\
&= \frac{\max(N^{q-1},1)}{w_j^q}\sum_{k\neq j} w_k^q \int |u| q_j \rho_k^q dx + \frac{(\max(N^{q-1},1)-1)}{w_j^q} w_j^q \int |u| q_j \rho_j^q dx \\
&\le \frac{\max(N^{q-1},1)}{w_j^q}\sum_{k\neq j} w_k^q \left(\int u^2 \rho_k^q dx\right)^{\frac{1}{2}} \left(\int q_j^2 \rho_k^q dx\right)^{\frac{1}{2}} \\
&+ (\max(N^{q-1},1)-1)\left(\int u^2 \rho_j^q dx\right)^{\frac{1}{2}} \left(\int q_j^2 \rho_j^q dx\right)^{\frac{1}{2}} \\
&\le \frac{\max(N^{q-1},1)}{\sqrt{w_j^q}}\sum_{k\neq j} w_k^q \left(\int u^2 \rho_k^q dx\right)^{\frac{1}{2}} \sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{w_j^q}\sqrt{\mathcal{S}_{\mathrm{w}}} \\
&\le \max(N^{q-1},1)\sqrt{\frac{\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q}{w_j^q}} + (\max(N^{q-1},1)-1)\sqrt{w_j^q \mathcal{S}_{\mathrm{w}}} \\
&\coloneqq I_j^N.
\end{aligned}$$

Define $v_j = u - \left(\frac{\langle u, q_j\rangle_{\rho_j^q}}{\langle q_j, q_j\rangle_{\rho_j^q}}\right) q_j$, then $v_j$ is orthogonal to $q_j$ w.r.t. $\langle\cdot,\cdot\rangle_{\rho_j^q}$. Thus

$$\int |\nabla v_j|^2 \rho_j^q dx = \int |\nabla u|^2 \rho_j^q dx - 2\frac{\langle u, q_j\rangle_{\rho_j^q}}{\langle q_j, q_j\rangle_{\rho_j^q}} \int \nabla q_j \nabla u \rho_j^q dx + \left(\frac{\langle u, q_j\rangle_{\rho_j^q}}{\langle q_j, q_j\rangle_{\rho_j^q}}\right)^2 \int |\nabla q_j|^2 \rho_j^q dx.$$

Also,

$$\int |\nabla q_j|^2 \rho_j^q dx = \frac{q^2}{4} w_j^q \int \left|\frac{\nabla \rho_j}{\rho_j} - \frac{\nabla \rho}{\rho}\right|^2 \rho_j^q dx = w_j^q \mathcal{C}_j \le \mathcal{C}_j.$$

Thus $\int |\nabla v_j|^2 \rho_j^q dx$ can be bounded by three parts:

$$\int |\nabla v_j|^2 \rho_j^q dx \le \int |\nabla u|^2 \rho_j^q dx + 2\left(\frac{I_j^N}{I_j^L}\right)\sqrt{\mathcal{C}}\left(\int |\nabla u|^2 \rho_j^q dx\right)^{\frac{1}{2}} + \left(\frac{I_j^N}{I_j^L}\right)^2 \mathcal{C},$$

where

$$\begin{aligned}
I_j^L &= I_j - \left(\max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q \mathcal{S}_{\mathrm{w}}\right), \\
I_j^U &= I_j + \left(\max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q \mathcal{S}_{\mathrm{w}}\right),
\end{aligned}$$

$$I_j^N = \max(N^{q-1},1)\sqrt{\frac{\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q}{w_j^q}} + (\max(N^{q-1},1)-1)\sqrt{w_j^q\mathcal{S}_{\mathrm{w}}}.$$

By Lemma 2, $\int |\nabla v_j|^2 \rho_j^q dx$ is lower bounded by

$$\frac{\Theta}{\alpha^{|p-q|}}\left(1-\left(\max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q\mathcal{S}_{\mathrm{w}}\right)\right)\langle v_j, v_j\rangle_{\rho_j^q}.$$

Also, we have

$$\langle v_j, v_j\rangle_{\rho_j^q} = \langle u,u\rangle_{\rho_j^q} - \frac{\langle u, q_j\rangle_{\rho_j^q}^2}{\langle q_j, q_j\rangle_{\rho_j^q}} \geq \langle u,u\rangle_{\rho_j^q} - \frac{(I_j^N)^2}{I_j^L}.$$

Thus

$$\begin{aligned}
&(1-\mathcal{S}_{\mathrm{adj}})\left(1-\left(\max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q\mathcal{S}_{\mathrm{w}}\right)\right)\left(\langle u,u\rangle_{\rho_j^q} - \frac{(I_j^N)^2}{I_j^L}\right)\\
&\leq \int |\nabla v_j|^2\rho_j^q dx \leq \int |\nabla u|^2\rho_j^q dx + 2\left(\frac{I_j^N}{I_j^L}\right)\sqrt{\mathcal{C}}\left(\int |\nabla u|^2\rho_j^q dx\right)^{\frac{1}{2}} + \left(\frac{I_j^N}{I_j^L}\right)^2\mathcal{C}.
\end{aligned}$$

Denote $\mathcal{S}_{\mathrm{adj}} \coloneqq \max_{j=1,2,\cdots,N}\left(\max(N^{q-1},1)\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q + \left(\max(N^{q-1},1)-1\right) w_j^q\mathcal{S}_{\mathrm{w}}\right)$. Then multiplying both sides of the above inequality by $w_j^q$ and adding over $j$.

Notice that

$$\begin{aligned}
\sum_{j=1}^N w_j^q I_j^N &= \sum_{j=1}^N w_j^q\left(\max(N^{q-1},1)\sqrt{\frac{\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q}{w_j^q}} + (\max(N^{q-1},1)-1)\sqrt{w_j^q\mathcal{S}_{\mathrm{w}}}\right)\\
&= \sum_{j=1}^N\left(\max(N^{q-1},1)\sqrt{w_j^q\mathcal{S}_{\mathrm{b}}\sum_{k\neq j} w_k^q} + (\max(N^{q-1},1)-1)\sqrt{w_j^{3q}\mathcal{S}_{\mathrm{w}}}\right)\\
&\leq N\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right),
\end{aligned}$$

and

$$\sum_{j=1}^{N} w_j^q (I_j^N)^2 = \sum_{j=1}^{N} w_j^q \left( \max(N^{q-1},1) \sqrt{\frac{\mathcal{S}_{\mathrm{b}} \sum_{k\neq j} w_k^q}{w_j^q}} + (\max(N^{q-1},1)-1)\sqrt{w_j^q \mathcal{S}_{\mathrm{w}}} \right)^2$$
$$= \sum_{j=1}^{N} \left( \max(N^{q-1},1) \sqrt{\mathcal{S}_{\mathrm{b}} \sum_{k\neq j} w_k^q} + (\max(N^{q-1},1)-1) w_j^q \sqrt{\mathcal{S}_{\mathrm{w}}} \right)^2$$
$$\leq \sum_{j=1}^{N} \left( N^{2q-2}\mathcal{S}_{\mathrm{b}} + w_j^{2q} (\max(N^{q-1},1)-1)^2 \mathcal{S}_{\mathrm{w}} + w_j^q \max(N^{q-1},1)(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{b}}\mathcal{S}_{\mathrm{w}}} \right)$$
$$\leq N \left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right)^2.$$

So the left hand side is lower bounded by

$$\sum_{j=1}^{N} \frac{w_j^q \Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}} \left( \langle u,u \rangle_{\rho_j^q} - \frac{(I_j^N)^2}{I_j^L} \right) \geq \frac{\Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}} \left( \frac{1}{\max(N^{q-1},1)} \langle u,u \rangle_{L^2(\nu)} - \sum_{j=1}^{N} \frac{w_j^q (I_j^N)^2}{I_j^L} \right)$$
$$\geq \frac{\Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}} \left( \frac{1}{\max(N^{q-1},1)} - \frac{N\left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right)^2}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \right).$$

For the right hand side,

$$\sum_{j=1}^{N} w_j^q \int |\nabla u|^2 \rho_j^q dx \leq \int |\nabla u|^2 \sum_{j=1}^{N} (w_j \rho_j)^q dx = \int |\nabla u|^2 \rho^q dx,$$

$$\sum_{j=1}^{N} w_j^q \left( \frac{I_j^N}{I_j^L} \right)^2 \mathcal{C} \leq \frac{\mathcal{C}}{(I_{\min} - \mathcal{S}_{\mathrm{b}})^2} \sum_{j=1}^{N} w_j^q (I_j^N)^2 \leq \frac{\mathcal{C}}{(I_{\min} - \mathcal{S}_{\mathrm{b}})^2} N \left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right)^2,$$

and

$$
\begin{aligned}
&\sum_{j=1}^{N} 2w_j^q \left(\frac{I_j^N}{I_j^L}\right) \sqrt{\mathcal{C}} \left(\int |\nabla u|^2 \rho_j^q dx\right)^{\frac{1}{2}} \\
&\leq 2\frac{\sqrt{\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \sum_{j=1}^{N} w_j^q I_j^N \left(\int |\nabla u|^2 \rho_j^q dx\right)^{\frac{1}{2}} \\
&\leq 2\frac{\sqrt{\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \sum_{i=1}^{N} \left(w_j^q (I_j^N)^2\right)^{\frac{1}{2}} \sum_{i=1}^{N} \left(w_j^q \int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}} \\
&\leq 2\frac{N\sqrt{\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(\sum_{i=1}^{N} w_j^q (I_j^N)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} w_j^q \int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}} \\
&\leq 2\frac{N\sqrt{\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(N \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2\right)^{\frac{1}{2}} \left(\int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}}.
\end{aligned}
$$

So the right hand side is upper bounded by

$$
\begin{aligned}
&\int |\nabla u|^2 \rho^q dx + 2\frac{N\sqrt{\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(N \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2\right)^{\frac{1}{2}} \left(\int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}} \\
&\qquad + \frac{\mathcal{C}}{(I_{\min} - \mathcal{S}_{\mathrm{b}})^2} N \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2 \\
&\qquad \leq N \left(\left(\int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}} + \frac{\sqrt{N\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)\right)^2.
\end{aligned}
$$

Combine them together:

$$
\begin{aligned}
&\frac{\Theta(1 - \mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}} \left(\frac{1}{\max(N^{q-1}, 1)} - \frac{N \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2}{I_{\min} - \mathcal{S}_{\mathrm{b}}}\right) \\
&\qquad \leq N \left(\left(\int |\nabla u|^2 \rho^q dx\right)^{\frac{1}{2}} + \frac{\sqrt{N\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)\right)^2 \\
&\qquad = N \left(\|\nabla u\|_{L^2(\nu)} + \frac{\sqrt{N\mathcal{C}}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left(\max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)\right)^2.
\end{aligned}
$$

Thus

$$\sqrt{\frac{\Theta(1-\mathcal{S}_{\text{adj}})}{\alpha^{|p-q|}}\left(\frac{1}{\max(N^q,N)}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)^2}{I_{\min}-\mathcal{S}_{\text{b}}}\right)}$$
$$-\frac{\sqrt{N\mathcal{C}}}{I_{\min}-\mathcal{S}_{\text{b}}}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)\leq\|\nabla u\|_{L^2(\nu)}.$$

Since the above inequality holds for all vector $u$ such that $\langle u,q_j\rangle_{L^2(\nu)}=0$ and $\langle u,u\rangle_{L^2(\nu)}=1$, we have

$$\begin{aligned}\lambda_{N+1}&\geq\min_{u\in Q^\perp}\frac{\int|\nabla u|^2\rho_k^q dx}{\int u^2\rho^q dx}\\&\geq\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{\text{adj}})}{\alpha^{|p-q|}}\left(\frac{1}{\max(N^q,N)}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)^2}{I_{\min}-\mathcal{S}_{\text{b}}}\right)}\right.\\&\left.-\frac{\sqrt{N\mathcal{C}}}{I_{\min}-\mathcal{S}_{\text{b}}}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)\right)^2.\end{aligned}$$

□

REMARK 10. *In the case of* $0<q\leq 1$*, simplified version of previous proposition is*

$$\lambda_{N+1}\geq\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{N^q}-\frac{\mathcal{S}_b}{I_{min}-\mathcal{S}_b}\right)}-\frac{\sqrt{N\mathcal{C}\mathcal{S}_b}}{I_{min}-\mathcal{S}_b}\right)^2.$$

Combining proposition 4 and proposition 5, we get the following corollary:

COROLLARY 1. *For every* $k=1,\cdots,N$*, we have*

$$\frac{1}{w_k^q}\|q_k-\pi_N(q_k)\|_{\rho^q}^2\leq\alpha^{|q-p|}\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}\mathcal{C}}\left(\frac{1}{\max(N^q,N)}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min}-\mathcal{S}_b}\right)}\right.$$
$$\left.-\frac{\sqrt{N}}{I_{min}-\mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^{-2}.$$

*When* $0<q\leq 1$*,*

$$\frac{1}{w_k^q}\|q_k-\pi_N(q_k)\|_{\rho^q}^2\leq\alpha^{|q-p|}\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}\mathcal{C}}\left(\frac{1}{N}-\frac{\mathcal{S}_b}{I_{min}-\mathcal{S}_b}\right)}-\frac{\sqrt{N\mathcal{S}_b}}{I_{min}-\mathcal{S}_b}\right)^{-2}.$$

Corollary 1 quantifies the distance of the mixture components and their projections on the space spanned by the $N$ eigenfunctions corresponding to the smallest $N$ eigenvalues. In particular, if the term on the right-hand side (which essentially means, if the bound from Lemma 1) tends to 0, then so does the distance on the left-hand side.

PROPOSITION 6. *The probability measure $\mu^Q = F_\sharp^Q \nu$ has an orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ for any $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \frac{\pi}{4})$, $\delta^* \leq \delta < 1$ and $r = \frac{1}{\sqrt{\max(N^{q-1},1) w_{\max}^q}}$ where*

$$\delta^* = \frac{N I_{max}}{I I_{min}} \left( \frac{w_{\max}}{w_{\min}} \right)^q \sum_{k=1}^N w_k^q \frac{\cos^2(\sigma_k)}{1 - \cos^2(\sigma_k)} \overline{\mathcal{S}_k}.$$

PROOF. For each $k = 1, \ldots, N$, let

$$C_k \coloneqq \left\{ z \in \mathbb{R}^N : \frac{z_k}{|z|} > \cos(\sigma_k), \quad |z| \geq r \right\}$$

with $r = \frac{1}{\sqrt{\max(N^{q-1},1) w_{\max}^q I_{\max}}}$ and fixed $\sigma_k \in (0, \pi/4)$ $(k = 1, 2, \cdots, N)$.

Also denote $A_k$ as the preimage of $C_k$ through $F^Q$, i.e.

$$A_k \coloneqq \left(F^Q\right)^{-1}(C_k) = \left\{ x \in \mathcal{M} : \frac{q_k(x)}{\sqrt{I_k w_k^q}} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2} > r \right\}.$$

Then we have

$$\mu^Q(C_k) = F_\sharp^Q \nu(C_k) = \nu(A_k),$$

and the condition $\left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2} > r$ is redundant because of the definition of $r$. Thus $A_k$ can be re-written as

$$A_k = \left\{ x \in \mathcal{M} : \sqrt{\frac{\rho_k^q(x)}{I_k \rho^q(x)}} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \sqrt{\frac{\rho_j^q(x)}{I_j \rho^q(x)}} \right)^2 \right)^{1/2} \right\}.$$

For an arbitrary $x_0 \in A_k^c \subseteq \Omega$ $(k = 1, 2, \cdots, N)$ we have

$$\frac{\rho_k^q(x_0)}{I_k \rho^q(x_0)} \leq \cos^2(\sigma_k) \sum_{j=1}^N \frac{\rho_j^q(x_0)}{I_j \rho^q(x_0)},$$

i.e.,

$$\left(1-\cos^2(\sigma_k)\right)\frac{\rho_k^q(x_0)}{I_k\rho^q(x_0)} \le \cos^2(\sigma_k)\sum_{j\neq k}\frac{\rho_j^q(x_0)}{I_j\rho^q(x_0)}.$$

So

$$\frac{\rho_k^q(x_0)}{I_k\rho^q(x_0)} \le \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\sum_{j\neq k}\frac{\rho_j^q(x_0)}{I_j\rho^q(x_0)}.$$

Thus

$$w_k^{2q}\frac{\rho_k^q(x_0)}{\rho^q(x_0)}\frac{\rho_k^q(x_0)}{\rho^q(x_0)} \le w_k^{2q}\frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\sum_{j\neq k}\frac{I_k}{I_j}\frac{\rho_j^q(x_0)}{\rho^q(x_0)}\frac{\rho_k^q(x_0)}{\rho^q(x_0)}.$$

Take the integral over $A_k^c$ on both sides:

$$\int_{A_k^c} w_k^{2q}\frac{\rho_k^q(x)}{\rho^q(x)}\frac{\rho_k^q(x)}{\rho^q(x)}\rho^q(x)dx \le \int_{A_k^c} w_k^{2q}\frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\sum_{j\neq k}\frac{I_k}{I_j}\frac{\rho_j^q(x)}{\rho^q(x)}\frac{\rho_k^q(x)}{\rho^q(x)}\rho^q(x)dx, \quad \forall k=1,\dots,N.$$

Take the sum over $k$:

$$\sum_{k=1}^N\int_{A_k^c} w_k^{2q}\frac{\rho_k^q(x)}{\rho^q(x)}\frac{\rho_k^q(x)}{\rho^q(x)}\rho^q(x)dx \le \sum_{k=1}^N\int_{A_k^c} w_k^{2q}\frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\sum_{j\neq k}\frac{I_k}{I_j}\frac{\rho_j^q(x)}{\rho^q(x)}\frac{\rho_k^q(x)}{\rho^q(x)}\rho^q(x)dx,$$

where

$$\begin{aligned}
\text{LHS} &= \sum_{k=1}^N\int_{A_k^c} w_k^{2q}\frac{\rho_k^{2q}(x)}{\rho^{2q}(x)}\rho^q(x)dx \\
&\ge \int_{A_k^c}\sum_{k=1}^N\left(\frac{w_k\rho_k}{\rho}\right)^{2q}\rho^q dx \\
&= \int_{\mathcal{M}}\sum_{k=1}^N\left(\frac{w_k\rho_k\mathbf{1}_{A_k^c}(x)}{\rho}\right)^{2q}\rho^q dx \\
&= \frac{1}{N}\int_{\mathcal{M}}\sum_{k=1}^N\left(\frac{w_k\rho_k\mathbf{1}_{A_k^c}(x)}{\rho}\right)^{2q}\sum_{k=1}^N 1\rho^q dx \\
&\ge \frac{1}{N}\int_{\mathcal{M}}\left(\sum_{k=1}^N\frac{w_k\rho_k\mathbf{1}_{A_k^c}(x)}{\rho}\right)^{2q}\rho^q dx \\
&\ge \frac{1}{N}\int_{\mathcal{M}}\mathbf{1}_{\bigcap_{l=1}^N A_l^c}(x)\left(\sum_{k=1}^N\frac{w_k\rho_k}{\rho}\right)^{2q}\rho^q dx \\
&= \frac{1}{N}\int_{\bigcap_{l=1}^N A_l^c}\left(\sum_{k=1}^N\frac{w_k\rho_k}{\rho}\right)^{2q}\rho^q dx \\
&= \frac{I}{N}\nu\left(\bigcap_{l=1}^N A_l^c\right).
\end{aligned}$$

$$
\begin{aligned}
\text{RHS} &= \sum_{k=1}^{N} \int_{A_k^c} w_k^{2q} \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \sum_{j\neq k} \frac{I_k}{I_j} \frac{\rho_j^q(x)}{\rho^q(x)} \frac{\rho_k^q(x)}{\rho^q(x)} \rho^q(x) dx \\
&= \sum_{k=1}^{N} \int_{\mathcal{M}} w_k^{2q} \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \sum_{j\neq k} \frac{I_k}{I_j} \frac{\rho_j^q(x)}{\rho^q(x)} \frac{\rho_k^q(x)}{\rho^q(x)} \mathbf{1}_{A_k^c}(x) \rho^q(x) dx \\
&\leq \int_{\mathcal{M}} \sum_{k=1}^{N} w_k^{2q} \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \sum_{j\neq k} \frac{I_k}{I_j} \frac{\rho_j^q(x)}{\rho^q(x)} \frac{\rho_k^q(x)}{\rho^q(x)} \rho^q(x) dx \\
&\leq \left(\frac{w_{\max}}{w_{\min}}\right)^q \int_{\mathcal{M}} \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \frac{\rho_k^q(x)}{\rho^q(x)} \sum_{j\neq k} \frac{I_k}{I_j} w_j^q \frac{\rho_j^q(x)}{\rho^q(x)} \rho^q(x) dx \\
&\leq \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \sum_{j\neq k} w_j^q \frac{I_k}{I_j} \mathcal{S}_{jk} \\
&\leq \frac{I_{\max}}{I_{\min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \sum_{j\neq k} w_j^q \mathcal{S}_{jk} \\
&= \frac{I_{\max}}{I_{\min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \overline{\mathcal{S}_k}.
\end{aligned}
$$

Thus we have

$$
\nu\left(\bigcap_{l=1}^{N} A_l^c\right) \leq \frac{N I_{\max}}{I I_{\min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \overline{\mathcal{S}_k},
$$

and this implies

$$
\mu^Q\left(\bigcup_{k=1}^{N} C_k\right) \geq 1 - \frac{N I_{\max}}{I I_{\min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{k=1}^{N} w_k^q \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \overline{\mathcal{S}_k},
$$

which completes the proof. □

The next auxiliary lemma will be used in the proof of Theorem 1 below:

LEMMA 3. *Let $V$ be a vector space of dimension $N$ and let $\langle \cdot, \cdot \rangle$ be an inner product on $V$ with associated norm $\| \cdot \|$. Suppose that $v_1, v_2, \cdots, v_N$ are linearly independent unit vectors in $V$ such that*

$$
|\langle v_j, v_l \rangle| \leq \delta, \quad \forall j \neq l,
$$

*for $\delta > 0$ satisfying*

$$
N\delta < 1.
$$

*Then, there exists an orthonormal basis for $V$, $\{\tilde{v}_1, \dots \tilde{v}_N\}$, such that for every $j = 1, \cdots, N$*

$$\|v_j - \tilde{v}_j\| \leq \tilde{\phi}(N, \delta),$$

*where*

$$\tilde{\phi}(N, \delta) \coloneqq \sqrt{N}\left[\frac{1}{\sqrt{1 - N\delta}} - 1\right].$$

Proof. This lemma is proved by Garcia Trillos, Hoffman and Hosseini (2019)( [**89**]) in the appendix. □

Now we can prove the original theorem under population setting as follows:

**Proof of Theorem 1**. The measure $\mu = F_\sharp \nu$ has the same orthogonal cone structure as the measure $(OF)_\sharp \nu$, where the map $(OF)_\sharp \nu$ is defined by $x \in \Omega \mapsto OF(x) \in \mathbb{R}^N$ with $O$ being an $N \times N$ orthogonal matrix. So we will consider the measure $(OF)_\sharp \nu$ where we construct the matrix $O$ such that $(OF)_\sharp \nu$ and $F^Q_\sharp \nu$ are close to each other in the 2-Wasserstein distance. Then combining with previous propositions, we can get the orthogonal cone structure for $(OF)_\sharp \nu$. Firstly, define the normalized projection of $q_i$'s as follows:

$$v_i \coloneqq \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_{L^2(\nu)}}, \quad i = 1, \dots, N,$$

where $\Pi_N : L^2(d\nu) \to U$ is the orthogonal projection onto $U$, the span of the $N$ eigenfunctions corresponding to the $N$ smallest eigenvalues of $\Delta_\rho$.

We first consider the degenarate case where $p = q$, now we have the following equivalence:

$$\|\cdot\|_{\rho^p} = \|\cdot\|_{\rho^q} = \|\cdot\|_{L^2(\nu)}.$$

Then based on the previous proposition, we have

$$
\begin{aligned}
\left\| \frac{q_i}{\sqrt{w_i^q I_i}} - v_i \right\|_{L^2(\nu)} &= \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_{L^2(\nu)}} \right\|_{L^2(\nu)} \\
&\le \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \frac{\Pi_N(q_i)}{\sqrt{w_i^q I_i}} \right\|_{L^2(\nu)} + \left\| \frac{\Pi_N(q_i)}{\sqrt{w_i^q I_i}} - \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_{L^2(\nu)}} \right\|_{L^2(\nu)} \\
&= \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \frac{\Pi_N(q_i)}{\sqrt{w_i^q I_i}} \right\|_{L^2(\nu)} + \frac{1}{\sqrt{w_i^q I_i}} \left| \|\Pi_N(q_i)\|_{L^2(\nu)} - \sqrt{w_i^q I_i} \right| \\
&\le 2 \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \frac{\Pi_N(q_i)}{\sqrt{w_i^q I_i}} \right\|_{L^2(\nu)} \\
&\le \frac{2\alpha^{\frac{|q-p|}{2}}}{\sqrt{I_i}} \left( \sqrt{\frac{\Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}\mathcal{C}} \left( \frac{1}{\max(N^q, N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \right)} \right. \\
&\quad \left. - \frac{\sqrt{N}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right) \right)^{-1} \\
&\le \frac{2\alpha^{\frac{|q-p|}{2}}}{\sqrt{I_{\min}}} \left( \sqrt{\frac{\Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}\mathcal{C}} \left( \frac{1}{\max(N^q, N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \right)} \right. \\
&\quad \left. - \frac{\sqrt{N}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right) \right)^{-1}.
\end{aligned}
$$

For a given pair $(i,j)$ with $i \neq j$, we have

$$
\begin{aligned}
\left| \langle v_i, v_j \rangle_{L^2(\nu)} \right| &= \left| \left\langle v_i - \frac{q_i}{\sqrt{w_i^q I_i}}, v_j \right\rangle_{L^2(\nu)} + \left\langle \frac{q_i}{\sqrt{w_i^q I_i}}, v_j - \frac{q_j}{\sqrt{w_j^q I_j}} \right\rangle_{L^2(\nu)} + \left\langle \frac{q_i}{\sqrt{w_i^q I_i}}, \frac{q_j}{\sqrt{w_j^q I_j}} \right\rangle_{L^2(\nu)} \right| \\
&\le \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - v_i \right\|_{L^2(\nu)} + \left\| \frac{q_j}{\sqrt{w_j^q I_j}} - v_j \right\|_{L^2(\nu)} + \sqrt{\frac{I}{I_i I_j}} \mathcal{S}_{\mathrm{b}}^{\frac{1}{2}} \\
&\le \frac{4\alpha^{\frac{|q-p|}{2}}}{\sqrt{I_{\min}}} \left( \sqrt{\frac{\Theta(1-\mathcal{S}_{\mathrm{adj}})}{\alpha^{|p-q|}\mathcal{C}} \left( \frac{1}{\max(N^q, N)} - \frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}}\right)^2}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \right)} \right. \\
&\quad \left. - \frac{\sqrt{N}}{I_{\min} - \mathcal{S}_{\mathrm{b}}} \left( \max(N^{q-1},1)\sqrt{\mathcal{S}_{\mathrm{b}}} + (\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\mathrm{w}}} \right) \right)^{-1} + \frac{\sqrt{I\mathcal{S}_{\mathrm{b}}}}{I_{\min}} \\
&\coloneqq \tau.
\end{aligned}
$$

Thus we can conclude by an application of Lemma 3, that there exists an orthonormal basis $\tilde{v}_1, \ldots, \tilde{v}_N$ for $\left(U, \langle \cdot, \cdot \rangle_{L^2(\nu)}\right)$ such that

$$\|v_i - \tilde{v}_i\|^2_{L^2(\nu)} \leq N \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right)^2, \quad i = 1, \ldots, N.$$

Thus for any $i = 1, \cdots, N$,

$$\begin{aligned}
\left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \tilde{v}_i \right\|^2_{L^2(\nu)} &= \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - v_i \right\|^2_{L^2(\nu)} + 2 \left\langle v_i - \tilde{v}_i, \frac{q_i}{\sqrt{w_i^q I_i}} \right\rangle_{L^2(\nu)} - \langle v_i + \tilde{v}_i, v_i - \tilde{v}_i \rangle_{L^2(\nu)} \\
&\leq \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - v_i \right\|^2_{L^2(\nu)} + 4 \|v_i - \tilde{v}_i\|_{L^2(\nu)} \\
&\leq \frac{4\alpha^{|q-p|}}{I_{\min}} \left( \sqrt{ \frac{\Theta(1 - \mathcal{S}_{\text{adj}})}{\alpha^{|p-q|}\mathcal{C}} \left( \frac{1}{\max(N^q, N)} - \frac{\left( \max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\text{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\text{w}}} \right)^2}{I_{\min} - \mathcal{S}_{\text{b}}} \right) } \right. \\
&\quad \left. - \frac{\sqrt{N}}{I_{\min} - \mathcal{S}_{\text{b}}} \left( \max(N^{q-1}, 1)\sqrt{\mathcal{S}_{\text{b}}} + (\max(N^{q-1}, 1) - 1)\sqrt{\mathcal{S}_{\text{w}}} \right) \right)^{-2} + 4\sqrt{N} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right) \\
&= \left( \frac{\tau - \frac{\sqrt{I\mathcal{S}_{\text{b}}}}{I_{\min}}}{2} \right)^2 + 4\sqrt{N} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right).
\end{aligned}$$

Now define $\tilde{F} : \Omega \mapsto \mathbb{R}^N$ as the map $\tilde{F}(x) = \sum_{j=1}^N \tilde{v}_j(x) e_j$. Since both $\{\tilde{v}_1, \ldots, \tilde{v}_N\}$ and $\{u_1, \ldots, u_N\}$ are orthonormal bases for $\left(U, \langle \cdot, \cdot \rangle_{L^2(\nu)}\right)$, there exists an orthogonal matrix $O \in \mathbb{R}^N \times \mathbb{R}^N$ such that

$$OF = \tilde{F}.$$

Let $\pi \coloneqq \left( F^Q \times \tilde{F} \right)_\sharp \nu$, then it is a coupling between $F^Q_\sharp \nu$ and $\tilde{F}_\sharp \nu$. Thus we have

$$\begin{aligned}
W_2^2 \left( F^Q_\sharp \nu, \tilde{F}_\sharp \nu \right) &\leq \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} |z - \tilde{z}|^2 d\pi(z, \tilde{z}) \\
&= \int_\Omega \left| F^Q(x) - \tilde{F}(x) \right|^2 d\nu(x) \\
&= \sum_{i=1}^N \left\| \frac{q_i}{\sqrt{w_i^q I_i}} - \tilde{v}_i \right\|^2_{L^2(\nu)} \\
&= N \left( \frac{\tau - \frac{\sqrt{I\mathcal{S}_{\text{b}}}}{I_{\min}}}{2} \right)^2 + 4N^{\frac{3}{2}} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right).
\end{aligned}$$

Also, it's easy to check the finite second moments condition of the probability measures $F^Q_\sharp\nu$ and $\tilde{F}_\sharp\nu$ as follows:

$$\int_\Omega \left|F^Q(x)\right|^2 d\nu(x) = \sum_{i=1}^N \left\| \frac{q_i}{\sqrt{w_i^q I_i}} \right\|^2_{L^2(\nu)} = N,$$

$$\int_\Omega \left|\tilde{F}(x)\right|^2 d\nu(x) = \sum_{i=1}^N \|\tilde{v}_i\|^2_{L^2(\nu)} = N.$$

By using the proposition 3 , we know that $\mu$ has an orthogonal cone structure with parameters $\left(\sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1-\sin(s)}{\sqrt{N} w_{\max}}\right)$ for any $\delta \in [\delta^*, 1)$ and $s, t > 0$ satisfying

$$\frac{t^2 \sin^2(s)}{N^q w_{\max}^q} \geq N \left( \frac{\tau - \frac{\sqrt{I\mathcal{S}_{\mathrm{b}}}}{I_{\min}}}{2} \right)^2 + 4N^{\frac{3}{2}} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right), \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N.$$

■

### 4.2. OCS of spectral embedding: The sample setting

We first introduce some basic assumptions and auxiliary results that are useful to prove major theorems and various lemmas.

Definition 6. *(injectivity radius)*

- *The injectivity radius at a point $x$ of a manifold $\mathcal{M}$ is the largest radius for which the exponential map at $x$ is a diffeomorphism, where the exponential map ( [**80**]) is a map from a subset of a tangent space $\mathrm{T}_x\mathcal{M}$ of a manifold $\mathcal{M}$ to $\mathcal{M}$ itself.*
- *The injectivity radius of a manifold $\mathcal{M}$ is the infimum of the injectivity radii at all points.*

Assumption 4. *Assume that*

$$\varepsilon < \min\{1, \frac{i_0}{10}, \frac{1}{\sqrt{mK}}, \frac{R}{\sqrt{27m}}\} \text{ and } (m+5)\delta_n < \varepsilon,$$

*where $i_0$ is the injectivity radius of the manifold $\mathcal{M}$, $K$ is a global upper bound on the absolute value of sectional curvatures of $\mathcal{M}$, $m$ is the dimension of $\mathcal{M}$, and $R$ is the reach of $\mathcal{M}$.*

PROPOSITION 7. *Let $R$ be the reach of the manifold $\mathcal{M} \subseteq \mathbb{R}^d$, which is assumed to be strictly postive. Let $x, y \in \mathcal{M}$ and suppose that $|x-y| \leq \frac{R}{2}$. Then,*

$$|x-y| \leq d_{\mathcal{M}}(x,y) \leq |x-y| + \frac{8}{R^2}|x-y|^3.$$

THEOREM 9. *Assume $\mathcal{M}, \rho, \varepsilon$ and $n$ satisfy the Assumption 2 and Assumption 3. For every $\beta > 1$ there exists a constant $C_\beta > 0$ such that with probability at least $1 - C_\beta n^{-\beta}$, there exists a map $T_n : \mathcal{M} \to \{x_1, \cdots, x_n\}$ satisfying*

$$\nu\left(T_n^{-1}\left(\{\mathbf{x}_i\}\right)\right) = \frac{1}{n}, \quad \forall i = 1, \ldots, n,$$

*and*

$$\|g_j - u_{n,j} \circ T_n\|^2_{L^2(\nu)} \leq c_{\mathcal{M}} \left( \left( \frac{\lambda_N}{\lambda_{N+1} - \lambda_N} \right) \left( \varepsilon + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} \right) + \lambda_N \varepsilon^{m+2} \left( \varepsilon + \varepsilon^2 + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} + \frac{\log(n)^{p_m}}{n^{1/m}} \right) \right),$$

*for some orthonormal functions $g_1, \cdots, g_N \in L^2(\nu)$ belonging to $U$, the span of the $N$ eigenfunctions corresponding to the smallest $N$ eigenvalues of $\Delta_\rho$ with respect to $\langle \cdot, \cdot \rangle_{L^2(\nu)}$, and a constant $c_{\mathcal{M}} > 0$ depending only on $\mathcal{M}, N, \alpha, C_\rho$ and $\eta$.*

Firstly, we define the Dirichlet forms associated to $\Delta_n$ and $\Delta_\rho$, respectively, as

$$b_n(u_n) \coloneqq \frac{1}{2n} \sum_{i,j} (W_n)_{ij} \left| u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j) \right|^2, \quad u_n \in L^2(\nu_n),$$

$$D(u) \coloneqq \frac{1}{2} \int_{\mathcal{M}} |\nabla u|^2 \rho^q(x) dx, \quad u \in H^1_q(\mathcal{M}, \rho).$$

PROPOSITION 8. *Denote the geodesic distance in $\mathcal{M}$ as $d_{\mathcal{M}}$, and let*

$$p_m = \begin{cases} \dfrac{3}{4} & \text{for} \quad m = 2, \\ \dfrac{1}{m} & \text{for} \quad m \geq 3. \end{cases}$$

*Then for a given $\beta > 1$, there exists a constant $C_\beta > 0$ depending only on $\beta$ so that with probability at least $1 - C_\beta n^{-\beta}$ there exists a map $T_n : \mathcal{M} \to \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ such that:*

- $\nu(T_n^{-1}(\mathbf{x}_i)) = \frac{1}{n}$ *for all* $i = 1, \cdots, n$,
- $\delta_n \coloneqq esssup_{x \in \mathcal{M}} d_{\mathcal{M}}(T_n(x), x) \leq C \frac{\log(n)^{p_m}}{n^{1/m}}$,

*where* $C = C(\mathcal{M}, \alpha, \beta) > 0$ *and* $d_{\mathcal{M}}$ *represents the geodesic distance in* $\mathcal{M}$.

PROOF. This proposition is proved by Garcia Trillos, et al. (2020) ( [**88**]). □

In order to know more about this constant $C$, we introduce the following definition.

DEFINITION 7. *(WP property) We say that a manifold* $\mathcal{M}$ *satisfies the WP property with k polytopes if there exists a finite family of closed convex polytopes* $\{A_i\}_{i=1}^k$ *covering* $\mathcal{M}$ *and they satisfy that for all* $i, j = 1, \cdots, k$*:*

- $\text{int}\,(A_i) \cap \mathcal{M} \neq \varnothing$,
- *if* $i \neq j$ *then int* $(A_i) \cap \text{int}\,(A_j) = \varnothing$,
- $A_i \cap \overline{\mathcal{M}}$ *is bi-Lipschitz homeomorphic to a closed cube,*

*where* $\text{int}(\cdot)$ *denotes the interior of the inside set.*

REMARK 11. *Assume that the manifold* $\mathcal{M}$ *satisfies the WP property, then by the Theorem 1.2 in [**91**], there exists a bi-Lipschitz differentiable homeomorphism* $\psi : \mathcal{M} \to [0,1]^m$ *between* $\mathcal{M}$ *and the unit cube. Thus the constant* $C$ *in Proposition 8 satisfies* $C \propto Lip(\psi^{-1})\alpha\sqrt{m}\|\psi\|_{op}\det\big(J\psi^{-1}(y)\big)$, *where* $J\psi^{-1}$ *denotes the Jacobian matrix of* $\psi^{-1}$.

Given the map $T_n$, we define the discretization map

$$P : L^2(\nu) \to L^2(\nu_n)$$

as the transformation

$$Pf\,(\mathbf{x}_i) \coloneqq n \int_{T_n^{-1}(\{\mathbf{x}_i\})} f(x)\rho^q(x)dx, \quad i = 1, \ldots, n.$$

We are going to define the interpolation map $I$. Note that the notation $I$ is also being used to denote scaling constants in our mixture model, but it should be clear from the context which object is being considered. Since there is no ambiguity between them and the notation is consistently used in related literature, we just keep both two notations in this thesis. To state it more clearly, $I$ denotes the interpolation map (operator) when it is applied to a function $u_n \in L^2(\nu_n)$. Besides

this case, $I$ and its variants form ($I_{\min}$) with or without superscript and subscript always denote scaling constants for our mixture model.

In order to define the interpolation map $I$, we start from $P^*$, the adjoint of $P$ with respect to $\langle \cdot, \cdot \rangle_{L^2(\nu_n)}$, i.e. the map that satisfies

$$\langle Pg, f_n \rangle_{L^2(\nu_n)} = \langle g, P^* f_n \rangle_{L^2(\nu)}, \quad \forall g \in L^2(\nu), \quad \forall f_n \in L^2(\nu_n),$$

which gives the following relationship

$$P^* f_n(x) = f_n \circ T_n(x), x \in \mathcal{M}.$$

Given the radial kernel $\eta$, let

$$\psi(t) \coloneqq \int_t^\infty \eta(s) ds,$$

and define a smoothing operator

$$\Lambda_{\varepsilon,n,0} f(x) \coloneqq \int_{\mathcal{M}} \frac{1}{(\varepsilon - 2\delta_n)^m} \psi\left(\frac{d_{\mathcal{M}}(x,y)}{\varepsilon - 2\delta_n}\right) f(y) dy, \quad x \in \mathcal{M}, \quad f \in L^2(\nu),$$

and its normalized version

$$\Lambda_{\varepsilon,n,f} \coloneqq \frac{\Lambda_{\varepsilon,n,0}}{\Lambda_{\varepsilon,n,0}\mathbf{1}} f, f \in L^2(\nu).$$

With this, we define $I$ by the composition of $P^*$ with $\Lambda_{\varepsilon,n}$ as

$$I u_n \coloneqq \Lambda_{\varepsilon,n} \circ P^* u_n, u_n \in L^2(\nu_n).$$

We can first define an intermediate, non-local continuum Dirichlet energy

$$E_r(f) \coloneqq \int_{\mathcal{M}} \int_{\mathcal{M}} \eta\left(\frac{d_{\mathcal{M}}(x,y)}{r}\right) |f(x) - f(y)|^2 \rho^q(y) \rho^q(x) dx dy, \quad f \in L^2(\mathcal{M}),$$

where $r > 0$ is a length scale to be chosen later on.

Given that the kernel $\eta$ is assumed to be normalized, we can treat $\frac{\hat{d}_\varepsilon(x_i)}{n}$ as a kernel density estimator of the density $\rho$. Formally, we have that

$$\max_{i=1,\cdots,n} \left| \frac{1}{n} \hat{d}_\varepsilon(x_i) - \rho(x_i) \right| \leq C \left(\varepsilon + \frac{\delta_n}{\varepsilon}\right),$$

where $\delta_n$ is the $\infty$-optimal transportation (OT) distance between measures $\nu_n$ and $\nu$.

Combining this with Proposition 8, we have that for a given $\beta > 1$, there exists a constant $C_\beta > 0$ depending only on $\beta$ so that with probability at least $1 - C_\beta n^{-\beta}$, the following bound holds

$$\max_{i=1,\cdots,n} \left| \frac{1}{n} \hat{d}_\varepsilon(x_i) - \rho(x_i) \right| \leq C \left( \varepsilon + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} \right).$$

This is not the optimal estimate on the error of approximation of a kernel density estimator, but has the advantage of only depending on the $\infty$-OT distance between empirical and ground-truth measures.

The next two lemmas (Lemma 4 and Lemma 5) play the key roles in the 'closeness' of the population embedding case and the sample embedding case. These two lemmas are useful to show that the empirical eigenvalues and eigenvectors of weighted Laplacian matrix converge to the eigenvalues and eigenfunctions of weighted Laplacian operator, respectively, under some suitable conditions. The proofs of these two lemmas are shown later after more auxiliary results are shown.

LEMMA 4. *(Discretization and interpolation errors). Under Assumption 2 and Assumption 3, the following four results hold.*

*1. For every $f \in H_q^1(\mathcal{M}, \rho)$,*

$$\left| \|Pf\|^2_{L^2(\nu_n)} - \|f\|^2_{L^2(\nu)} \right| \leq \alpha^q \left( \|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho \right) (1 + 2q\alpha^q L_\rho \delta_n) \|f\|^2_{L^2(\nu)} + \tilde{C}' \delta_n \|f\|_{L^2(\nu)} D(f)^{\frac{1}{2}},$$

*where $\tilde{C}'$ has the form*

$$\tilde{C}' = \frac{C\alpha \left(1 + q\alpha^q L_\rho\right) \left(1 + mq\alpha^q L_\rho\right) m 2^{m/2} \sigma_\eta^{1/2}}{\sqrt{\eta(1/2)\omega_m}}$$

*for some universal constant $C > 0$.*

*2. For every $f \in H_q^1(\mathcal{M}, \rho)$,*

$$b_n(Pf) \leq \left( 1 + C_1' \varepsilon + C_2' \frac{\delta_n}{\varepsilon} + C_3' \varepsilon^2 \right) D(f),$$

*where the constants $C_1'$, $C_2'$, $C_3'$ can be written as*

$$C_1' = Cq\alpha^q L_\rho, \quad C_2' = C\left(m + \frac{2^{m+1}L_\eta\left(1+q\alpha^q L_\rho\right)}{\eta(1/2)}\right), \quad C_3' = Cm\left(K + \frac{1}{R^2}\right),$$

*where $C$ is a universal constant.*

*3. For every $u \in L^2(\nu_n)$,*

$$\left|\|Iu\|^2_{L^2(\nu)} - \|u\|^2_{L^2(\nu_n)}\right| \leq \tilde{C}''\varepsilon\|u\|_{L^2(\nu_n)} b_n(u)^{\frac{1}{2}} + 2\alpha\left(1+q\alpha^{2q-1}L_\rho\delta_n\right)\cdot\left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right)\|u\|^2_{L^2(\nu_n)},$$

*where $\tilde{C}'' = C\alpha\left(1+q\alpha^q L_\rho\right)\cdot\left(1+q\alpha^q L_\rho\right)\cdot\left(1+c''\right), \quad c'' = \frac{L_\eta 8^m\left(1+q\alpha^q L_\rho\right)^2}{\eta(1/2)}$.*

*4. For every $u \in L^2(\nu_n)$,*

$$D(Iu) \leq (1 + C_1''\varepsilon + C_2''\frac{\delta_n}{\varepsilon} + C_3''\varepsilon^2) b_n(u),$$

*where*

$$C_1'' = q\alpha^q L_\rho, \quad C_2'' = C\cdot 4^m\left(m + C_2'\right), \quad C_3'' = C\left(1+1/\sigma_\eta\right)mK.$$

Lemma 4 provides the error bounds of discretization and interpolation. Furthermore, notice that the quantity $\delta_n$ (defined in Proposition 8) can be arbitrarily small as long as the sample size $n$ is large enough, and thus by the following Lemma 6, $\|\mathbf{m}^q - \rho^q\|_\infty$ can also be arbitrarily small. As a result, Lemma 4 indicates that the discretization and interpolation errors can be arbitrarily small, which indicates the convergence for eigenvectors of the weighted Laplacian matrix towards the eigenfunctions of the weighted Laplacian operator. Similar analysis on the following Lemma 7 illustrates the convergence for eigenvalues of the weighted Laplacian matrix towards the eigenvalues of the weighted Laplacian operator.

Lemma 5. *For $i \in \mathbb{N}$, recall that $\lambda_{n,i}$ is the $i$-th eigenvalue of the empirical weighted Laplacian $\Delta_n$ and $\lambda_i$ is the $i$-th eigenvalue of the differential operator $\Delta_\rho$. Recall that $\delta_n$ is the $\infty-OT$ distance between measures $\nu_n$ and $\nu$ and assume that $h > 0$ satisfies Assumption 4. Then*

*1. (Upper bound) If $\delta_n$ and $\|\mathbf{m}^q - \rho^q\|_\infty$ satisfy*

$$\sqrt{\lambda_i}\delta_n + \|\mathbf{m}^q - \rho^q\|_\infty < c$$

*for a positive constant $c$ that depends only on $m, \alpha, L_\rho$, and $\eta$, then*

$$\frac{\lambda_{n,i} - \lambda_i}{\lambda_i} \leq \tilde{C}\left(L_\rho \varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right),$$

*where $\tilde{C}$ only depends on $m, \alpha, L_\rho$, and $\eta$.*

*2. (Lower bound) If $\varepsilon$ and $\|\mathbf{m}^q - \rho^q\|_\infty$ satisfy*

$$\sqrt{\lambda_i}\varepsilon + \|\mathbf{m}^q - \rho^q\|_\infty < c$$

*for a positive constant $c$ that depends only on $m, \alpha, L_\rho$ and $\eta$, then*

$$\frac{\lambda_{n,i} - \lambda_i}{\lambda_i} \geq -\tilde{C}\left(L_\rho \varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right),$$

*where $\tilde{C}$ only depends on $m, q, \alpha, L_\rho$, and $\eta$.*

To make notation convenience, we denote $T_n^{-1}(\{\mathbf{x}_i\})$ by $U_i$, and obviously $\nu(U_i) = \frac{1}{n}$ for all $i = 1, \cdots, n$.

LEMMA 6. *Consider $\eta : \mathbb{R} \to \mathbb{R}$, nonincreasing, supported on $[0,1]$, and normalized: $\int_{\mathbb{R}^m} \eta(|x|)dx = 1$. Consider $\varepsilon > 0$ satisfying Assumption 4. Then there exists a universal constant $C > 0$ such that*

$$\|\mathbf{m}^q - \rho^q\|_\infty \coloneqq \max_{i=1,\ldots,n} |m_i^q - \rho(\mathbf{x}_i)^q| \leq C\alpha^q L_\rho^q \varepsilon^q + C\alpha^q \eta(0)^q m^q \omega_m^q \frac{\delta_n^q}{\varepsilon^q} + C\alpha^q m^q \left(K + \frac{1}{R^2}\right)^q \varepsilon^{2q},$$

*where the weights $m_i = \frac{1}{n\varepsilon^m}\sum_{j=1}^n \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right), \quad i = 1, \ldots, n.$*

PROOF. For every $i, j$, if $|x_i - x_j| \leq \varepsilon$, then $|x_i - x_j| \leq \frac{R}{2}$, thus we have

$$d(x_i, x_j) \leq |x_i - x_j| + \frac{8}{R^2}|x_i - x_j|^3 \leq \left(1 + \frac{8\varepsilon^2}{R^2}\right)|x_i - x_j|.$$

Thus for every $i, j$ and every $y \in U_j$,

$$\eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) \leq \eta\left(\frac{d(x_i, x_j)}{\hat{\varepsilon}}\right) \leq \eta\left(\frac{(d(x_i, y) - \delta_n)_+}{\hat{\varepsilon}}\right),$$

where $\hat{\varepsilon} \coloneqq \varepsilon + \frac{27\varepsilon^3}{R^2}$.

Then we can bound $m_i$ as

$$m_i = \frac{1}{n\varepsilon^m}\sum_{j=1}^{n}\eta\left(\frac{|x_i-x_j|}{\varepsilon}\right) \le \frac{1}{\varepsilon^m}\int_{\mathcal{M}}\eta\left(\frac{(d(x_i,y)-\delta_n)_+}{\hat{\varepsilon}}\right)p(y)dVol(y) \le (p(x_i)+10L_\rho\varepsilon)\frac{1}{\varepsilon^m}\int_{\mathcal{M}}\eta\left(\frac{(d(x_i,y)-\delta_n)_+}{\hat{\varepsilon}}\right)dVol(y).$$

Also, we have

$$\frac{1}{\varepsilon^m}\int_{\mathcal{M}}\eta\left(\frac{(d(x_i,y)-\delta_n)_+}{\hat{\varepsilon}}\right)dVol(y) = \frac{1}{\varepsilon^m}\int_{B(\hat{\varepsilon}+\delta_n)}\eta\left(\frac{(|z|-\delta_n)_+}{\hat{\varepsilon}}\right)J_{x_i}(z)dz \le (1+CmK\varepsilon^2)\frac{1}{\varepsilon^m}\int_{B(\hat{\varepsilon}+\delta_n)}\eta\left(\frac{(|z|-\delta_n)_+}{\hat{\varepsilon}}\right)dz,$$

where the last integral can be estimated as

$$\begin{aligned}
\frac{1}{\varepsilon^m}\int_{B(\hat{\varepsilon}+\delta_n)}\eta\left(\frac{(|z|-\delta_n)_+}{\hat{\varepsilon}}\right)\mathrm{d}z =& \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m} + \frac{1}{\varepsilon^m}\int_{b_n(\hat{\varepsilon}+\delta_n)\setminus B(\delta_n)}\eta\left(\frac{|z|-\delta_n}{\hat{\varepsilon}}\right)\mathrm{d}z \\
&= \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m} + \frac{\hat{\varepsilon}^m}{\varepsilon^m}\int_0^1 m\omega_m\left(r+\frac{\delta_n}{\hat{\varepsilon}}\right)^{m-1}\eta(r)\mathrm{d}r \\
&\le \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m} + \left(1+\frac{16m\varepsilon^2}{R^2}\right)\int_0^1 m\omega_m\left(r+\frac{\delta_n}{\varepsilon}\right)^{m-1}\eta(r)\mathrm{d}r.
\end{aligned}$$

Also, by using the binomial theorem, we have

$$\begin{aligned}
m\omega_m\int_0^1\left(r+\frac{\delta_n}{\varepsilon}\right)^{m-1}\eta(r)\mathrm{d}r &\le m\omega_m\int_0^1 r^{m-1}\eta(r)\mathrm{d}r + m\omega_m\eta(0)\sum_{k=1}^{m-1}\binom{m-1}{k}\left(\frac{\delta_n}{\varepsilon}\right)^k\frac{1}{m-k} \\
&= 1+\omega_m\eta(0)\sum_{k=1}^{m-1}\binom{m}{k}\left(\frac{\delta_n}{\varepsilon}\right)^k \\
&= 1+\omega_m\eta(0)\left(\left(1+\frac{\delta_n}{\varepsilon}\right)^m - 1 - \frac{\delta_n^m}{\varepsilon^m}\right) \\
&\le 1+2m\eta(0)\omega_m\frac{\delta_n}{\varepsilon} - \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m}.
\end{aligned}$$

Combining all above equations, we have

$$m_i \le (\rho(x_i)+10L_\rho\varepsilon)(1+C_mK\varepsilon^2)\left(\eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m} + \left(1+\frac{16m\varepsilon^2}{R^2}\right)\right)\left(1+2m\eta(0)\omega_m\frac{\delta_n}{\varepsilon} - \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m}\right).$$

Thus we also have

$$m_i^q \le \rho(x_i)^q\left(1+\frac{10L_\rho\varepsilon}{p(x_i)}\right)^q(1+C_mK\varepsilon^2)^q\left(\eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m} + \left(1+\frac{16m\varepsilon^2}{R^2}\right)\right)^q\left(1+2m\eta(0)\omega_m\frac{\delta_n}{\varepsilon} - \eta(0)\omega_m\frac{\delta_n^m}{\varepsilon^m}\right)^q$$

and

$$m_i^q - \rho(x_i)^q \le C\alpha^qL_\rho^q\varepsilon^q + C\alpha^q\eta(0)^qm^q\omega_m^q\frac{\delta_n^q}{\varepsilon^q} + C\alpha^qm^q\left(K+\frac{1}{R^2}\right)^q\varepsilon^{2q}$$

for an absolute constant $C>0$.

By similar steps, we can also find an upper bound for $\rho(x_i)^q - m_i^q$. Combining them together, we have

$$\max_{i=1,\ldots,n}|m_i^q - \rho(\mathbf{x}_i)^q| \le C\alpha^qL_\rho^q\varepsilon^q + C\alpha^q\eta(0)^qm^q\omega_m^q\frac{\delta_n^q}{\varepsilon^q} + C\alpha^qm^q\left(K+\frac{1}{R^2}\right)^q\varepsilon^{2q}.$$

□

LEMMA 7. *(Convergence rate for eigenvalues). Suppose $\varepsilon$ satisfies Assumption 4. Let $\lambda_i$ be the i-th eigenvalue of $\Delta_\rho$ and let $\lambda_{n,i}$ be the i-th eigenvalue of $\Delta_n$. Let $\beta > 1$, then there exist constants $C, C_\beta > 0$ such that for sufficiently large $n$, with probability at least $1 - C_\beta n^{-\beta}$, we have*

$$|\lambda_{n,i} - \lambda_i| \leq C \left( L_\rho \varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty \right) \lambda_i,$$

*where $C$ only depends on $\mathcal{M}, \beta, m, \alpha, L_\rho, L_\rho$, and $\eta$.*

PROOF. By Proposition 8, for a given $\beta > 1$, there exists a constant $C_\beta > 0$ depending only on $\beta$ so that with probability at least $1 - C_\beta n^{-\beta}$, $\delta_n \leq C\frac{\log(n)^{p_m}}{n^{1/m}}$. By this result and Lemma 18, the condition of Lemma 5 holds and the conclusion is attained by just multiple $\lambda_i$ on both two sides of the inequalities. □

Some useful lemmas are introduced in the following parts, and the proofs of these lemmas (Lemma 8, Lemma 9, Lemma 10, Lemma 11, Lemma 12, Lemma 13, and Lemma 14) can be found in [**88**].

LEMMA 8. *Suppose $\varepsilon$ satisfies Assumption 4. Then there exists a universal constant $C > 0$ such that for every $0 < r < 2\varepsilon$ and every $f \in L^2(\nu)$*

$$E_r(f) \leq C2^m(1 + q\alpha^q L_\rho)E_{r/2}(f).$$

LEMMA 9. *Suppose $\varepsilon$ satisfies Assumption 4. Then there exists a universal constant $C > 0$ such that*

$$E_r(f) \leq (1 + L_\rho q\alpha^q r)(1 + CmKr^2)\sigma_\eta r^{m+2} D(f),$$

*for every $f \in H_q^1(\mathcal{M}, \rho)$ and $0 < r < 2\varepsilon$.*

LEMMA 10. *Suppose $\varepsilon$ satisfies Assumption 4. Let $\delta_n < r < 2\varepsilon$, $f \in L^2(\nu)$ and $V \subseteq \mathcal{M}$ a Borel set such that $\nu(V) > 0$ and diam$(V) \leq 2\delta_n$. Then we have*

$$\int_V \left| f(x) - \frac{1}{\nu(V)} \int_V f d\nu \right|^2 d\nu(x) \leq \frac{2\left(1 + CmKr^2\right)}{\eta(1/2)\omega_m(r - \delta_n)^m} E_{2r}(f, V).$$

*For every $r > 0$, define the operator $\Lambda_r^0$ by*

$$(\Lambda_r^0 f)(x) \coloneqq \int_{\mathcal{M}} f(y)k_r(x, y)dVol(y),$$

*where*

$$k_r(x,y) \coloneqq \frac{1}{r^m}\psi(\frac{d_{\mathcal{M}}(x,y)}{r}).$$

*Also define the smoothing operator* $\Lambda_r$ *as*

$$\Lambda_r f(x) \coloneqq (\theta(x))^{-1}\Lambda_r^0 f(x),$$

*where* $\theta \coloneqq \Lambda_r^0 \mathbf{1}$.

LEMMA 11. *There exists an absolute constant* $C > 0$ *such that*

$$(1 + CmKr^2)^{-1} \leq \theta(x) \leq 1 + CmKr^2.$$

LEMMA 12. *Suppose that* $h$ *satisfies Assumption 4. Then there exists a universal constant* $C > 0$ *such that*

$$\|\Lambda_r f\|^2_{L^2(\nu)} \leq (1 + q\alpha^q L_\rho r)\,(1 + q\alpha^q L_\rho r)\left(1 + CmKr^2\right)\|f\|^2_{L^2(\nu)}$$

*and*

$$\|\Lambda_r f - f\|^2_{L^2(\nu)} \leq \frac{C\alpha^2}{\sigma_\eta r^m} E_r(f)$$

*for all* $f \in L^2(\nu)$ *and all* $r < 2\varepsilon$.

LEMMA 13. *Suppose that* $\varepsilon$ *satisfies Assumption 4. Then there exists a universal constant* $C > 0$ *such that*

$$D\left(\Lambda_r f\right) \leq (1 + q\alpha^q L_\rho r) \cdot \left(1 + C\left(1 + 1/\sigma_\eta\right) mKr^2\right) \frac{1}{\sigma_\eta r^{m+2}} E_r(f)$$

*for all* $f \in L^2(\nu)$ *and all* $r < 2\varepsilon$.

LEMMA 14. *Assume the support of* $\eta$ *is contained in* $[0,1]$ *and* $\eta$ *is Lipschitz in* $[0,1]$. *Then for all* $r, s > 0$ *and* $t \geq 0$ *we have*

- $\eta(\frac{t}{r+s}) \leq \eta(\frac{(t-s)_+}{r}) \leq \eta(\frac{t}{r+s}) + L_\eta \frac{s}{r}\mathbf{1}_{\{t \geq r+s\}}$,
- $\eta(\frac{t+s}{r}) \geq \eta(\frac{t}{r-s}) - L_\eta \frac{s}{r}\mathbf{1}_{\{t \geq r-s\}}$ *provided that* $s < r$,

*where* $L_\eta > 0$ *denotes the Lipschitz constant of* $\eta$ *restricted to* $[0,1]$.

LEMMA 15. *For all $u \in L^2(\nu_n)$ and $f \in L^2(\nu)$, we have*

$$\left|\langle P^*u, f\rangle_{L^2(\nu)} - \langle u, Pf\rangle_{L^2(\nu_n)}\right| \le \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right) \langle P^*|u|, |f|\rangle_{L^2(\nu)}$$

*and*

$$\left|\|P^*u\|^2_{L^2(\nu)} - \|u\|^2_{L^2(\nu_n)}\right| \le \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right) \|P^*u\|^2_{L^2(\nu)}.$$

*In addition, if $\alpha^q\|\mathbf{m}^q - \rho^q\|_\infty \le \frac{1}{2}$, then $\forall u \in L^2(\nu_n)$,*

$$\|P^*u\|^2_{L^2(\nu)} \le 2\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right)\|u\|^2_{L^2(\nu_n)}$$

*for some universal constant $C > 0$.*

PROOF. By the definition of $P$ and $P^*$, we have that

$$\begin{aligned}
\left|\langle u, Pf\rangle_{L^2(\nu_n)} - \langle P^*u, f\rangle_{L^2(\nu)}\right| &= \left|\sum_{i=1}^n \frac{m_i^q}{n}u(x_i)\cdot n\int_{U_i} f dx - \int_{\mathcal{M}}\sum_{i=1}^n u(x_i)\mathbf{1}_{U_i} f\rho^q dx\right| \\
&\le \int_{\mathcal{M}}\sum_{i=1}^n |u(x_i)|\mathbf{1}_{U_i}|f(x)|\cdot|m_i^q - \rho^q(x_i) + \rho^q(x_i) - \rho^q(x)|\,dx \\
&\le \alpha^q\left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right)\langle P^*|u|, |f|\rangle_{L^2(\nu)}.
\end{aligned}$$

Also,

$$\begin{aligned}
\left|\|P^*u\|^2_{L^2(\nu)} - \|u\|^2_{L^2(\nu_n)}\right| &= \left|\sum_{i=1}^n\left(\int_{U_i} u^2(x_i)\rho^q dx - \int_{U_i}\frac{m_i^q}{n}u^2(x_i)dx\right)\right| \\
&\le \alpha^q\left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right)\|P^*u\|^2_{L^2(\nu)}.
\end{aligned}$$

The last part of this lemma is obtained by following that

$$\|P^*u\|^2_{L^2(\nu)} = \sum_{i=1}^n u(x_i)^2\int_{U_i}\rho^q(y)dy \le \frac{2\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right)}{n}\sum_{i=1}^n u(x_i)^2 m_i^q = 2\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right)\|u\|^2_{L^2(\nu_n)}.$$

□

LEMMA 16. *For every $f \in L^2(\nu)$, we have*

$$\|P^*Pf\|^2_{L^2(\nu)} \le (1 + 2q\alpha^q L_\rho\delta_n)\|f\|^2_{L^2(\nu)}.$$

*And there exists a universal constant $C > 0$ such that*

$$\|f - P^*Pf\|_{L^2(\nu)} \leq \frac{C\left(1 + mq\alpha^q L_\rho \delta_n\right) m 2^{m/2} \sigma_\eta^{1/2}}{\sqrt{\eta(1/2)\omega_m}} \delta_n D(f)^{\frac{1}{2}}$$

*for all $f \in H_q^1(\mathcal{M}, \rho)$.*

PROOF. By Jensen's inequality, we have

$$\begin{aligned}
\int_{\mathcal{M}} \left(P^*Pf(x)\right)^2 \frac{\rho^q(x)}{I} dx &\leq \sum_{i=1}^n \int_{U_i} \int_{U_i} nf(y)^2 \frac{\rho^q(x)}{I} dy dx \\
&\leq \left(1 + 2q\alpha^q L_\rho \delta_n\right) \sum_{i=1}^n \int_{U_i} \int_{U_i} nf(y)^2 \frac{\rho^q(y)}{I} dy dx \\
&= \left(1 + 2q\alpha^q L_\rho \delta_n\right) \int_{\mathcal{M}} f(y)^2 \frac{\rho^q(y)}{I} dy.
\end{aligned}$$

Also,

$$\|f - P^*Pf\|_{L^2(\nu)}^2 \leq \frac{2\left(1 + CmKr^2\right)}{\eta(1/2)\omega_m (r - \delta_n)^m} E_{2r}(f) \leq \frac{C\left(1 + 2q\alpha^q L_\rho r\right) 2^m \sigma_\eta}{\eta(1/2)\omega_m} \frac{r^m}{(r - \delta_n)^m} r^2 D(f)$$

for any $r \in (\delta_n, 2\varepsilon)$. By choosing $r = (m+1)\delta_n$, $\frac{r^m}{(r-\delta_n)^m}$ is bounded by a constant and the assertion thus follows. □

Combining the previous auxiliary results, we first prove parts 1 and 2 of Lemma 4 and use it to prove the part 1 of Lemma 5, then prove the parts 3 and 4 of Lemma 4 and use it to prove the part 2 of Lemma 5.

PROOF. **Proofs of Lemma 4, parts 1 and 2**

1. From Lemma 15, we know $P^*$ is almost an isometry. Thus

$$\begin{aligned}
\left|\|Pf\|_{L^2(\nu_n)}^2 - \|f\|_{L^2(\nu)}^2\right| &\leq \left|\|Pf\|_{L^2(\nu_n)}^2 - \|P^*Pf\|_{L^2(\nu)}^2\right| + \left|\|P^*Pf\|_{L^2(\nu)}^2 - \|f\|_{L^2(\nu)}^2\right| \\
&\leq \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right) \|P^*Pf\|_{L^2(\nu)}^2 \\
&\quad + \left(\|P^*Pf\|_{L^2(\nu)} + \|f\|_{L^2(\nu)}\right) \|P^*Pf - f\|_{L^2(\nu)} \\
&\leq \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right) \left(1 + 2q\alpha^q L_\rho \delta_n\right) \|f\|_{L^2(\nu)}^2 \\
&\quad + \frac{C\alpha\left(2 + q\alpha^q L_\rho \delta_n\right)\left(1 + mq\alpha^q L_\rho \delta_n\right) m 2^{m/2} \sigma_\eta^{1/2}}{\sqrt{\eta(1/2)\omega_m}} \delta_n \|f\|_{L^2(\nu)} D(f)^{\frac{1}{2}}.
\end{aligned}$$

2. Notice that

$$|Pf(x_j) - Pf(x_i)|^2 \leq \frac{n^2}{I} \int_{U_i} \int_{U_j} |f(y) - f(x)|^2 \rho^q(x) \rho^q(y) dy dx.$$

Let $\hat{\varepsilon} \coloneqq (1 + \frac{27}{R^2}\varepsilon^2)\varepsilon$, then we have

$$\begin{aligned}
b_n(Pf) &\leq \frac{1}{\sigma_\eta \varepsilon^{m+2}} \sum_i \sum_j \int_{U_i} \int_{U_j} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) |f(y) - f(x)|^2 \rho^q(y) dy \rho^q(x) dx \\
&\leq \frac{1}{\sigma_\eta \varepsilon^{m+2}} \sum_i \sum_j \int_{U_i} \int_{U_j} \eta\left(\frac{d(x_i, x_j)}{\hat{\varepsilon}}\right) |f(y) - f(x)|^2 \rho^q(y) dy \rho^q(x) dx \\
&\leq \frac{1}{\sigma_\eta \varepsilon^{m+2}} \int_{\mathcal{M}} \int_{\mathcal{M}} \eta\left(\frac{(d_{\mathcal{M}}(x, y) - 2\delta_n)_+}{\hat{\varepsilon}}\right) |f(y) - f(x)|^2 \rho^q(y) dy \rho^q(x) dx \\
&\leq \frac{1}{\sigma_\eta \varepsilon^{m+2}} \int_{\mathcal{M}} \int_{\mathcal{M}} \left(\eta\left(\frac{d_{\mathcal{M}}(x, y)}{\hat{\varepsilon} + 2\delta_n}\right) + 2L_\eta \frac{\delta_n}{\hat{\varepsilon}} \mathbf{1}_{B_{\mathcal{M}}(x, \hat{\varepsilon} + 2\delta_n)}(y)\right) |f(y) - f(x)|^2 \rho^q(y) dy \rho^q(x) dx \\
&= \frac{1}{\sigma_\eta \varepsilon^{m+2}} \left(E_{\hat{\varepsilon} + 2\delta_n}(f) + \frac{2L_\eta}{\eta(1/2)} \frac{\delta_n}{\varepsilon} E_{2(\hat{\varepsilon} + 2\delta_n)}(f)\right).
\end{aligned}$$

In addition,

$$\begin{aligned}
\frac{1}{\sigma_\eta \varepsilon^{m+2}} E_{\hat{\varepsilon} + 2\delta_n}(f) &\leq (1 + Cq\alpha^q L_\rho \varepsilon)\left(1 + CmK\varepsilon^2\right)\left(1 + \frac{27\varepsilon^2}{R^2} + 2\frac{\delta_n}{\varepsilon}\right)^{m+2} D(f) \\
&\leq (1 + Cq\alpha^q L_\rho \varepsilon)\left(1 + CmK\varepsilon^2\right)\left(1 + Cm\frac{\varepsilon^2}{R^2} + Cm\frac{\delta_n}{\varepsilon}\right) D(f),
\end{aligned}$$

and

$$\frac{1}{\sigma_\eta \varepsilon^{m+2}} \frac{2L_\eta}{\eta(1/2)} \frac{\delta_n}{\varepsilon} E_{2(\hat{\varepsilon} + 2\delta_n)}(f) \leq \frac{2^{m+1} L_\eta}{\eta(1/2)} (1 + Cq\alpha^q L_\rho \varepsilon)\left(1 + CmK\varepsilon^2\right)\left(1 + Cm\frac{\varepsilon^2}{R^2} + Cm\frac{\delta_n}{\varepsilon}\right) \frac{\delta_n}{\varepsilon} D(f).$$

□

Proof. **Proof of Lemma 5, part 1** For a fixed non-negative integer $i$. By the minimax principle, we have

$$\lambda_{n,i} \leq \sup_{u \in L \setminus \{0\}} \frac{b_n(u)}{\|u\|^2_{L^2(\nu_n)}}$$

holds for every $i$-dimensional subspace $L \subseteq L^2(\nu_n)$. We denote by $W$ the span of the first $i$ orthonormal eigenfunctions of $\Delta_\rho$. Set $L \coloneqq P(W)$, then for every $f \in W$, by the Courant minimax principle, we have

$$D(f) \leq \lambda_N \|f\|^2_{\rho^q}.$$

Combining this with the part 1 of Lemma 4, we have

$$\|Pf\|^2_{L^2(\nu_n)} \geq \left(1 - \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right)(1 + 2q\alpha^q L_\rho \delta_n) - \tilde{C}' \sqrt{\lambda_i}\delta_n\right) \|f\|^2_{L^2(\nu)}.$$

So if the condition $\alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right)(1 + 2q\alpha^q L_\rho \delta_n) + \tilde{C}' \sqrt{\lambda_i}\delta_n \leq \frac{1}{2}$ holds, then $P$ is injective on $W$. Thus $\dim(L) = i$ and by applying part 2 of Lemma 4 to $u = Pf \in L$, we have

$$\begin{aligned}
&\frac{b_n(u)}{\|u\|^2_{L^2(\nu_n)}} \leq \frac{\left(1 + C_1'\varepsilon + C_2'\frac{\delta_n}{\varepsilon} + C_3'\varepsilon^2\right)}{1 - \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right)(1 + 2q\alpha^q L_\rho \delta_n) - \tilde{C}' \sqrt{\lambda_i}\delta_n} \lambda_i \\
&\leq \left(1 + C_1'\varepsilon + C_2'\frac{\delta_n}{\varepsilon} + C_3'\varepsilon^2 + \frac{\alpha C}{I}\left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right)(1 + 2q\alpha^q L_\rho \delta_n) + \tilde{C}' \sqrt{\lambda_i}\delta_n\right) \lambda_i.
\end{aligned}$$

The above inequality holds for all $u = Pf$ for $f \in W$, so

$$\frac{\lambda_{n,i} - \lambda_i}{\lambda_i} \leq \tilde{C}\left(L_\rho \varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right).$$

□

PROOF. **Proofs of Lemma 4, parts 3 and 4**

$$\begin{aligned}
&\left|\|Iu\|^2_{L^2(\nu)} - \|u\|^2_{L^2(\nu_n)}\right| \\
&\quad \leq \left|\|Iu\|^2_{L^2(\nu)} - \|P^*u\|^2_{L^2(\nu)}\right| + \left|\|P^*u\|^2_{L^2(\nu)} - \|u\|^2_{L^2(\nu_n)}\right| \\
&\quad \leq \left(\|Iu\|_{L^2(\nu)} + \|P^*u\|_{L^2(\nu)}\right)\|Iu - P^*u\|_{L^2(\nu)} + \alpha^q \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q \alpha^{q-1} L_\rho\right)\|P^*u\|^2_{L^2(\nu)}.
\end{aligned}$$

By using Lemma 12, we have

$$\|Iu - P^*u\|^2_{L^2(\nu)} = \|\Lambda_{\varepsilon - 2\delta_n} P^*u - P^*u\|^2 \leq \frac{C\alpha^2}{\sigma_\eta \varepsilon^m} E_{\varepsilon - 2\delta_n}(P^*u)$$

for some universal constant $C > 0$.

Pick a kernel $\tilde{\eta} = \mathbf{1}_{[0,1]}$ and let $\tilde{b}$ and $\tilde{E}$ denote the discrete Dirichlet form and the energy $E$ when using the kernel $\tilde{\eta}$ and $b_\varepsilon$ denote the forms $b$ with bandwidth $\varepsilon$. (Except that $b_n$ denotes Dirichlet form associated to $\Delta_n$.) Then

$$
\begin{aligned}
\tilde{b}_\varepsilon(u) &= \frac{1}{\sigma_{\tilde{\eta}}\varepsilon^{m+2}}\frac{1}{n^2}\sum_i\sum_j \tilde{\eta}\left(\frac{|x_i-x_j|}{\varepsilon}\right)|u(x_i)-u(x_j)|^2 \\
&= \frac{1}{I^2\sigma_{\tilde{\eta}}\varepsilon^{m+2}}\sum_{i,j}\int_{U_i}\int_{U_j}\tilde{\eta}\left(\frac{|T(x)-T(y)|}{\varepsilon}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&\geq \frac{1}{I^2\sigma_{\tilde{\eta}}\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\tilde{\eta}\left(\frac{d(T(x),T(y))}{\varepsilon}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&\geq \frac{1}{I^2\sigma_{\tilde{\eta}}\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\tilde{\eta}\left(\frac{d_{\mathcal{M}}(x,y)}{\varepsilon-2\delta_n}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&= \frac{1}{I^2\sigma_{\tilde{\eta}}\varepsilon^{m+2}}\tilde{E}_{\varepsilon-2\delta_n}(P^*u) \\
&= \frac{m+2}{I^2w_m\varepsilon^{m+2}}\tilde{E}_{\varepsilon-2\delta_n}(P^*u).
\end{aligned}
$$

Recall that $\eta$ is decreasing and thus $\eta(t)\geq\eta(\frac{1}{2})>0$ for all $t\in[0,\frac{1}{2})$, so

$$
\tilde{b}_{\varepsilon/2}(u)\leq\frac{\sigma_\eta(m+2)2^{m+2}}{\eta(1/2)\omega_m}b_\varepsilon(u).
$$

On the other hand, we have

$$
\begin{aligned}
b_\varepsilon(u) &\geq \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\eta\left(\frac{d(T(x),T(y))}{\varepsilon}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&\geq \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\eta\left(\frac{d_{\mathcal{M}}(x,y)+2\delta_n}{\varepsilon}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&\geq \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\eta\left(\frac{d_{\mathcal{M}}(x,y)}{\varepsilon-2\delta_n}\right)|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&- \frac{L_\eta}{I^2\sigma_\eta}\frac{\delta_n}{\varepsilon}\frac{1}{\varepsilon^{m+2}}\int_{\mathcal{M}}\int_{\mathcal{M}}\mathbf{1}_{\{d_{\mathcal{M}}(x,y)\leq\varepsilon-2\delta_n\}}|(P^*u)(x)-(P^*u)(y)|^2\rho^q(y)dy\rho^q(x)dx \\
&= \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}(P^*u)-\frac{L_\eta}{I^2\sigma_\eta}\frac{\delta_n}{\varepsilon}\frac{1}{\varepsilon^{m+2}}\tilde{E}_{\varepsilon-2\delta_n}(P^*u) \\
&\geq \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}(P^*u)-\frac{CL_\eta4^m(1+q\alpha^qL_\rho)^2}{I^2\sigma_\eta}\frac{\delta_n}{\varepsilon}\frac{1}{\varepsilon^{m+2}}\tilde{E}_{\frac{\varepsilon}{2}-2\delta_n}(P^*u) \\
&\geq \frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}(P^*u)-\frac{CL_\eta4^m\omega_m(1+q\alpha^qL_\rho)^2}{I^2(m+2)\sigma_\eta}\frac{\delta_n}{\varepsilon}\tilde{b}_{\frac{\varepsilon}{2}}(u).
\end{aligned}
$$

Combining the above inequalities together, we have

$$
\left(1+\frac{CL_\eta8^m(1+q\alpha^qL_\rho)^2}{I^2\eta(1/2)}\frac{\delta_n}{\varepsilon}\right)b_\varepsilon(u)\geq\frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}(P^*u),
$$

equivalantly,

$$E_{\varepsilon-2\delta_n}\left(P^*u\right) \leq \left(1+\frac{CL_\eta 8^m\omega_m^2\left(1+q\alpha^qL_\rho\right)^2}{I^2\eta(1/2)(m+2)^2}\frac{\delta_n}{\varepsilon}\right) I^2\sigma_\eta\varepsilon^{m+2}b_n(u),$$

and thus

$$\|Iu-P^*u\|^2 \leq \frac{C\alpha^2}{\sigma_\eta\varepsilon^m}E_{\varepsilon-2\delta_n}\left(P^*u\right) \leq C\alpha^2\left(1+\frac{L_\eta 8^m\left(1+q\alpha^qL_\rho\right)^2}{\eta(1/2)}\frac{\delta_n}{\varepsilon}\right)\varepsilon^2 b_n(u).$$

Recall that from Lemma 15, we have

$$\|P^*u\|^2_{L^2(\nu)} \leq 2\left(1+q\alpha^{2q-1}L_\rho\delta_n\right)\|u\|^2_{L^2(\nu_n)}.$$

ALso, by Lemma 8,

$$\begin{aligned}\|Iu\|_{L^2(\nu)} &= \|\Lambda_{\varepsilon-2\delta_n}P^*u\|_{L^2(\nu)}\\ &\leq C\left(1+q\alpha^qL_\rho\varepsilon\right)^{1/2}\cdot\left(1+q\alpha^qL_\rho\varepsilon\right)^{1/2}\|P^*u\|_{L^2(\nu)}\\ &\leq C\left(1+q\alpha^qL_\rho\varepsilon\right)\cdot\left(1+q\alpha^qL_\rho\varepsilon\right)\|u\|_{L^2(\nu_n)}.\end{aligned}$$

Combining all these inequalities to the first one, we get the desired bound for assertion 1.

For assertion 2, by using Lemma 13, we have

$$\begin{aligned}D(Iu) &\leq \left(1+q\alpha^qL_\rho\varepsilon\right)\cdot\left(1+C\left(1+\frac{1}{\sigma_\eta}\right)mK\varepsilon^2\right)\frac{1}{I^2\sigma_\eta(\varepsilon-2\delta_n)^{m+2}}E_{\varepsilon-2\delta_n}\left(P^*u\right)\\ &\leq \left(1+q\alpha^qL_\rho\varepsilon\right)\cdot\left(1+C\left(1+\frac{1}{\sigma_\eta}\right)mK\varepsilon^2\right)\left(1+Cm\frac{\delta_n}{\varepsilon}\right)\frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}\left(P^*u\right)\\ &\leq \left(1+q\alpha^qL_\rho\varepsilon+C\left(1+\frac{1}{\sigma_\eta}\right)mK\varepsilon^2+Cm\frac{\delta_n}{\varepsilon}\right)\frac{1}{I^2\sigma_\eta\varepsilon^{m+2}}E_{\varepsilon-2\delta_n}\left(P^*u\right)\\ &\leq \left(1+q\alpha^qL_\rho\varepsilon+C\left(1+\frac{1}{\sigma_\eta}\right)mK\varepsilon^2+Cm\frac{\delta_n}{\varepsilon}\right)\left(1+\frac{CL_\eta 8^m\omega_m^2\left(1+q\alpha^qL_\rho\right)^2}{I^2\eta(1/2)(m+2)^2}\frac{\delta_n}{\varepsilon}\right)b_n(u).\end{aligned}$$

□

PROOF. **Proof of Lemma 5, part 2** For a fixed non-negative integer $i$. By the minimax principle, we have

$$\lambda_i \leq \sup_{f\in L\setminus\{0\}}\frac{D(f)}{\|f\|^2_{L^2(\nu)}}$$

holds for every $i$-dimensional subspace $L \subseteq H^q(\mathcal{M})$. Denote by $W$ as the span of the first $i$ orthonormal eigenfunctions of $\Delta_n$. Set $L \coloneqq I(W)$, then for every $u \in W$, by the Courant minimax principle, we have

$$b_n(u) \leq \lambda_{n,i} \|u\|^2_{L^2(\nu_n)}.$$

Combining this with the part 3 of Lemma 4, we have

$$\|Iu\|^2_{L^2(\nu)} \geq \left(1 - 2\alpha\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right) \cdot \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right) - \tilde{C}''\sqrt{\lambda_{n,i}}\varepsilon\right)\|u\|^2_{L^2(\nu_n)}.$$

So if the condition $2\alpha\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right) \cdot \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right) + \tilde{C}''\sqrt{\lambda_{n,i}}\varepsilon \leq \frac{1}{2}$ holds, then $I$ is injective on $W$. Thus $\dim(L) = i$ and by applying part 4 of Lemma 4 to $u = Pf \in L$, we have

$$\begin{aligned}
\frac{D(f)}{\|f\|^2_{L^2(\nu)}} &\leq \frac{\left(1 + C_1''\varepsilon + C_2''\frac{\delta_n}{\varepsilon} + C_3''\varepsilon^2\right)}{1 - \frac{2\alpha}{I} \cdot \left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right)\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right) - \tilde{C}''\sqrt{\lambda_{n,i}}\varepsilon}\lambda_{n,i} \\
&\leq \left(1 + C_1''\varepsilon + C_2''\frac{\delta_n}{\varepsilon} + C_3''\varepsilon^2 + \frac{\alpha C}{I}\left(\|\mathbf{m}^q - \rho^q\|_\infty + \delta_n q\alpha^{q-1}L_\rho\right)\left(1 + q\alpha^{2q-1}L_\rho\delta_n\right) + \tilde{C}''\sqrt{\lambda_{n,i}}\varepsilon\right)\lambda_{n,i}.
\end{aligned}$$

The above inequality holds for all $f = Iu$ for $u \in U_n$, so

$$\frac{\lambda_{n,i} - \lambda_i}{\lambda_i} \geq -\tilde{C}\left(L_\rho\varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right).$$

□

**Proof of Theorem 9**.

Recall that $u_{n,1}, u_{n,2}, \cdots, u_{n,N}$ are the unit eigenvectors corresponding to the $N$ smallest eigenvalues of $\Delta_n$. They form an orthonormal basis with respect to $\langle\cdot,\cdot\rangle_{L^2(\nu)}$.

From Lemma 4, Lemma 7 in previous parts and Lemma 7.3 of [**21**], we have

$$\|Iu_{n,j} - \Pi_N(Iu_{n,j})\|^2_{L^2(\nu)} \leq \frac{C_{\mathcal{M},N}\lambda_N}{\lambda_{N+1} - \lambda_N}\left(\varepsilon + \frac{\delta_n}{\varepsilon}\right) =: \gamma_0^2,$$

where $\Pi_N$ denotes the projection onto $U$ and $C_{\mathcal{M},N} > 0$ is a constant depending on $\mathcal{M}$ and $N$ only. Then we have

$$\|Iu_{n,j}\|_{L^2(\nu)} - \gamma_0 \leq \|\Pi_N(Iu_{n,j})\|_{L^2(\nu)} \leq \|Iu_{n,j}\|_{L^2(\nu)} + \gamma_0.$$

In order to bound $\|Iu_{n,j}\|_{L^2(\nu)}$ by using Lemma 4, we first bound $b_n(u_{n,j})$ by using the convergence of eigenvalues as follows:

$$\begin{aligned} b_n\left(u_{n,j}\right) &= \left\langle u_{n,j}, \Delta_n u_{n,j}\right\rangle_{L^2(\nu_n)} = \lambda_{n,j} \\ &\leq C\left(1 + L_\rho\varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right)\lambda_N. \end{aligned}$$

Thus by using Lemma 4 and the fact that $u_{n,j}$ are normalized, we have

$$\begin{aligned} \left|\|Iu_{n,j}\|^2_{L^2(\nu)} - 1\right| \leq C\Bigg(&h\sqrt{\left(1 + L_\rho\varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_i}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right)\lambda_N} \\ &+2\alpha\left(1 + q\alpha^{2q-1}L_\rho\varepsilon\right)\cdot\left(\|\mathbf{m}^q - \rho^q\|_\infty + \varepsilon q\alpha^{q-1}L_\rho\right)\Bigg) \coloneqq \gamma_1. \end{aligned}$$

Combining previous estimate together, we have

$$1 - \gamma_2 \leq \left\|\Pi_N\left(Iu_{n,j}\right)\right\|_{L^2(\nu)} \leq 1 + \gamma_2 \quad \forall j = 1, \ldots, N,$$

where

$$\gamma_2 \coloneqq \sqrt{\gamma_1} + \gamma_0.$$

Notice that the following two equation hold for all $i \neq j$:

$$\left\langle Iu_{n,j}, Iu_{n,i}\right\rangle_{L^2(\nu)} = \frac{1}{2}\left(\|Iu_{n,j}\|^2_{L^2(\nu)} + \|Iu_{n,i}\|^2_{L^2(\nu)} - \|Iu_{n,j} - Iu_{n,i}\|^2_{L^2(\nu)}\right),$$

and

$$0 = \left\langle u_{n,j}, u_{n,i}\right\rangle_{L^2(\nu_n)} = \frac{1}{2}\left(\|u_{n,j}\|^2_{L^2(\nu_n)} + \|u_{n,i}\|^2_{L^2(\nu_n)} - \|u_{n,j} - u_{n,i}\|^2_{L^2(\nu_n)}\right).$$

Then take the difference of these two equations on both side, and use Lemma 4.15 again to obtain the following bound:

$$\left|\left\langle Iu_{n,j}, Iu_{n,i}\right\rangle_{L^2(\nu)}\right| \leq \gamma_1.$$

Thus by combining the previous inequalities and Cauchy-Schwarz inequalities, we get

$$\begin{aligned} &\left|\left\langle \Pi_N Iu_{n,i}, \Pi_N Iu_{n,j}\right\rangle_{L^2(\nu)}\right| \\ &\leq \left|\left\langle Iu_{n,i}, Iu_{n,j}\right\rangle_{L^2(\nu)}\right| + \left|\left\langle Iu_{n,j}, Iu_{n,i} - \Pi_N Iu_{n,i}\right\rangle_{L^2(\nu)}\right| + \left|\left\langle \Pi_N Iu_{n,i} Iu_{n,j} - \Pi_N Iu_{n,j}\right\rangle_{L^2(\nu)}\right| \\ &\leq \gamma_1 + 2(1 + \gamma_1)\cdot\gamma_0 =: \gamma_3, \quad \forall i \neq j. \end{aligned}$$

Then by Lemma A.1, there exists an orthonormal system $g_1, \cdots, g_N$ for $U$ satisfying:

$$\|\Pi_N I u_{n,j} - g_j\|_{L^2(\nu)} \leq \sqrt{N}\left(\frac{1}{\sqrt{1-N\gamma_3}} - 1\right), \forall j = 1, \ldots, N.$$

Combining with the first bound in this proof, we have

$$\begin{aligned}
\|Iu_{n,j} - g_j\|^2_{L^2(\nu)} &\leq \left(\|Iu_{n,j} - \Pi_N I u_{n,j}\|_{L^2(\nu)} + \|\Pi_N I u_{n,j} - g_j\|_{L^2(\nu)}\right)^2 \\
&\leq \left(\gamma_0 + \sqrt{N}\left(\frac{1}{\sqrt{1-N\gamma_3}} - 1\right)\right)^2 \\
&\leq C\left(\gamma_0 + \frac{N^{3/2}}{2}\gamma_3 + \frac{3N^{5/2}}{8}\gamma_3^2\right)^2 =: \gamma_4, \quad \forall j = 1, \ldots, N.
\end{aligned}$$

Then

$$\begin{aligned}
\|Iu_{n,j} - u_{n,j} \circ T_n\|^2_{L^2(\nu)} &= \|\Lambda_{\varepsilon - 2\delta_n} P^* u_{n,j} - P^* u_{n,j}\|^2_{L^2(\nu)} \\
&\leq \frac{C\alpha^2}{(\varepsilon - 2\delta_n)^{m+2}} E_{\varepsilon - 2\delta_n}\left(P^* u_{n,j}\right) \\
&\leq \frac{C\alpha^2}{(\varepsilon - 2\delta_n)^{m+2}}\left(1 + \frac{CL_\eta 8^m \left(1 + q\alpha^q L_\rho\right)^2}{I^2 \eta(1/2)}\frac{\varepsilon}{\delta_n}\right) I^2 \sigma_\eta \varepsilon^{m+2} b_n(u_{n,j}) \\
&\leq \frac{C\alpha^2}{(\varepsilon - 2\delta_n)^{m+2}}\left(1 + \frac{CL_\eta 8^m \left(1 + q\alpha^q L_\rho\right)^2}{I^2 \eta(1/2)}\frac{\varepsilon}{\delta_n}\right) I^2 \sigma_\eta \varepsilon^{m+2} \\
&\quad \cdot \left(1 + L_\rho \varepsilon + \frac{\delta_n}{\varepsilon} + \sqrt{\lambda_N}\delta_n + K\varepsilon^2 + \frac{\varepsilon^2}{R^2} + \|\mathbf{m}^q - \rho^q\|_\infty\right)\lambda_N \coloneqq \gamma_5.
\end{aligned}$$

Combining previous two estimates and triangle inequality, we have

$$\begin{aligned}
\|u_{n,j} \circ T_n - g_j\|^2_{L^2(\nu)} &\leq 2\left(\gamma_4 + \gamma_5\right) \\
&\leq C\left(1 + N^{3/2} + N^{5/2}\right) C_{\mathcal{M},N}\left(\frac{\lambda_N}{\lambda_{N+1} - \lambda_N}\right)(\varepsilon + \delta_n/\varepsilon) \\
&\quad + C\varepsilon^{m+2}(\varepsilon + \frac{\delta_n}{\varepsilon} + \delta_n + \varepsilon^2)\lambda_N \\
&\leq c_{\mathcal{M}}\left(\left(\frac{\lambda_N}{\lambda_{N+1} - \lambda_N}\right)\left(\varepsilon + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}}\right) + \lambda_N \varepsilon^{m+2}\left(\varepsilon + \varepsilon^2 + \frac{\log(n)^{p_m}}{\varepsilon n^{1/m}} + \frac{\log(n)^{p_m}}{n^{1/m}}\right)\right).
\end{aligned}$$

PROPOSITION 9. ***(Upper bound for $\lambda_N$):***

$$\lambda_N \leq \frac{N\mathcal{C}}{I_{min} - N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}.$$

PROOF. Denote $Q \coloneqq \mathrm{span}\{q_1, \cdots, q_N\}$ as $N$-dimensional subspace, then by the minimax theorem, we obtain that

$$\lambda_N \leq \max_{u \in Q} \frac{\langle \Delta_\rho u, u \rangle_{L^2(\nu)}}{\langle u, u \rangle_{L^2(\nu)}}.$$

Take $u \in Q$ that satisfies $\langle u, u \rangle_{L^2(\nu)} = 1$. Then there exists $a_k$'s that satisfies

$$1 = \sum_{k=1}^{N} a_k^2 \left\| q_k \right\|_{L^2(\nu)}^2 + \sum_{k=1}^{N} \sum_{j \neq k} a_k a_j \left\langle q_k, q_j \right\rangle_{L^2(\nu)} = \sum_{k=1}^{N} a_k^2 w_k^q I_k + \sum_{k=1}^{N} \sum_{j \neq k} a_k a_j \left\langle q_k, q_j \right\rangle_{L^2(\nu)},$$

such that

$$u = \sum_{i=1}^{N} a_k q_k.$$

For the last term, the norm can be bounded as

$$\left| \sum_{k=1}^{N} \sum_{j \neq k} a_k a_j \left\langle q_k, q_j \right\rangle_{L^2(\nu)} \right| \leq \sum_{k=1}^{N} \sum_{j \neq k} |a_k| |a_j| \sqrt{w_k^q} \sqrt{w_j^q} \mathcal{S}_{\mathrm{b}}^{1/2} \mathcal{I}^{*1/2} \leq N \mathcal{S}_{\mathrm{b}}^{1/2} \mathcal{I}^{*1/2} \sum_{k=1}^{N} a_k^2 w_k^q.$$

Then

$$\sum_{k=1}^{N} a_k^2 w_k^q \leq \frac{1}{I_{\min} - N \mathcal{S}_{\mathrm{b}}^{1/2} \mathcal{I}^{*1/2}}.$$

Also,

$$\langle \Delta_\rho u, u \rangle_{L^2(\nu)} = \sum_{k=1}^{N} \sum_{j=1}^{N} a_k a_j \int_{\mathcal{M}} \nabla q_k \cdot \nabla q_j \rho^q dx.$$

Recall that $\left\| \nabla q_k \right\|_{L^2(\nu)}^2 = w_k^q \mathcal{C}_k$, so we have

$$\begin{aligned}
\langle \Delta_\rho u, u \rangle_{L^2(\nu)} &\leq \sum_{k=1}^{N} \sum_{j=1}^{N} |a_k a_j| \left( \int_{\mathcal{M}} |\nabla q_k|^2 \rho^q dx \right)^{1/2} \left( \int_{\mathcal{M}} |\nabla q_j|^2 \rho^q dx \right)^{1/2} \\
&\leq \mathcal{C} N \sum_{k=1}^{N} a_k^2 w_k^q \\
&\leq \frac{N \mathcal{C}}{I_{\min} - N \mathcal{S}_{\mathrm{b}}^{1/2} \mathcal{I}^{*1/2}},
\end{aligned}$$

and thus we have

$$\lambda_N \leq \frac{N \mathcal{C}}{I_{\min} - N \mathcal{S}_{\mathrm{b}}^{1/2} \mathcal{I}^{*1/2}}.$$

□

Combining the lower bound of $\lambda_{N+1}$ and the upper bound of $\lambda_N$, we get

$$\lambda_{N+1}-\lambda_N \geq \left(\sqrt{\frac{\Theta(1-\mathcal{S}_{\text{adj}})}{\alpha^{|p-q|}}\left(\frac{1}{N^q}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)^2}{I_{\min}-\mathcal{S}_{\text{b}}}\right)}\right.$$
$$\left.-\frac{\sqrt{N\mathcal{C}}}{I_{\min}-\mathcal{S}_{\text{b}}}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_{\text{b}}}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_{\text{w}}}\right)\right)^2-\frac{N\mathcal{C}}{I_{\min}-N\mathcal{S}_{\text{b}}^{1/2}\mathcal{I}^{*1/2}}.$$

Then we have the following corollary:

COROLLARY 2. *The inequality in Theorem 9 can be replaced by*

$$\int_{\mathcal{M}}|g_j(x)-u_{n,j}\circ T_n(x)|^2\,d\nu(x)=\|g_j-u_{n,j}\circ T_n\|^2_{L^2(\nu)}\leq\phi$$

*for all* $j=1,\cdots,N$, *where*

$$\phi=\phi(\mathcal{S}_b,\mathcal{C},\Theta,I_{min},\mathcal{I}^*,N,\varepsilon,n,m)=c_{\mathcal{M}}\left(\left(\frac{N\mathcal{C}}{I_{min}-N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\left(\varepsilon+\frac{\log(n)^{p_m}}{\varepsilon n^{1/m}}\right)\psi^{-1}\right.$$
$$\left.+\left(\frac{N\mathcal{C}}{I_{min}-N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}\right)\varepsilon^{m+2}\left(\varepsilon+\varepsilon^2+\frac{\log(n)^{p_m}}{\varepsilon n^{1/m}}+\frac{\log(n)^{p_m}}{n^{1/m}}\right)\right),$$

*and*

$$\psi=\psi(\mathcal{S}_b,\mathcal{S}_{adj},\mathcal{C},\Theta,I_{min},\mathcal{I}^*,N,\varepsilon,n,m)\coloneqq\left(\sqrt{\frac{\Theta(1-\mathcal{S}_{adj})}{\alpha^{|p-q|}}\left(\frac{1}{\max(N^q,N)}-\frac{\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)^2}{I_{min}-\mathcal{S}_b}\right)}\right.$$
$$\left.-\frac{\sqrt{N\mathcal{C}}}{I_{min}-\mathcal{S}_b}\left(\max(N^{q-1},1)\sqrt{\mathcal{S}_b}+(\max(N^{q-1},1)-1)\sqrt{\mathcal{S}_w}\right)\right)^2-\frac{N\mathcal{C}}{I_{min}-N\mathcal{S}_b^{1/2}\mathcal{I}^{*1/2}}.$$

PROOF. **Proof of Theorem 2** By Theorem 9 and previous corollary, we have that for a given $\beta>1$, then with probability larger than $1-C_\beta n^{-\beta}$, there exists a transportation map $T_n:\mathcal{M}\to\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ that pushes forward $\nu$ into $\nu_n$ and an orthonormal set of functions $g_1,\cdots,g_n$ in $U$ satisfying

- $\sup_{x\in\mathcal{M}} d_{\mathcal{M}}\left(x,T_n(x)\right)\leq c_{\mathcal{M}}\frac{\log(n)^{p_m}}{n^{1/m}}$,
- $\int_{\mathcal{M}}|g_i(x)-u_{i,n}\left(T_n(x)\right)|^2\,d\nu(x)\leq\phi$.

For $x\in\mathcal{M}$, denote $G(x)\coloneqq(g_1(x),\cdots,g_N(x))$, then we have

$$\int_{\mathcal{M}}|F_n\circ T_n(x)-G(x)|^2\,d\nu(x)=\sum_{i=1}^N\int_{\mathcal{M}}|g_i(x)-u_{n,i}\left(T_n(x)\right)|^2\,d\nu(x)\leq N\phi.$$

Let $\tilde{\pi}_n \coloneqq (Id \times T_n)_\sharp \nu \in \mathcal{P}(\mathcal{M} \times \mathcal{M})$ and $G \times F_n : (x, y) \longmapsto (G(x), F_n(y))$. Denote $\pi_n \coloneqq (G \times F_n)_\sharp \tilde{\pi}_n$, the push-forward of $\tilde{\pi}_n$ by the map $G \times F_n$. Then $\pi_n$ is a transportation plan between $G_\sharp \nu$ and $F_{n\sharp}\nu_n$. Also,

$$\begin{aligned}
\int_{\mathbb{R}^N \times \mathbb{R}^N} |x - y|^2 d\pi_n(x, y) &= \int_{\mathbb{R}^N \times \mathbb{R}^N} |x - y|^2 d\left(G \times F_n\right)_\sharp \tilde{\pi}_n(x, y) \\
&= \int_{\mathcal{M} \times \mathcal{M}} \left|G(x) - F_n(y)\right|^2 d\tilde{\pi}_n(x, y) \\
&= \int_{\mathcal{M} \times \mathcal{M}} \left|G(x) - F_n(y)\right|^2 d\left(Id \times T_n\right)_\sharp \nu(x, y) \\
&= \int_{\mathcal{M}} \left|G(x) - F_n \circ T_n(x)\right|^2 d\nu(x).
\end{aligned}$$

Thus

$$\left(W_2\left(G_\sharp \nu, F_{n\sharp}\nu_n\right)\right)^2 \leq \int_{\mathcal{M}} \left|F_n \circ T_n(x) - G(x)\right|^2 d\nu(x) \leq N\phi.$$

As $g_1, \cdots, g_N$ is an orthonormal basis for $U$, there exists an orthogonal matrix $R$ such that for every $x \in \mathcal{M}$, $G(x) = RF(x)$.

Now we can choose an orthogonal transformation $O$ such that $OF(x) = \tilde{F}(x) = \sum_{j=1}^N \tilde{v}_j(x) e_j$. Thus $G = RO^{-1}\tilde{F}$, and we have

$$\begin{aligned}
W_2\left(OR^{-1}F_{n\sharp}\nu_n, F^Q_\sharp \nu\right) &\leq W_2\left(OR^{-1}F_{n\sharp}\nu_n, \tilde{F}_\sharp \nu\right) + W_2\left(\tilde{F}_\sharp \nu, F^Q_\sharp \nu\right) \\
&= W_2\left(F_{n\sharp}\nu_n, RO^{-1}\tilde{F}_\sharp \nu\right) + W_2\left(\tilde{F}_\sharp \nu, F^Q_\sharp \nu\right) \\
&= W_2\left(F_{n\sharp}\nu_n, G_\sharp \nu\right) + W_2\left(\tilde{F}_\sharp \nu, F^Q_\sharp \nu\right) \\
&\leq \sqrt{N\phi} + \sqrt{N\left(\frac{\tau - \frac{\sqrt{I\mathcal{S}_{\mathrm{b}}}}{I_{\min}}}{2}\right)^2 + 4N^{\frac{3}{2}}\left(\frac{1}{\sqrt{1 - N\tau}} - 1\right)}.
\end{aligned}$$

So the measures $OR^{-1}F_{n\sharp}\nu_n$ and $F^Q_\sharp \nu$ are close to each other with respect to the 2-Wasserstein distance. Similar to the population setting, the closeness of these two measures and the orthogonal cone structure of $F^Q_\sharp \nu$ lead to the conclusion for the orthogonal cone structure of $OR^{-1}F_{n\sharp}\nu_n$, and thus for the orthogonal cone structure of $F_{n\sharp}\nu_n$. The parameters in this theorem can be derived directly from Proposition 3. □

### 4.3. OCS of kernel PCA embedding: The population setting

The basic idea underlying the proof of OCS quantification for Kernel PCA is similar to the one used above for spectral embeddings. However, the details are quite different. The following auxiliary propositions use notations introduced in Chapter 2.

PROPOSITION 10. *For every $k = 1, \cdots, N$ we have*

$$\|q_k - \Pi_N(q_k)\|_\nu^2 \leq \frac{\Lambda}{w_{\min}},$$

*where $\Pi_N$ stands for the projection onto $U$, the span of the $N$ eigenfunctions corresponding to the largest $N$ eigenvalues of $\Sigma_\nu$.*

PROOF. For every $k = 1, \cdots, N$, $q_k$ can be written in the orthonormal basis of eigenfunctions $\{u_1, u_2, \cdots\}$ of $\Sigma_\nu$ as

$$q_k = \sum_{l=1}^{\infty} a_{lk} u_l$$

with coefficients $a_{lk}, l = 1, 2, \cdots$.

Also we have

$$\Pi_N(q_k) = \sum_{l=1}^{N} a_{lk} u_l,$$

and

$$q_k - \Pi_N(q_k) = \sum_{l=N+1}^{\infty} a_{lk} u_l.$$

So

$$\|q_k - \Pi_N(q_k)\|_\nu^2 = \sum_{l=N+1}^{\infty} a_{lk}^2,$$

where

$$\begin{aligned}
a_{lk} &= \langle q_k, u_l \rangle_\nu \\
&= \int_\Omega q_k(x) u_l(x) \nu(dx) \\
&= \int_\Omega \int_\Omega k(x,y) \nu_k(dy) u_l(x) \nu(dx) \\
&= \int_\Omega \left( \int_\Omega k(x,y) u_l(x) \nu(dx) \right) \nu_k(dy) \\
&= \int_\Omega \left( \lambda_l u_l(y) \right) \nu_k(dy) \\
&= \lambda_l \int_\Omega u_l(y) \nu_k(dy) \\
&\leq \lambda_l \sqrt{\int_\Omega u_l^2(y) \nu_k(dy) \int_\Omega \nu_k(dy)} \\
&= \lambda_l \sqrt{\int_\Omega u_l^2(y) \nu_k(dy)}.
\end{aligned}$$

The above inequality comes from Cauchy-Schwarz inequality. Recall the constraint $\int_\Omega u_l^2(y)\nu(dy) = 1$, we have that

$$\begin{aligned}
1 &= \int_\Omega u_l^2(y) \nu(dy) \\
&= \sum_{i=1}^N w_i \int_\Omega u_l^2(y) \nu_i(dy) \\
&\geq w_{\min} \sum_{i=1}^N \int_\Omega u_l^2(y) \nu_i(dy) \\
&\geq w_{\min} \int_\Omega u_l^2(y) \nu_k(dy).
\end{aligned}$$

So

$$a_{lk} = \lambda_l \sqrt{\int_\Omega u_l^2(y) \nu_k(dy)} \leq \frac{\lambda_l}{\sqrt{w_{\min}}},$$

then

$$\|q_k - \Pi_N(q_k)\|_\nu^2 = \sum_{l=N+1}^\infty a_{lk}^2 \leq \sum_{l=N+1}^\infty \left( \frac{\lambda_l}{\sqrt{w_{\min}}} \right)^2 = \frac{\Lambda}{w_{\min}}.$$

□

Recall the definition of $F_\sharp^Q \nu$ given in section 2.3.3. The next result presents an OCS for this measure.

PROPOSITION 11. *The probability measure $\mu^Q = F^Q_\sharp \nu$ with $F^Q$ defined above has an orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ for any $\sigma \in (0, \pi/4)$, $\delta^* \le \delta < 1$ and $r = \frac{1}{\sqrt{N} w_{\max}}$ where*

$$\delta^* \coloneqq \frac{N w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{w,up}}{\mathcal{S}^*_w} \sum_{k=1}^N w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \overline{\mathcal{S}_k}.$$

PROOF. For each $k = 1, \ldots, N$, let

$$C_k \coloneqq \left\{ z \in \mathbb{R}^N : \frac{z_k}{|z|} > \cos(\sigma_k), \quad |z| \ge r \right\}$$

with $r = \frac{1}{\sqrt{N} w_{\max}}$ and fixed $\sigma_k \in (0, \pi/4)$ $(k = 1, 2, \cdots, N)$.

Also denote $A_k$ as the preimage of $C_k$ through $F^Q$, i.e.

$$A_k \coloneqq \left(F^Q\right)^{-1}(C_k) = \left\{ x \in \Omega : \frac{q_k(x)}{\|q_k\|_\nu} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2} > r \right\}.$$

Then we have

$$\mu^Q(C_k) = F^Q_\sharp \nu(C_k) = \nu(A_k),$$

and $A_k$ can be re-written as

$$A_k = \left\{ x \in \Omega : \frac{q_k(x)}{\|q_k\|_\nu \, q(x)} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\|q_j\|_\nu \, q(x)} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2} > r \right\}.$$

For an arbitrary $x_0 \in A_k^c \subseteq \Omega$ $(k = 1, 2, \cdots, N)$ we have

$$\left( \frac{q_k(x_0)}{\|q_k\|_\nu \, q(x_0)} \right)^2 \le \cos^2(\sigma_k) \sum_{j=1}^N \left( \frac{q_j(x_0)}{\|q_j\|_\nu \, q(x_0)} \right)^2,$$

i.e.,

$$\left(1 - \cos^2(\sigma_k)\right) \left( \frac{q_k(x_0)}{\|q_k\|_\nu \, q(x_0)} \right)^2 \le \cos^2(\sigma_k) \sum_{j \ne k} \left( \frac{q_j(x_0)}{\|q_j\|_\nu \, q(x_0)} \right)^2.$$

So

$$\sqrt{1-\cos^2(\sigma_k)} \frac{q_k(x_0)}{\|q_k\|_\nu \, q(x_0)} \le \cos(\sigma_k) \sqrt{\sum_{j \ne k} \left( \frac{q_j(x_0)}{\|q_j\|_\nu \, q(x_0)} \right)^2} \le \cos(\sigma_k) \sum_{j \ne k} \frac{q_j(x_0)}{\|q_j\|_\nu \, q(x_0)}.$$

Thus

$$w_k^2 \left( \frac{q_k(x_0)}{q(x_0)} \right)^2 \leq w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j \neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x_0)}{q(x_0)} \frac{q_k(x_0)}{q(x_0)}.$$

Take the integral over $A_k^c$ on both sides:

$$\int_{A_k^c} w_k^2 \left( \frac{q_k(x)}{q(x)} \right)^2 \nu(dx) \leq \int_{A_k^c} w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j \neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \frac{q_k(x)}{q(x)} \nu(dx), \quad \forall k = 1, \ldots, N.$$

Take the sum over $k$:

$$\sum_{k=1}^N \int_{A_k^c} w_k^2 \left( \frac{q_k(x)}{q(x)} \right)^2 \nu(dx) \leq \sum_{k=1}^N \int_{A_k^c} w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j \neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \frac{q_k(x)}{q(x)} \nu(dx),$$

where

$$\begin{aligned}
\text{LHS} &= \sum_{k=1}^N \int_{A_k^c} w_k^2 \left( \frac{q_k(x)}{q(x)} \right)^2 \nu(dx) \\
&= \sum_{k=1}^N \int_\Omega w_k^2 \left( \frac{q_k(x) \mathbf{1}_{A_k^c}(x)}{q(x)} \right)^2 \nu(dx) \\
&= \int_\Omega \sum_{k=1}^N w_k^2 \left( \frac{q_k(x) \mathbf{1}_{A_k^c}(x)}{q(x)} \right)^2 \nu(dx) \\
&= \frac{1}{N} \int_\Omega \sum_{k=1}^N w_k^2 \left( \frac{q_k(x) \mathbf{1}_{A_k^c}(x)}{q(x)} \right)^2 \sum_{k=1}^N 1 \nu(dx) \\
&\geq \frac{1}{N} \int_\Omega \left( \sum_{k=1}^N w_k \left( \frac{q_k(x) \mathbf{1}_{A_k^c}(x)}{q(x)} \right) \right)^2 \nu(dx) \\
&\geq \frac{1}{N} \int_\Omega \mathbf{1}_{\bigcap_{l=1}^N A_l^c}(x) \left( \sum_{k=1}^N w_k \left( \frac{q_k(x)}{q(x)} \right) \right)^2 \nu(dx) \\
&= \frac{1}{N} \int_{\bigcap_{l=1}^N A_l^c} \left( \sum_{k=1}^N w_k \left( \frac{q_k(x)}{q(x)} \right) \right)^2 \nu(dx) \\
&= \frac{1}{N} \nu \left( \bigcap_{l=1}^N A_l^c \right).
\end{aligned}$$

$$
\begin{aligned}
\text{RHS} &= \sum_{k=1}^{N} \int_{A_k^c} w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j\neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \frac{q_k(x)}{q(x)} \nu(dx) \\
&= \sum_{k=1}^{N} \int_{\Omega} w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j\neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \frac{q_k(x)}{q(x)} \mathbf{1}_{A_k^c}(x) \nu(dx) \\
&\leq \int_{\Omega} \sum_{k=1}^{N} w_k^2 \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j\neq k} \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \frac{q_k(x)}{q(x)} \nu(dx) \\
&\leq \frac{w_{\max}}{w_{\min}} \int_{\Omega} \sum_{k=1}^{N} \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} w_k \frac{q_k(x)}{q(x)} \sum_{j\neq k} w_j \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \frac{q_j(x)}{q(x)} \nu(dx) \\
&= \frac{w_{\max}}{w_{\min}} \sum_{k=1}^{N} w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j\neq k} w_j \frac{\|q_k\|_\nu}{\|q_j\|_\nu} \mathcal{S}_{jk} \\
&\leq \frac{w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{\text{w,up}}}{\mathcal{S}^*_{\text{w}}} \sum_{k=1}^{N} w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \sum_{j\neq k} w_j \mathcal{S}_{jk} \\
&= \frac{w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{\text{w,up}}}{\mathcal{S}^*_{\text{w}}} \sum_{k=1}^{N} w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \overline{\mathcal{S}_k}.
\end{aligned}
$$

Thus we have

$$
\nu\left(\bigcap_{l=1}^{N} A_l^c\right) \leq \frac{N w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{\text{w,up}}}{\mathcal{S}^*_{\text{w}}} \sum_{k=1}^{N} w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \overline{\mathcal{S}_k},
$$

and this implies

$$
\mu^Q\left(\bigcup_{k=1}^{N} C_k\right) \geq 1 - \frac{N w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{\text{w,up}}}{\mathcal{S}^*_{\text{w}}} \sum_{k=1}^{N} w_k \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \overline{\mathcal{S}_k},
$$

which completes the proof. □

Now we turn to the proof of Theorem 3. The basic ideas are similar to the proof of Theorem 1.

**Proof of Theorem 3**. The measure $\mu = F_\sharp \nu$ has the same orthogonal cone structure as the measure $(OF)_\sharp \nu$, where the map $(OF)_\sharp \nu$ is defined by $x \in \Omega \mapsto OF(x) \in \mathbb{R}^N$ with $O$ being an $N \times N$ orthogonal matrix. So we will consider the measure $(OF)_\sharp \nu$ where we construct the matrix $O$ such that $(OF)_\sharp \nu$ and $F^Q_\sharp \nu$ are close to each other in the 2-Wasserstein distance. Then combining with previous propositions, we can get the orthogonal cone structure for $(OF)_\sharp \nu$. Firstly, define

the normalized projection of $q_i$'s as follows:

$$v_i \coloneqq \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_\nu}, \quad i = 1, \ldots, N,$$

where $\Pi_N : L^2(d\nu) \to U$ is the orthogonal projection onto $U$, the span of the $N$ eigenfunctions corresponding to the $N$ largest eigenvalues of $\Sigma_\nu$. Then based on the previous proposition, we have

$$\begin{aligned}
\left\| \frac{q_i}{\|q_i\|_\nu} - v_i \right\|_\nu &= \left\| \frac{q_i}{\|q_i\|_\nu} - \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_\nu} \right\|_\nu \\
&\le \left\| \frac{q_i}{\|q_i\|_\nu} - \frac{\Pi_N(q_i)}{\|q_i\|_\nu} \right\|_\nu + \left\| \frac{\Pi_N(q_i)}{\|q_i\|_\nu} - \frac{\Pi_N(q_i)}{\|\Pi_N(q_i)\|_\nu} \right\|_\nu \\
&= \left\| \frac{q_i}{\|q_i\|_\nu} - \frac{\Pi_N(q_i)}{\|q_i\|_\nu} \right\|_\nu + \frac{1}{\|q_i\|_\nu} \left| \, \|\Pi_N(q_i)\|_\nu - \|q_i\|_\nu \, \right| \\
&\le 2 \left\| \frac{q_i}{\|q_i\|_\nu} - \frac{\Pi_N(q_i)}{\|q_i\|_\nu} \right\|_\nu \\
&\le \frac{2}{\|q_i\|_\nu} \sqrt{\frac{\Lambda}{w_{\min}}} \\
&= \frac{2}{\sqrt{\mathcal{S}^*_{ii}}} \sqrt{\frac{\Lambda}{w_{\min}}} \\
&\le \frac{2}{\sqrt{\mathcal{S}^*_{\mathrm{w}}}} \sqrt{\frac{\Lambda}{w_{\min}}} \\
&= 2 \sqrt{\frac{\Lambda}{\mathcal{S}^*_{\mathrm{w}} w_{\min}}}.
\end{aligned}$$

For a given pair $(i, j)$ with $i \neq j$, we have

$$\begin{aligned}
\left| \langle v_i, v_j \rangle_\nu \right| &= \left| \left\langle v_i - \frac{q_i}{\|q_i\|_\nu}, v_j \right\rangle_\nu + \left\langle \frac{q_i}{\|q_i\|_\nu}, v_j - \frac{q_j}{\|q_j\|_\nu} \right\rangle_\nu + \left\langle \frac{q_i}{\|q_i\|_\nu}, \frac{q_j}{\|q_j\|_\nu} \right\rangle_\nu \right| \\
&\le \left\| \frac{q_i}{\|q_i\|_\nu} - v_i \right\|_\nu + \left\| \frac{q_j}{\|q_j\|_\nu} - v_j \right\|_\nu + \frac{\mathcal{S}^*_{ij}}{\sqrt{\mathcal{S}^*_{ii} \mathcal{S}^*_{jj}}} \\
&\le 4 \sqrt{\frac{\Lambda}{\mathcal{S}^*_{\mathrm{w}} w_{\min}}} + \frac{\mathcal{S}^*_{\mathrm{b}}}{\mathcal{S}^*_{\mathrm{w}}} \coloneqq \tau.
\end{aligned}$$

Thus we can conclude that there exists an orthonormal basis $\tilde{v}_1, \ldots, \tilde{v}_N$ for $(U, \langle \cdot, \cdot \rangle_\nu)$ such that

$$\|v_i - \tilde{v}_i\|_\nu^2 \le N \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right)^2, \quad i = 1, \ldots, N.$$

Thus for any $i = 1, \cdots, N$,

$$\begin{aligned}
\left\| \frac{q_i}{\|q_i\|_\nu} - \tilde{v}_i \right\|_\nu^2 &= \left\| \frac{q_i}{\|q_i\|_\nu} - v_i \right\|_\nu^2 + 2 \left\langle v_i - \tilde{v}_i, \frac{q_i}{\|q_i\|_\nu} \right\rangle_\nu - \langle v_i + \tilde{v}_i, v_i - \tilde{v}_i \rangle_\nu \\
&\leq \left\| \frac{q_i}{\|q_i\|_\nu} - v_i \right\|_\nu^2 + 4 \|v_i - \tilde{v}_i\|_\nu \\
&\leq \frac{4\Lambda}{\mathcal{S}_{\mathrm{w}}^* w_{\min}} + 4\sqrt{N} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right) \\
&= \left( \frac{\tau - \frac{\mathcal{S}_{\mathrm{b}}^*}{\mathcal{S}_{\mathrm{w}}^*}}{2} \right)^2 + 4\sqrt{N} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right).
\end{aligned}$$

Now define $\tilde{F} : \Omega \mapsto \mathbb{R}^N$ as the map $\tilde{F}(x) = \sum_{j=1}^N \tilde{v}_j(x) e_j$. Since both $\{\tilde{v}_1, \ldots, \tilde{v}_N\}$ and $\{u_1, \ldots, u_N\}$ are orthonormal bases for $(U, \langle \cdot, \cdot \rangle_\nu)$, there exists an orthogonal matrix $O \in \mathbb{R}^N \times \mathbb{R}^N$ such that

$$OF = \tilde{F}.$$

Let $\pi \coloneqq \left( F^Q \times \tilde{F} \right)_\sharp \nu$, then it is a coupling between $F_\sharp^Q \nu$ and $\tilde{F}_\sharp \nu$. Thus we have

$$\begin{aligned}
W_2^2 \left( F_\sharp^Q \nu, \tilde{F}_\sharp \nu \right) &\leq \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} |z - \tilde{z}|^2 d\pi(z, \tilde{z}) \\
&= \int_\Omega \left| F^Q(x) - \tilde{F}(x) \right|^2 d\nu(x) \\
&= \sum_{i=1}^N \left\| \frac{q_i}{\|q_i\|_\nu} - \tilde{v}_i \right\|_\nu^2 \\
&= N \left( \frac{\tau - \frac{\mathcal{S}_{\mathrm{b}}^*}{\mathcal{S}_{\mathrm{w}}^*}}{2} \right)^2 + 4N^{\frac{3}{2}} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right).
\end{aligned}$$

Also, it's easy to check the finite second moments condition of the probability measures $F_\sharp^Q \nu$ and $\tilde{F}_\sharp \nu$ as follows:

$$\begin{aligned}
\int_\Omega \left| F^Q(x) \right|^2 d\nu(x) &= \sum_{i=1}^N \left\| \frac{q_i}{\|q_i\|_\nu} \right\|_\nu^2 = N, \\
\int_\Omega \left| \tilde{F}(x) \right|^2 d\nu(x) &= \sum_{i=1}^N \|\tilde{v}_i\|_\nu^2 = N.
\end{aligned}$$

By using the previous propositions about wasserstein distance, we can derive that $\mu$ has an orthogonal cone structure with parameters $\left( \sigma_1 + s, \sigma_2 + s, \cdots, \sigma_N + s, \delta + t^2, \frac{1 - \sin(s)}{\sqrt{N} w_{\max}} \right)$ for any $\delta \in [\delta^*, 1)$

and $s, t > 0$ satisfying

$$\frac{t^2 \sin^2(s)}{N^2 w_{\max}^2} \geq N \left( \frac{\tau - \frac{\mathcal{S}_{\mathrm{b}}^*}{\mathcal{S}_{\mathrm{w}}^*}}{2} \right)^2 + 4N^{\frac{3}{2}} \left( \frac{1}{\sqrt{1 - N\tau}} - 1 \right), \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N.$$

■

### 4.4. OCS of kernel PCA embedding: The sample setting

Then we will prove Theorem 4 under the sample setting. Firstly, we need some auxiliary results to show closeness of spectrum between the covariance operator and its empirical version, including Lemma 17, Lemma 18 and Lemma 19. Lemma 17 is given in equation (2.5) in [**58**]. Lemma 18 is Theorem 9 of [**60**]. Lemma 19 is proved by Lemma 1 in [**59**].

LEMMA 17. *When $\Sigma_{\nu_n}$ is close to $\Sigma_\nu$ in the operator norm, the spectrum $\sigma(\Sigma_{\nu_n})$ of $\Sigma_{\nu_n}$ is a small perturbation of the spectrum $\sigma(\Sigma_\nu)$ of $\Sigma_\nu$. This can be expressed by the following inequality:*

$$\sup_{j \geq 1} |\lambda_{n,j} - \lambda_j| \leq \|\Sigma_{\nu_n} - \Sigma_\nu\|_\infty.$$

In the following, let $r(\Sigma_\nu) \coloneqq \frac{\mathrm{tr}(\Sigma_\nu)}{\|\Sigma_\nu\|_\infty}$.

LEMMA 18. *If $\Phi(x) = k(\cdot, x)$ with $x \sim \nu$ is sub-Gaussian and pre-Gaussian, then*

$$\|\Sigma_{\nu_n} - \Sigma_\nu\|_\infty \leq C \|\Sigma_\nu\|_\infty \left( \sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n} \right)$$

*for some numerical constant $C > 0$ with probability at least $1 - e^{-\beta}, \beta > 0$.*

Recall that $\Sigma_\nu = \sum_{i=1}^\infty \lambda_i u_i \otimes u_i$ and $\Sigma_{\nu_n} = \sum_{i=1}^\infty \lambda_i u_{n,i} \otimes u_{n,i}$. Define

$$g_i \coloneqq g_i(\Sigma_\nu) \coloneqq \lambda_i - \lambda_{i+1} > 0, i \geq 1,$$

and define the $i$-th spectral gap $\bar{g}_i \coloneqq \bar{g}_i(\Sigma_\nu) \coloneqq \min(g_{i-1}, g_i)$ for $i \geq 2$ and $\bar{g}_1 \coloneqq g_1$. Then we have the following bound:

LEMMA 19.

$$\|u_{n,i} \otimes u_{n,i} - u_i \otimes u_i\|_\infty \leq \frac{4\|\Sigma_{\nu_n} - \Sigma_\nu\|_\infty}{\bar{g}_i}.$$

The following lemma shows that the closeness of the spectra of the covariance operator and its empirical version implies the closeness of their eigenfunctions.

LEMMA 20. *If $\Phi(x) = k(\cdot, x)$ with $x \sim \nu$ is sub-Gaussian and pre-Gaussian, then*

$$\|u_i - \text{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}})u_{n,i}\|_{\mathcal{H}} \leq \frac{8C\|\Sigma_\nu\|_\infty \left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)}{\bar{g}_i}$$

*with probability at least $1 - e^{-\beta}$.*

PROOF. For any $f, g, h \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} = \|g\|_{\mathcal{H}} = \|h\|_{\mathcal{H}} = 1$ and $\langle f, g\rangle_{\mathcal{H}} \geq 0$, we have

$$(f \otimes f - g \otimes g)h = f\langle f, h\rangle - g\langle g, h\rangle.$$

So

$$\|f\langle f, h\rangle_{\mathcal{H}} - g\langle g, h\rangle_{\mathcal{H}}\|_{\mathcal{H}} = \|(f \otimes f - g \otimes g)h\|_{\mathcal{H}} \leq \|f \otimes f - g \otimes g\|_\infty \|h\|_{\mathcal{H}}.$$

Let $h = f$ and $h = g$, respectively, we get

$$\|f - g\langle g, f\rangle_{\mathcal{H}}\|_{\mathcal{H}} = \|(f \otimes f - g \otimes g)h\|_{\mathcal{H}} \leq \|f \otimes f - g \otimes g\|_\infty \|f\|_{\mathcal{H}},$$

and

$$\|f\langle f, g\rangle_{\mathcal{H}} - g\|_{\mathcal{H}} = \|(f \otimes f - g \otimes g)h\|_{\mathcal{H}} \leq \|f \otimes f - g \otimes g\|_\infty \|g\|_{\mathcal{H}}.$$

Add them together, then

$$\|f - g\langle g, f\rangle_{\mathcal{H}}\|_{\mathcal{H}} + \|f\langle f, g\rangle_{\mathcal{H}} - g\|_{\mathcal{H}} \leq \|f \otimes f - g \otimes g\|_\infty (\|f\|_{\mathcal{H}} + \|g\|_{\mathcal{H}}) = 2\|f \otimes f - g \otimes g\|_\infty.$$

Also,

$$\begin{aligned}
\|f - g\langle g, f\rangle_{\mathcal{H}}\|_{\mathcal{H}} + \|f\langle f, g\rangle_{\mathcal{H}} - g\|_{\mathcal{H}} &\geq \|f - g\langle g, f\rangle_{\mathcal{H}} + f\langle f, g\rangle_{\mathcal{H}} - g\|_{\mathcal{H}} \\
&= \|(1 + \langle f, g\rangle_{\mathcal{H}})(f - g)\|_{\mathcal{H}} \\
&= (1 + \langle f, g\rangle_{\mathcal{H}})\|(f - g)\|_{\mathcal{H}} \\
&\geq \|f - g\|_{\mathcal{H}}.
\end{aligned}$$

So

$$\|f - g\|_{\mathcal{H}} \leq 2\|f \otimes f - g \otimes g\|_\infty.$$

Now let $f = u_i$ and $g = \text{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}})u_{n,i}$, then by using Lemmas 19 and 20, we have

$$\begin{aligned}\|u_i - \text{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}})u_{n,i}\|_{\mathcal{H}} &\leq 2\|u_i \otimes u_i - u_{n,i} \otimes u_{n,i}\|_\infty \\ &\leq \frac{8\|\Sigma_{\nu_n} - \Sigma_\nu\|_\infty}{\bar{g}_i} \\ &\leq \frac{8C\|\Sigma_\nu\|_\infty \left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)}{\bar{g}_i}\end{aligned}$$

holds with probability at least $1 - e^{-\beta}$. □

**Proof of Theorem 4**. Lemmas 17, 18, 19 and 20 describe how eigenvalues and eigenvectors of the empirical covariance operator approximate eigenvalues and eigenfunctions of the population covariance operator. These properties will now be used in this proof.

Then we are able to bound the Wasserstein distance between $F_{n\sharp}\nu_n$ and $F_\sharp\nu$ by the following steps. Firstly, define function $T_n$ as

$$T_n(x) = \sum_{i=1}^{n} x_i \mathbf{1}_{\{x \in V_i\}}, x \in \Omega,$$

where $V_i, i = 1, \cdots, n$ are Voronoi cells

$$V_i \coloneqq \left\{x \in \Omega : |x - x_i| = \min_{j=1,\cdots,n} |x - x_j|\right\}.$$

Then, let $\tilde{\pi}_n \in \mathcal{P}(\Omega \times \Omega)$ be given by

$$\tilde{\pi}_n \coloneqq (Id \times T_n)_\sharp \nu,$$

and let $F \times \tilde{F}_n : \Omega \times \Omega \to \mathbb{R}^N \times \mathbb{R}^N$ be given by

$$F \times \tilde{F}_n : (x, y) \longmapsto \left(F(x), \tilde{F}_n(y)\right).$$

Let $\pi_n \coloneqq \left(F \times \tilde{F}_n\right)_\sharp \tilde{\pi}_n$ (i.e. the push-forward of $\tilde{\pi}_n$ by the map $F \times \tilde{F}_n$). Thus $\pi_n$ is a transportation plan between $F_\sharp \nu$ and $\tilde{F}_{n\sharp}\nu_n$. Then

$$
\begin{aligned}
&W_2^2\left(F_\sharp \nu, \tilde{F}_{n\sharp}\nu_n\right) \\
&\leq \int_{\mathbb{R}^N}\int_{\mathbb{R}^N} |x-y|^2 d\pi_n(x,y) \\
&= \int_{\mathbb{R}^N\times\mathbb{R}^N} |x-y|^2 d\left(F \times \tilde{F}_n\right)_\sharp \tilde{\pi}_n(x,y) \\
&= \int_{\Omega\times\Omega} \left|F(x) - \tilde{F}_n(y)\right|^2 d\tilde{\pi}_n(x,y) \\
&= \int_{\Omega\times\Omega} \left|F(x) - \tilde{F}_n(y)\right|^2 d\left(Id \times T_n\right)_\sharp \nu(x,y) \\
&= \int_{\Omega} \left|F(x) - \tilde{F}_n \circ T_n(x)\right|^2 d\nu(x) \\
&= \sum_{i=1}^N \int_\Omega |u_i(x) - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}(T_n(x))|^2 d\nu(x) \\
&= \sum_{i=1}^N \int_\Omega |u_i(x) - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}(x) + \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}(x) - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}(T_n(x))|^2 d\nu(x) \\
&\leq 2\sum_{i=1}^N \int_\Omega \left(|u_i(x) - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}(x)|^2 + |u_{n,i}(x) - u_{n,i}(T_n(x))|^2\right) d\nu(x) \\
&= 2\sum_{i=1}^N \|u_i - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}\|_\nu^2 + 2\sum_{i=1}^N \int_\Omega |u_{n,i}(x) - u_{n,i}(T_n(x))|^2 d\nu(x) \\
&\leq 2\sum_{i=1}^N \|u_i - \mathrm{sign}(\langle u_i, u_{n,i}\rangle_{\mathcal{H}}) u_{n,i}\|_{\mathcal{H}}^2 + 2\sum_{i=1}^N \int_\Omega |u_{n,i}(x) - u_{n,i}(T_n(x))|^2 d\nu(x) \\
&\leq \frac{128C^2N\|\Sigma_\nu\|_\infty^2 \left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)^2}{(\min_{i=1,2,\cdots,n} \bar{g}_i)^2} + 2\sum_{i=1}^N \int_\Omega |u_{n,i}(x) - u_{n,i}(T_n(x))|^2 d\nu(x)
\end{aligned}
$$

holds with probability at least $1 - e^{-\beta}$, where

$$
u_{n,i}(x) - u_{n,i}(T_n(x)) = \frac{1}{\sqrt{n}\lambda_{n,i}} \sum_{j=1}^n v_{ji} \left(\bar{k}(x, x_j) - \bar{k}(T_n(x), x_j)\right) \leq \frac{2M}{\lambda_{n,i}},
$$

and the last inequaility comes from Lemma 20.

Denote $\bar{g}_{\min} = \min_{i=1,2,\cdots,n} \bar{g}_i$, and recall that (from Lemma 17)

$$\sup_{j\geq 1} |\lambda_{n,j} - \lambda_j| \leq \|\Sigma_{\nu_n} - \Sigma_\nu\|_\infty.$$

We have

$$\begin{aligned}
&W_2^2\left(F_\sharp\nu, \tilde{F}_{n\sharp}\nu_n\right) \\
&\leq \frac{128C^2N\|\Sigma_\nu\|_\infty^2\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)^2}{(\bar{g}_{\min})^2} + 8M^2\sum_{i=1}^N \frac{1}{\lambda_{n,i}^2} \\
&\leq \frac{128C^2N\|\Sigma_\nu\|_\infty^2\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)^2}{(\bar{g}_{\min})^2} + 8M^2\sum_{i=1}^N \frac{1}{\left(\lambda_i - C\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)\right)^2}
\end{aligned}$$

holds with probability at least $1 - 2e^{-\beta}$, where the sencond inequality comes from Lemma 18 and we assume that $n$ is large enough such that $C\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right) < \lambda_i$, as required at the end of Theorem 4.

Then we have

$$\begin{aligned}
&W_2\left(O\tilde{F}_{n\sharp}\nu_n, F_\sharp^Q\nu\right) \\
&\leq W_2\left(\tilde{F}_\sharp\nu, O\tilde{F}_{n\sharp}\nu_n\right) + W_2\left(\tilde{F}_\sharp\nu, F_\sharp^Q\nu\right) \\
&= W_2\left(O^{-1}\tilde{F}_\sharp\nu, \tilde{F}_{n\sharp}\nu_n\right) + W_2\left(\tilde{F}_\sharp\nu, F_\sharp^Q\nu\right) \\
&= W_2\left(F_\sharp\nu, \tilde{F}_{n\sharp}\nu_n\right) + W_2\left(\tilde{F}_\sharp\nu, F_\sharp^Q\nu\right) \\
&\leq \sqrt{\frac{128C^2N\|\Sigma_\nu\|_\infty^2\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)^2}{(\bar{g}_{\min})^2} + 8M^2\sum_{i=1}^N \frac{1}{\left(\lambda_i - C\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{\beta}{n}} \vee \frac{\beta}{n}\right)\right)^2}} \\
&+ \sqrt{N\left(\frac{\tau - \frac{\mathcal{S}_{\mathrm{b}}^*}{\mathcal{S}_{\mathrm{w}}^*}}{2}\right)^2 + 4N^{\frac{3}{2}}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right)}
\end{aligned}$$

holds with probability at least $1 - 2e^{-\beta}$.

Also, it's easy to check the finite second moments condition of the probability measures $O\tilde{F}_{n\sharp}\nu_n$ as follows:

$$\int_\Omega \left|O\tilde{F}_n(x)\right|^2 d\nu_n(x) = \int_\Omega \left|\tilde{F}_n(x)\right|^2 d\nu_n(x) = \sum_{i=1}^N \|u_{n,i}\|_{\nu_n}^2 = \sum_{i=1}^N \sum_{j=1}^n u_{n,i}^2(x_j) = N.$$

By using Proposition 3 about Wasserstein distance again, we can derive that with probability at least $1-2e^{-\beta}$, the probability measure $\tilde{F}_{n\sharp}\nu_n$ (equivalently, $\mu_n = F_{n\sharp}\nu_n$) has an orthogonal cone structure with parameters $\left(\sigma_1+s, \sigma_2+s, \cdots, \sigma_N+s, \delta+t^2, \frac{1-\sin(s)}{\sqrt{N}w_{\max}}\right)$ for any $\delta \in [\delta^*, 1)$ and $s, t > 0$ satisfying

$$\frac{t\sin(s)}{Nw_{\max}} \geq \sqrt{\frac{128C^2N\|\Sigma_\nu\|_\infty^2\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)^2}{(\bar{g}_{\min})^2} + 8M^2\sum_{j=1}^N \frac{1}{\left(\lambda_i - C\left(\sqrt{\frac{r(\Sigma_\nu)}{n}} \vee \frac{r(\Sigma_\nu)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)\right)^2}}$$
$$+ \sqrt{N\left(\frac{\tau - \frac{\mathcal{S}_{\mathrm{b}}^*}{\mathcal{S}_{\mathrm{w}}^*}}{2}\right)^2 + 4N^{\frac{3}{2}}\left(\frac{1}{\sqrt{1-N\tau}} - 1\right)}, \quad s + \sigma_i < \frac{\pi}{4}, i = 1, \cdots, N.$$

■

### 4.5. Strong Version of OCS

As stated in Remark 1, all the four major theorems proved above can be generalized to the strong version of OCS, i.e. each cone covers one specific component, which requires only a small modification in Proposition 6 and Proposition 11. We obtained an upper bound of $\nu\left(\bigcap_{l=1}^N A_l^c\right)$ and thus get an lower bound of $\mu_k^Q\left(\bigcup_{k=1}^N C_k\right)$. Instead of considering all cones together, we can obtain upper bounds of $\nu_k\left(A_k^c\right)$ for $k = 1, 2, \cdots, N$ respectively, and thus get lower bounds of $\mu^Q\left(C_k\right)$ for $k = 1, 2, \cdots, N$ respectively. In order to distinguish it from the original definition, we call it as **strong OCS**. In Proposition 1, we have already used the concept of strong OCS since we assumed that the embedded points have latent labels.

The following Proposition 12 and Proposition 13 gives similar results to Proposition 6 and Proposition 11. We can find the difference of the $\delta^*$'s defined in these four Propositions. Proposition 12 defines a significantly larger $\delta^*$ than Proposition 6 showing in the coefficient $N^{2q}$ and Proposition

13 also defines a significantly larger $\delta^*$ than Proposition 11 showing in the coefficient $N^2$. This is consistent with our intuition that strong versions of the OCS may have smaller coverage than the original version.

PROPOSITION 12. *(modified from proposition 6) The probability measure $\mu^Q = F_\sharp^Q \nu$ has a strong orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ for any $\sigma_1, \sigma_2, \cdots, \sigma_N \in (0, \frac{\pi}{4})$, $\delta^* \leq \delta < 1$ and $r = \frac{1}{\sqrt{\max(N^{q-1},1) w_{\max}^q}}$ where*

$$\delta^* = \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \frac{I_{max} N^{2q+1}}{I_{min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{l=1}^N w_l^q \overline{\mathcal{S}_l}.$$

PROOF. For each $k = 1, \ldots, N$, let

$$C_k \coloneqq \left\{ z \in \mathbb{R}^N : \frac{z_k}{|z|} > \cos(\sigma_k), \quad |z| \geq r \right\}$$

with $r = \frac{1}{\sqrt{\max(N^{q-1},1) w_{\max}^q I_{\max}}}$ and fixed $\sigma_k \in (0, \pi/4)$ $(k = 1, 2, \cdots, N)$.

Also denote $A_k$ as the preimage of $C_k$ through $F^Q$, i.e.

$$A_k \coloneqq \left(F^Q\right)^{-1}(C_k) = \left\{ x \in \mathcal{M} : \frac{q_k(x)}{\sqrt{I_k w_k^q}} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2} > r \right\}.$$

Then we have

$$\mu^Q(C_k) = F_\sharp^Q \nu(C_k) = \nu(A_k),$$

and the condition $\left( \sum_{j=1}^N \left( \frac{q_j(x)}{\sqrt{I_j w_j^q}} \right)^2 \right)^{1/2} > r$ is redundant because of the definition of $r$. Thus $A_k$ can be re-written as

$$A_k = \left\{ x \in \mathcal{M} : \sqrt{\frac{\rho_k^q(x)}{I_k \rho^q(x)}} > \cos(\sigma_k) \left( \sum_{j=1}^N \left( \sqrt{\frac{\rho_j^q(x)}{I_j \rho^q(x)}} \right)^2 \right)^{1/2} \right\}.$$

For an arbitrary $x_0 \in A_k^c \subseteq \Omega$ $(k = 1, 2, \cdots, N)$ we have

$$\frac{\rho_k^q(x_0)}{I_k \rho^q(x_0)} \leq \cos^2(\sigma_k) \sum_{j=1}^N \frac{\rho_j^q(x_0)}{I_j \rho^q(x_0)},$$

i.e.,

$$\left(1-\cos^2(\sigma_k)\right)\frac{\rho_k^q(x_0)}{I_k} \le \cos^2(\sigma_k)\sum_{j\neq k}\frac{\rho_j^q(x_0)}{I_j}.$$

Since $\sum_{k=1}^N \frac{w_k\rho_k(x_0)}{\rho(x_0)} = 1$, we know that there exists a $\hat{k} \in \{1, 2, \cdots, N\}$ for which

$$\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)} \ge \frac{1}{N}.$$

Thus we have

$$\begin{aligned}
\frac{1-\cos^2(\sigma_k)}{N^{2q}} &\le (1-\cos^2(\sigma_k))\left(\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)}\right)^{2q} \\
&\le \cos^2(\sigma_k)\frac{I_k}{\rho_k^q(x_0)}\sum_{j\neq\hat{k}}\frac{\rho_j^q(x_0)}{I_j}\left(\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)}\right)^{2q} \\
&= \cos^2(\sigma_k)I_k\sum_{j\neq\hat{k}}\frac{1}{I_j}\frac{\rho_j^q(x_0)}{\rho_{\hat{k}}^q(x_0)}\left(\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)}\right)^{2q} \\
&= \cos^2(\sigma_k)I_k\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_{j\neq\hat{k}}\frac{1}{I_j}\left(\frac{w_j\rho_j(x_0)}{\rho(x_0)}\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)}\right)^q \\
&\le \cos^2(\sigma_k)I_k\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_k\sum_{j\neq k}\frac{1}{I_j}\left(\frac{w_j\rho_j(x_0)}{\rho(x_0)}\frac{w_{\hat{k}}\rho_{\hat{k}}(x_0)}{\rho(x_0)}\right)^q.
\end{aligned}$$

This is true for every $x_0 \in A_k^c$, so

$$\begin{aligned}
\frac{1-\cos^2(\sigma_k)}{N^{2q}}\nu(A_k^c) &\le \cos^2(\sigma_k)I_k\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_l\sum_{j\neq l}\frac{w_l^q w_j^q}{I_j}\int_{\mathcal{M}}\left(\frac{\rho_l\rho_j}{\rho^2}\right)^q\rho^q dx \\
&\le \cos^2(\sigma_k)\frac{I_{\max}}{I_{\min}}\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_{l=1}^N w_l^q\overline{\mathcal{S}_l},
\end{aligned}$$

and thus

$$\nu(A_k^c) \le \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\frac{I_{\max}N^{2q}}{I_{\min}}\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_{l=1}^N w_l^q\overline{\mathcal{S}_l}.$$

Recall that $\nu_k(A_k^c) \le \frac{\nu(A_k^c)}{w_k}$, we have

$$\sum_{k=1}^N w_k\nu_k(A_k^c) \le \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)}\frac{I_{\max}N^{2q+1}}{I_{\min}}\left(\frac{w_{\max}}{w_{\min}}\right)^q\sum_{l=1}^N w_l^q\overline{\mathcal{S}_l},$$

and this implies

$$\sum_{k=1}^{N} w_k \mu_k^Q(C_k) \geq 1 - \frac{\cos^2(\sigma_k)}{1-\cos^2(\sigma_k)} \frac{I_{\max} N^{2q+1}}{I_{\min}} \left(\frac{w_{\max}}{w_{\min}}\right)^q \sum_{l=1}^{N} w_l^q \overline{\mathcal{S}_l},$$

which completes the proof. □

PROPOSITION 13. *(modified from proposition 11) The probability measure $\mu^Q = F_\sharp^Q \nu$ with $F^Q$ defined above has a strong orthogonal cone structure with parameters $(\sigma_1, \sigma_2, \cdots, \sigma_N, \delta, r)$ for any $\sigma \in (0, \pi/4)$, $\delta^* \leq \delta < 1$ and $r = \frac{1}{\sqrt{N} w_{\max}}$ where*

$$\delta^* \coloneqq \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \frac{N^3 w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{w,up}}{\mathcal{S}^*_w} \sum_l w_l \overline{\mathcal{S}_l}.$$

PROOF. For each $k = 1, \ldots, N$, let

$$C_k \coloneqq \left\{ z \in \mathbb{R}^N : \frac{z_k}{|z|} > \cos(\sigma_k), \quad |z| \geq r \right\}$$

with $r = \frac{1}{\sqrt{N} w_{\max}}$ and fixed $\sigma_k \in (0, \pi/4)$ $(k = 1, 2, \cdots, N)$.

Also denote $A_k$ as the preimage of $C_k$ through $F^Q$, i.e.

$$A_k \coloneqq \left(F^Q\right)^{-1}(C_k) = \left\{ x \in \Omega : \frac{q_k(x)}{\|q_k\|_\nu} > \cos(\sigma_k) \left( \sum_{j=1}^{N} \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^{N} \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2} > r \right\}.$$

Then we have

$$\mu^Q(C_k) = F_\sharp^Q \nu(C_k) = \nu(A_k),$$

and $A_k$ can be re-written as

$$A_k = \left\{ x \in \Omega : \frac{q_k(x)}{\|q_k\|_\nu \, q(x)} > \cos(\sigma_k) \left( \sum_{j=1}^{N} \left( \frac{q_j(x)}{\|q_j\|_\nu \, q(x)} \right)^2 \right)^{1/2}, \left( \sum_{j=1}^{N} \left( \frac{q_j(x)}{\|q_j\|_\nu} \right)^2 \right)^{1/2} > r \right\}.$$

For an arbitrary $x_0 \in A_k^c \subseteq \Omega$ $(k = 1, 2, \cdots, N)$ we have

$$\left( \frac{q_k(x_0)}{\|q_k\|_\nu \, q(x_0)} \right)^2 \leq \cos^2(\sigma_k) \sum_{j=1}^{N} \left( \frac{q_j(x_0)}{\|q_j\|_\nu \, q(x_0)} \right)^2,$$

i.e.,

$$\left(1-\cos^2(\sigma_k)\right)\left(\frac{q_k(x_0)}{\|q_k\|_\nu\, q(x_0)}\right)^2 \le \cos^2(\sigma_k)\sum_{j\neq k}\left(\frac{q_j(x_0)}{\|q_j\|_\nu\, q(x_0)}\right)^2.$$

By the fact that $\sum_{k=1}^N \frac{w_k\rho_k(x_0)}{\rho(x_0)} = 1$ and the definition of $q(\cdot)$ and $q_k(\cdot)$, we know that $\sum_{k=1}^N \frac{w_k q_k(x_0)}{q(x_0)} = 1$, and thus there exists a $\hat{k} \in \{1, 2, \cdots, N\}$ for which

$$\frac{w_{\hat{k}} q_{\hat{k}}(x_0)}{q(x_0)} \ge \frac{1}{N}.$$

Thus we have

$$\begin{aligned}
\frac{\sqrt{1-\cos^2(\sigma_k)}}{N^2} &\le \sqrt{1-\cos^2(\sigma_k)}\left(\frac{w_{\hat{k}} q_{\hat{k}}(x_0)}{q(x_0)}\right)^2 \\
&\le \cos(\sigma_k)\frac{\|q_k\|_\nu\, q(x_0)}{q_k(x_0)}\sqrt{\sum_{j\neq\hat{k}}\left(\frac{q_j(x_0)}{\|q_j\|_\nu\, q(x_0)}\right)^2}\left(\frac{w_{\hat{k}} q_{\hat{k}}(x_0)}{q(x_0)}\right)^2 \\
&\le \cos(\sigma_k)\frac{\|q_k\|_\nu\, q(x_0)}{q_k(x_0)}\sum_{j\neq\hat{k}}\left(\frac{q_j(x_0)}{\|q_j\|_\nu\, q(x_0)}\right)\left(\frac{w_{\hat{k}} q_{\hat{k}}(x_0)}{q(x_0)}\right)^2 \\
&\le \cos(\sigma_k)\,\|q_k\|_\nu\,\frac{w_{\max}}{w_{\min}}\sum_k\sum_{j\neq k}\frac{w_k w_j}{\|q_j\|_\nu}\frac{q_k(x_0)q_j(x_0)}{q^2(x_0)}.
\end{aligned}$$

This is true for every $x_0 \in A_k^c$, so

$$\begin{aligned}
\frac{\sqrt{1-\cos^2(\sigma_k)}}{N^2}\nu(A_k^c) &\le \cos(\sigma_k)\frac{w_{\max}}{w_{\min}}\frac{\mathcal{S}^*_{\mathrm{w,up}}}{\mathcal{S}^*_{\mathrm{w}}}\sum_l\sum_{j\neq l} w_l w_j\int_{\mathcal{M}}\frac{q_k q_j}{q^2}\nu(dx) \\
&= \cos(\sigma_k)\frac{w_{\max}}{w_{\min}}\frac{\mathcal{S}^*_{\mathrm{w,up}}}{\mathcal{S}^*_{\mathrm{w}}}\sum_l w_l\overline{\mathcal{S}_l},
\end{aligned}$$

and thus

$$\nu(A_k^c) \le \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}}\frac{N^2 w_{\max}}{w_{\min}}\frac{\mathcal{S}^*_{\mathrm{w,up}}}{\mathcal{S}^*_{\mathrm{w}}}\sum_l w_l\overline{\mathcal{S}_l}.$$

Recall that $\nu_k(A_k^c) \le \frac{\nu(A_k^c)}{w_k}$, we have

$$\sum_{k=1}^N w_k\nu_k(A_k^c) \le \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}}\frac{N^3 w_{\max}}{w_{\min}}\frac{\mathcal{S}^*_{\mathrm{w,up}}}{\mathcal{S}^*_{\mathrm{w}}}\sum_l w_l\overline{\mathcal{S}_l},$$

and this implies

$$\sum_{k=1}^{N} w_k \mu_k^Q(C_k) \geq 1 - \frac{\cos(\sigma_k)}{\sqrt{1-\cos^2(\sigma_k)}} \frac{N^3 w_{\max}}{w_{\min}} \frac{\mathcal{S}^*_{\mathrm{w,up}}}{\mathcal{S}^*_{\mathrm{w}}} \sum_l w_l \overline{\mathcal{S}_l},$$

which competes the proof. □

CHAPTER 5

# Simulations

This chapter consists of two parts. The first part presents numerical experiments illustrating the above theoretical contributions. In the second part, we present some explorations of the relationships between well-separateness of the mixture model and the OCS. In particular, we address the question whether it is possible to draw conclusions about the number of clusters (mixture components) by using OCS features of the embeddings. Note that this is the inverse of the problem studied theoretically above.

## 5.1. Numerical experiments illustrating the theoretical results

The theoretical results about the OCS presented above depend on various parameters, and it is not straightforward to immediately see how the statements vary in these parameters. Therefore, we explore this dependence numerically in this section. First, as a sanity check, we numerically compute $\delta^*$ from Theorem 2 and Theorem 4 in different situations in order to verify that the dependence of $\delta^*$ on these parameters is as expected, or that the values of $\delta^*$ are 'reasonable'. Overall, the behavior of $\delta^*$ as a function of the various parameters is as expected. However, as it turns out, for some combinations of parameters, the value of $\delta^*$ is larger than 1, rendering the presented inequality trivial. All the numerical examples considered here are in dimension 2.

### 5.1.1. Weighted Laplacian case.

**Equal mixture of two Gaussians**. We consider a simple 2-dimensional Gaussian mixture model consisting of an equal mixture of a standard normal and a shifted standard normal with mean $(\gamma, \gamma)^T$ and try different values of $\gamma$ and different values of the parameter $q$. The considered values for $\gamma$ are 2, 3, and 4, and the values for $q$ are 0.5, 1, and 2.

Indeed, the numerical experiments verify the expected behavior: Fixing the remaining parameters,

- $\delta^*$ is decreasing in $\gamma$,

- $\delta^*$ is decreasing in $q$,
- $\delta^*$ is decreasing in $\sigma$.

As can be seen in Table A.1, the value of $\delta^*$ sometimes is larger than 1, so that in these cases the obtained bound is trivial. A plot of $\log(\delta^*)$ and $\sigma$ and of $\log(\delta^*)$ and $q$, respectively, appears to be approximately linear, again confirming the intuition (see figure 5.1).

**Equal mixture of two Gaussians with different variances**. We consider two Gaussians with different variances, i.e., one Gaussian is more concentrate than another one. Here the behavior is qualitatively similar to the previous case of equal variances if the difference of two variances are not too large compared to the difference of two means. See Table A.2.

**Equal Mixture of two Uniforms**. Here we consider the equal mixture of two uniform distributions on squares, one of them is uniform on $[0,1]^2$ and the other is uniform on $[a,b]^2$, where we consider different values of $a$ and $b$ (see Table A.3). Assume $0 < a < 1 < b$, then in this case, we observe that: Fixing the remaining parameters,

- $\delta^*$ is decreasing in $a$,
- $\delta^*$ is decreasing in $b$,
- $\delta^*$ is decreasing in $q$,
- $\delta^*$ is decreasing in $\sigma$.

The first and the second bullet points simply say that a a higher overlap has a negative effect.

**Equal mixture of a Gaussian and a Uniform**. Next, we consider the equal mixture of a 2-dimensional standard normal with a uniform on $[a,b]^2$, where we consider different choices of $a$ and $b$, and again different powers $q$. When $b-a$ is fixed, and $[a,b]^2$ moves away from the center of the Gaussian, then the OCS gets stronger. Since we keep the angle $\sigma$ fixed, we observe that $\delta^*$ is decreasing, as expected. Again, $\delta^*$ is decreasing in q, except in the case $[a,b]^2 = [0,1]^2$, where the behavior of $\delta^*$ is not monotonically decreasing in $q$. See Table A.4.

**Equal mixture of three Gaussians**. Then we consider three Gaussian components with different variance (1, 3, and 2, respectively). The distance between the center of component 1 and 2, 2 and 3 are denoted as $\gamma_1$ and $\gamma_2$. The behavior of incorrect coverage ratio $\delta^*$ with respect to them

and power parameter $q$ are explored. Most behaviors are similar with two Gaussian case and are omitted here. An additional interesting finding comes from the case where $\gamma_1 = 2$, $\gamma_2 = 3$ and the case where $\gamma_1 = 3$, $\gamma_2 = 2$. Their behaviors are slightly different because of the different variance of components 1 and 3. If their positions are fixed, then when component 2 is closer to more concentrate component (with smaller variance) is more probable to be well-separated (see Table A.5). All these results are based on reasonable choices of the variances such that the pairwise differences of the three variances are not too large compared to the pairwise differences of the three means.

**Equal mixture of an annulus and a ball inside the annulus**. This example has two connected components. The first connected component consists of points lying in a ball centered at the origin with unit radius, and the second connected component consists of points lying in an annulus (also centered at the origin) with radius $r$ and 'thickness' $\eta$. The boundary of this annulus is a small circle with radius $r - \eta$ and a large circle with radius $r + \eta$. We generate the samples with noise, and more specifically, the first group of points is drawn from a two-dimensional standard normal distribution, and the second group of points is drawn from uniformly distributed points on the boundary of a circle with radius $r$ and then added by two-dimensional Gaussian noise with common standard deviations of $\eta$ on both two dimensions. Note that the annulus is a non-convex connected component, and the two connected components only have a small overlap caused by the random noise. This is a standard example for non-linear clustering, because linear methods are unable to separate the clusters, and so one is interested to explore in how far the non-linear methods can. Here we consider $\delta^*$ as a function of $r$ and $\eta$. The behavior is as expected (see Table A.6) and can be summarized as follows: Fixing the remaining parameters,

- $\delta^*$ is decreasing in $r$,
- $\delta^*$ is increasing in $\eta$,
- $\delta^*$ is decreasing in $q$.

**Mixture of two ellipses and the figure shaped as '$\infty$'**. The last example is a generalized version of previous case, and there are three connected components. The first connected component is an ellipse centered at the origin with long axis lying on the angle bisector of x-axis and y-axis with length $\eta$ and eccentricity $\rho$. The second connected component is an ellipse centered at $(12, 0)^T$

with long axis lying on the angle bisector of x-axis and the straight line of $x = 12$ (parallel with y-axis) with length $\eta$ and eccentricity $\rho$. The third connected component is an area combined by two annuli and shaped as $\infty$, where the two annuli touch each other just one point. These two annulus have the same centers with the first two components, respectively, and their radii and thickness are kept as 5.5 and 1. We generate the samples with mixture weights $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{2}$, where $\infty$ has weight $\frac{1}{2}$, and noises are also added in this example. More specifically, the first group of points is drawn from a bivariate normal distribution with common standard deviation of $\eta$ on both two dimensions and correlation $\rho$ of the two dimensions, and the second group of points is drawn from the same distribution with a location shift of length 12. The third group of points is drawn from uniformly distributed points on the boundary of two circles with radius 5.5 and then added by Gaussian noise with a standard deviation of 1, which have the same centers with the first two components, respectively. Here we consider $\delta^*$ as a function of $\eta$ and $\rho$. The variance parameter $\eta$ can represent the magnitude of the two middle components while the correlation parameter $\rho$ quantifies the shape. Note that the extreme cases $\rho = 0$ and $\rho = 1$ correspond to a ball and a line segment. The behavior is as expected (see Table A.6) and can be summarized as follows: Fixing the remaining parameters,

- $\delta^*$ is increasing in $\eta$,
- $\delta^*$ is first decreasing then increasing in $\rho$, and this behavior depends on the value of $\eta$ and the increasing part may degenerate ($\delta^*$ is decreasing in $\rho$ in that case),
- $\delta^*$ is decreasing in $q$.

Here, the behaviors are more complicated then in the previous examples. A small decrease of $\eta$ (from 1.1 to 1) leads to strong decrease of $\delta^*$. On the other hand, a significant increase of $\rho$ (0 to 0.5) leads to relatively small decrease of $\delta^*$. The small change in $\eta$ hardly detectable visually, while changes of $\rho$ are more obvious. Thus the variance parameter $\eta$ plays more important role than correlation parameter $\rho$ in this setting.

For fixed $\eta$, when $\rho$ increases from 0 to 1, the circles are squeezed to be ellipses and finally segments. During this process, the overlapping parameter of rings and ellipse decreases on the direction of major axis of the ellipse and increases on the direction of minor axis. Such trade-off between two directions explains the unimodal change of $\delta^*$ with respect to $\rho$.

**5.1.2. Kernel PCA case. Equal Mixture of two Gaussians**. In this part, we use the same mixture models as in weighted Laplacian case with two Gaussians with covariance matrices the identity, centered at $(0,0)$ and $(\gamma,\gamma)^T$, respectively, and use the following kernel

$$k(x,y) = \frac{1}{\sqrt{\pi h}} \exp\left\{-\frac{(x-y)^2}{h}\right\}.$$

We consider all the combinations of $\gamma = 1,2,3,5$ and $h = 2,5,10$. Again, suppose the other parameters are fixed, we can find the following relationships (see Table A.8):

- $\delta^*$ is decreasing in $\gamma$,
- $\delta^*$ is increasing in $h$,
- $\delta^*$ is decreasing in $\sigma$.

All similar examples discussed above in the context of the weighted Laplacian give a similar behavior of $\delta^*$. Note also that the bandwidth parameter $h$ in the Kernel PCA case plays a similar role as the reciprocal of the power parameter $q$ in the weighted Laplacian case.

**5.1.3. Summary.** In previous examples, the relationship between the upper bound of coverage ratio $\delta^*$ and the location and the shape of the underlying mixtures can be derived in a straightforward manner, and all the simulation results are as expected. However, it is worth thinking about the influence of the power $q$ (in case of the Laplacian) and the bandwidth $h$ (in case of kernel PCA). Simple monotonic relationships have been observed in the examples discussed above, but these might not always hold. In order to explore the relationships further, we can simply consider the two components case and check the behavior of the (weighted) overlapping parameter, which is highly positive correlated with the coverage ratio $\delta^*$.

In the weighted Laplacian case, our results show that larger $q$ often leads to a larger coverage, i.e. better OCS. However, this only holds when the two components are not too concentrated, i.e. the variance of each component is not too small. We used standard Gaussian distributions with variances equal to 1, so that the maximum value of each component equals $\frac{1}{\sqrt{2\pi}}$. In such case, a large value of $q$ makes the rescaled density flatter and overlapping of each pair of components reduces. This monotone relationship still holds for variances larger than 1, but it might no longer

hold for smaller variances. Suppose the components of previous example has variance $\sigma$, then

$$\begin{aligned}
\mathcal{S}_{12} &= \int \left( \frac{\rho_1(x)\rho_2(x)}{\rho(x)} \right)^q dx \\
&= \left(\frac{2}{\pi}\right)^{\frac{q}{2}} \frac{1}{\sigma^q} \int_{-\infty}^{\infty} \left( \frac{\exp\{-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}(x-\gamma)^2\}}{\exp\{-\frac{1}{2\sigma^2}x^2\} + \exp\{-\frac{1}{2\sigma^2}(x-\gamma)^2\}} \right)^q dx \\
&= \left(\frac{2}{\pi}\right)^{\frac{q}{2}} \sigma^{1-q} \int_{-\infty}^{\infty} \left( \frac{\exp\{-\frac{1}{2}x^2 - \frac{1}{2}(x-\frac{\gamma}{\sigma})^2\}}{\exp\{-\frac{1}{2}x^2\} + \exp\{-\frac{1}{2}(x-\frac{\gamma}{\sigma})^2\}} \right)^q dx,
\end{aligned}$$

which depends on the variance even just for gaussian example. Theoretical analysis and simulation results show that $\mathcal{S}_{12}|_{\sigma=1}$ is decreasing w.r.t $q$, so $\mathcal{S}_{12}$ is decreasing w.r.t $q$ when $\sigma > 1$ but may increase when $\sigma \ll 1$. More generally (not restricted to the Gaussian case), this phenomenon also appears when more concentrated components (peak values of densities are much larger than 1) are included in mixture models. Practically, kernel density estimations of standardized data are usually not too concentrated, (otherwise the choice of parameter is not a necessary agenda since concentrated components are easily clustered with good performance), and in such cases, we tend to choose larger power $q$ to obtain better coverage ratio $\delta^*$.

### 5.2. The inverse problem: Does the OCS contain information about the separateness of the mixture model?

So far, we have explored the OCS of different spectral and kernel embeddings based on mixture models, both theoretically and numerically. One might also ask the question: Can we say something about the separateness of mixture components based on the observed embedding. For instance, if the embedding displays a strong OCS (well separated clusters along orthogonal axis), can we conclude that the mixture components are well separated? Again the explorations are conducted numerically.

Given an embedding, we attempt to numerically estimate the OCS. We do this by first fixing a coverage proportion and then finding orthogonal cones achieving this coverage with opening angles as small as possible. To simplify the computations somewhat the approach taken here is as follows:

Given embedded data $y_1, y_2, \cdots, y_n \in \mathbb{R}^N (N \geq 2)$, we first find an "optimal" orthogonal basis $e_1^*, e_2^*, \cdots, e_N^*$ (or an optimal rotation) by minimizing the criterion

$$\sum_{i=1}^{N} \left( \min_{j=1,2,\cdots,N} \frac{\langle y_i, e_j \rangle}{\|y_i\| \|e_j\|} \right)$$

over all possible orthogonal basis. Then, given $e_1^*, e_2^*, \cdots, e_N^*$, and given a desired coverage proportion $1 - \delta^*$, we find the smallest angle $\sigma^*$ such that at least $(1 - \delta) \times 100\%$ of the data are covered by the orthogonal cones with exes $e_1^*, e_2^*, \cdots, e_N^*$ and opening angle $\sigma^*$. Here, for computational reasons, we fix the angle to be the same for all cones. The angle $\sigma^*$ then serves as an observed measure of the quality of the OCS. We found this to be a good compromise between computational complexity and measuring the OCS.

We then applied this approach to the examples discussed above. It turns out that our measure of quality of the OCS shows some relation to the choice of the number of clusters. Indeed, the angle $\sigma^*$ always was clearly the smallest when the number of clusters was chosen correctly. (It should perhaps be noted in this context that in our theoretical results, the OCS always was based on the correct number $N$ of mixture components.)

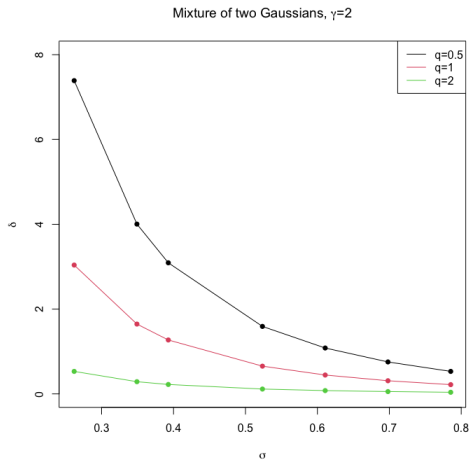
In order to illustrate the useful contribution based on OCS, let's think of a traditional clustering problem: Given data $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$ and assume that the true number of clusters is $N$. In practice, one needs to choose the embedding dimension ($k$) and the number of clusters ($M$). We expect to choose the correct number of clusters ($M = N$) and hope that the choice of $k$ is good enough to get reasonable clustering result. Now if original data is generated from a well-separated mixture model, previous idea of how to choose the correct number of clusters can be further explained as follows:

- If we choose correct number of clusters (i.e., $M = N$), then the major theorems can be applied and the $M$ concentrated cones give good clustering result.

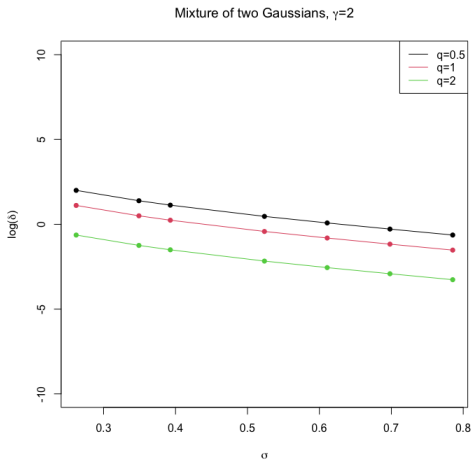
- If we overestimate the number of clusters (i.e., $M > N$), then we can also apply the major theorems in an $N$-dimensional subspace and get $N$ orthogonal concentrated cones in that subspace.

- If we underestimate the number of clusters (i.e., $M < N$), then there exists a set of rotation angles $\beta_1, \beta_2, \cdots, \beta_{\binom{M}{2}}$ such that under the rotated space, there are $M$ cones concentrated around new $M$ axes which cover most parts of original $M$ clusters. Most parts of all the embeddings based on the other $N - M$ clusters are concentrated around the origin. (They can be seen as projection from higher dimensional embedding space.)

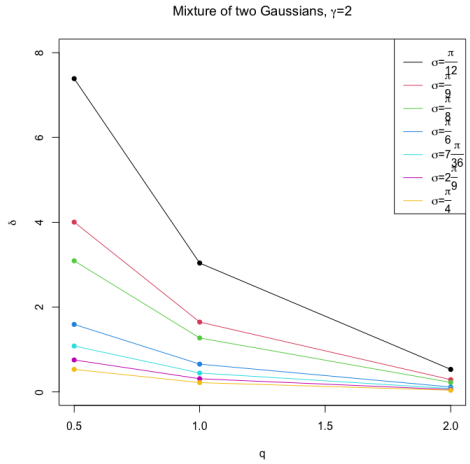
The first two cases are straightforward and the third one can be proved by the similar ideas of the major theorems and can be explained heuristically that $M$-dimensional embedding can be also treated as a further projection of $N$-dimensional embedding. Since the OCS of the latter case was proved, as long as we project $N$-dimensional embedding based on the direction generated by the axes that cones are concentrated around, then $M$ clusters (out of $N$) are still concentrated around given axes and other $N - M$ clusters are concentrated around the origin. This set of conclusions informs a practical way to choose the number of clusters: Find embeddings in dimensions $k = 1, \ldots, N_1$ for some upper bound $N_1$ (to be specified). For each $k$, consider two criteria: The concentration of embeddings around the origin, and the performance measure $\sigma^*$ from above. The goal is to find a $k$ for which simultaneously, $\sigma^*$ is small, and also the concentration of embeddings around the origin is neither high nor small. We have not yet developed an explicit practical criterion based on these ideas. This will be addressed in future work.



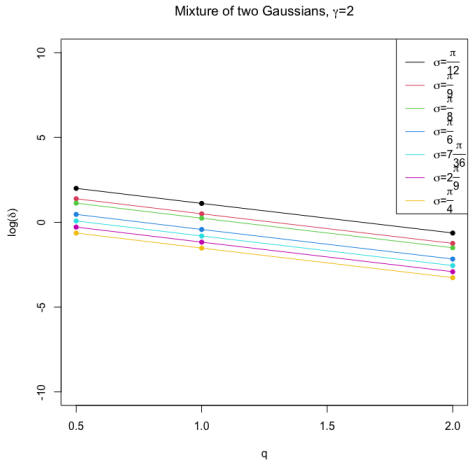

(a) $\delta \sim \sigma$



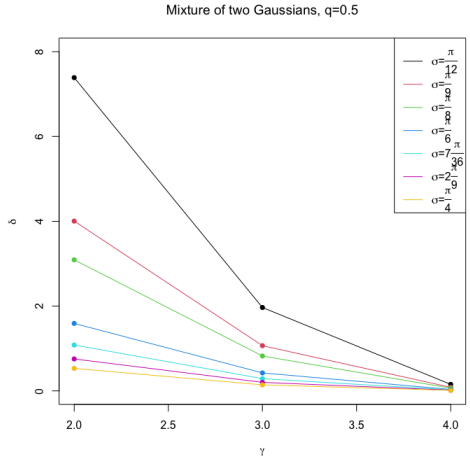

(b) $\log(\delta) \sim \sigma$



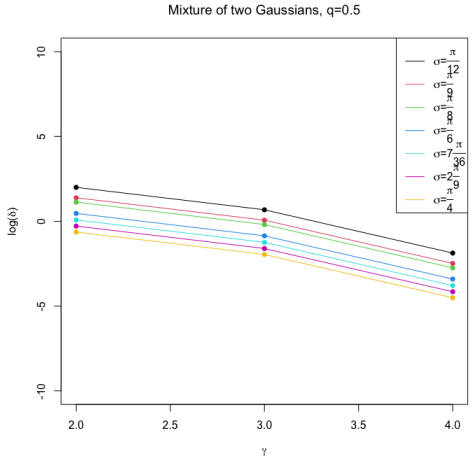

(c) $\delta \sim q$




(d) $\log(\delta) \sim q$




(e) $\delta \sim \gamma$




(f) $\log(\delta) \sim \gamma$

**Figure 5.1.** Behavior of $\delta$ with respect to parameters $\sigma$, $q$ and $\gamma$ in the Equal mixture of Two Gaussians case.

CHAPTER 6

# Conclusions and future work

We proved orthogonal cone structure (OCS) in two low-dimensional embeddings including weighted Laplacian embedding and Kernel PCA embedding, in both population case and sample case. Moreover, precise definition of mixture model and the well-separation property with similarity parameter, coupling parameter, indivisibility parameter and eigen-tail parameter are given. Angles, coverage and radius of OCS were also explored based on theoretical analysis and simulation study. Our results about OCS are useful in clustering algorithm and especially guarantee the correct ratio of clustering result based on $k$-means algorithm. Some theoretical examples about mixture of Gaussians or Uniforms are checked in both weighted Laplacian case and Kernel PCA case and the behavior of OCS w.r.t the change of corresponding tunning parameters (e.g. power parameter $q$ in weighted Laplacian case and bandwidth parameter $h$ in Kernel PCA case) are also illustrated. Detailed proof of all four cases are given, where rescaled densities and corresponding induced measure play essential role in the population setting and spectral convergence is the key part in the sample setting. Meanwhile, Control of interpolation errors and discretization errors are important under weighted Laplacian case while basic concepts of RKHS and related norm bounds are frequently used under Kernel PCA case. All theoretical statements are checked based on simulation study and the effects of parameters are explored in the same part. Finally, reverse problem about the inference of original model or data based on embedded data is explored. By choosing optimal rotation matrix in the embedded space, people can find the best axes that most embedded data are concentrated around, and thus it is helpful about the choice of correct number of clusters.

As stated at the beginning, OCS is observed in many different dimensional reduction embeddings, and there also exist some different geometric structures that can be considered. Better geometric summarization with less parameters is one possible direction. There are also many possible subproblems based on the inverse problem stated in the last part. Besides the choice of the number of clusters, can we infer more properties (e.g. well-separation, geometric relative position) of the

original model (data)? Also, the choice of optimal rotation matrix is stable only when original clusters are well-separated, it is worth exploring the case when original clusters are somehow mixed. We also hope to extend our results to more general groups of operators. Graph Laplacian operator, weighted Laplacian operator and p-Laplacian operator are differential operators while Kernel PCA operator and Kernel CCA operator are integral operators. The proof techniques and ideas are different in these two groups but are similar within each group. A latent set of conditions of these operators (respectively in both two groups) could be summarized and generalized to most kinds of operators. Angle parameter and coverage ratio parameter can also be considered as new criteria to evaluate the performance of clustering and classification, which requires the prerequisite step of density estimation and need to be improved in both aspects of methodology and computational efficiency. The last but not the least, real data application of OCS should also be considered, including improving the clustering results, choosing best tuning parameters, quantifying the true cluster numbers, and, most importantly, explaining the real data in an appropriate manner.

APPENDIX A

# Behaviors of incorrect coverage ratio with respect to some selected parameters

In this appendix, we put all examples about the behaviors of incorrect coverage ratio $\delta^*$ and other parameters. The first example about mixture of two gaussians in the weighted Laplacian setting is used to plot the Figure 5.1. Other tables show similar information for some different parameters in other cases.

| $\gamma$ | $q$ | $\sigma = \frac{\pi}{12}$ | $\sigma = \frac{\pi}{9}$ | $\sigma = \frac{\pi}{8}$ | $\sigma = \frac{\pi}{6}$ | $\sigma = \frac{7\pi}{36}$ | $\sigma = \frac{2\pi}{9}$ | $\sigma = \frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.5 | 7.3879 | 4.0040 | 3.0915 | 1.5913 | 1.0819 | 0.7533 | 0.5304 |
| 2 | 1 | 3.0384 | 1.6467 | 1.2715 | 0.6544 | 0.4449 | 0.3098 | 0.2181 |
| 2 | 2 | 0.5304 | 0.2875 | 0.2220 | 0.1142 | 0.0777 | 0.0541 | 0.0381 |
| 3 | 0.5 | 1.9666 | 1.0658 | 0.8229 | 0.4236 | 0.2880 | 0.2005 | 0.1412 |
| 3 | 1 | 0.5117 | 0.2773 | 0.2141 | 0.1102 | 0.0749 | 0.0522 | 0.0367 |
| 3 | 2 | 0.0405 | 0.0219 | 0.0169 | 0.0087 | 0.0059 | 0.0041 | 0.0029 |
| 4 | 0.5 | 0.1535 | 0.0832 | 0.0642 | 0.0331 | 0.0225 | 0.0157 | 0.0110 |
| 4 | 1 | 0.0112 | 0.0061 | 0.0047 | 0.0024 | 0.0016 | 0.0011 | 0.0008 |
| 4 | 2 | 0.0001 | 0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |

**Table A.1.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of two Gaussians in the weighted Laplacian case. $\gamma$: off-set parameter, $q$: power parameter, $\sigma$: angle parameter. The last column is the average $\delta^*$ for each row, the same for the following tables.

| $\gamma$ | $q$ | $\sigma = \frac{\pi}{12}$ | $\sigma = \frac{\pi}{9}$ | $\sigma = \frac{\pi}{8}$ | $\sigma = \frac{\pi}{6}$ | $\sigma = \frac{7\pi}{36}$ | $\sigma = \frac{2\pi}{9}$ | $\sigma = \frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.5 | 16.2891 | 8.8282 | 6.8164 | 3.5085 | 2.3853 | 1.6610 | 1.1695 |
| 2 | 1 | 4.6312 | 2.5100 | 1.9380 | 0.9975 | 0.6782 | 0.4723 | 0.3325 |
| 2 | 2 | 2.6981 | 1.4623 | 1.1290 | 0.5811 | 0.3951 | 0.2751 | 0.1937 |
| 3 | 0.5 | 8.9057 | 4.8266 | 3.7267 | 1.9182 | 1.3041 | 0.9081 | 0.6394 |
| 3 | 1 | 2.0203 | 1.0949 | 0.8454 | 0.4352 | 0.2958 | 0.2060 | 0.1450 |
| 3 | 2 | 0.7833 | 0.4245 | 0.3278 | 0.1687 | 0.1147 | 0.0799 | 0.0562 |
| 4 | 0.5 | 3.5421 | 1.9197 | 1.4823 | 0.7629 | 0.5187 | 0.3612 | 0.2543 |
| 4 | 1 | 0.6532 | 0.3540 | 0.2733 | 0.1407 | 0.0956 | 0.0666 | 0.0469 |
| 4 | 2 | 0.1908 | 0.1034 | 0.0798 | 0.0411 | 0.0279 | 0.0195 | 0.0137 |

**Table A.2.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of two Gaussians with different variances. $\gamma$: off-set parameter, $q$: power parameter, $\sigma$: angle parameter.

| $a$ | $b$ | $q$ | $\sigma=\frac{\pi}{12}$ | $\sigma=\frac{\pi}{9}$ | $\sigma=\frac{\pi}{8}$ | $\sigma=\frac{\pi}{6}$ | $\sigma=\frac{7\pi}{36}$ | $\sigma=\frac{2\pi}{9}$ | $\sigma=\frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.5 | 1.5 | 0.5 | 12.6117 | 6.8351 | 5.2775 | 2.7164 | 1.8468 | 1.2860 | 0.9055 |
| 0.5 | 1.5 | 1 | 6.0459 | 3.2767 | 2.5300 | 1.3022 | 0.8853 | 0.6165 | 0.4341 |
| 0.5 | 1.5 | 2 | 1.7028 | 0.9228 | 0.7125 | 0.3668 | 0.2493 | 0.1736 | 0.1223 |
| 0.5 | 1.2 | 0.5 | 19.7555 | 10.7069 | 8.2669 | 4.2552 | 2.8929 | 2.0145 | 1.4184 |
| 0.5 | 1.2 | 1 | 8.2104 | 4.4498 | 3.4357 | 1.7684 | 1.2023 | 0.8372 | 0.5895 |
| 0.5 | 1.2 | 2 | 3.4436 | 1.8663 | 1.4410 | 0.7417 | 0.5043 | 0.3511 | 0.2472 |
| 0.8 | 1.5 | 0.5 | 8.7774 | 4.7571 | 3.6730 | 1.8906 | 1.2853 | 0.8950 | 0.6302 |
| 0.8 | 1.5 | 1 | 2.8662 | 1.5534 | 1.1994 | 0.6173 | 0.4197 | 0.2923 | 0.2058 |
| 0.8 | 1.5 | 2 | 0.8954 | 0.4853 | 0.3747 | 0.1929 | 0.1311 | 0.0913 | 0.0643 |
| 0.8 | 1.2 | 0.5 | 14.4122 | 7.8110 | 6.0310 | 3.1043 | 2.1105 | 1.4696 | 1.0348 |
| 0.8 | 1.2 | 1 | 4.3795 | 2.3735 | 1.8326 | 0.9433 | 0.6413 | 0.4466 | 0.3144 |
| 0.8 | 1.2 | 2 | 2.1962 | 1.1903 | 0.9190 | 0.4730 | 0.3216 | 0.2240 | 0.1577 |

**Table A.3.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of two Uniforms. $a$&$b$: support parameters (left endpoint and right endpoint), $q$: power parameter, $\sigma$: angle parameter.

| $a$ | $b$ | $q$ | $\sigma=\frac{\pi}{12}$ | $\sigma=\frac{\pi}{9}$ | $\sigma=\frac{\pi}{8}$ | $\sigma=\frac{\pi}{6}$ | $\sigma=\frac{7\pi}{36}$ | $\sigma=\frac{2\pi}{9}$ | $\sigma=\frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.5 | 23.8397 | 12.9203 | 9.9760 | 5.1348 | 3.4910 | 2.4310 | 1.7116 |
| 0 | 1 | 1 | 5.4092 | 2.9316 | 2.2636 | 1.1651 | 0.7921 | 0.5516 | 0.3883 |
| 0 | 1 | 2 | 6.5453 | 3.5473 | 2.7390 | 1.4098 | 0.9585 | 0.6674 | 0.4699 |
| 1 | 2 | 0.5 | 10.2587 | 5.5599 | 4.2929 | 2.2096 | 1.5023 | 1.0461 | 0.7365 |
| 1 | 2 | 1 | 1.3345 | 0.7232 | 0.5584 | 0.2874 | 0.1954 | 0.1361 | 0.0958 |
| 1 | 2 | 2 | 0.6536 | 0.3542 | 0.2735 | 0.1408 | 0.0957 | 0.0666 | 0.0469 |
| 2 | 3 | 0.5 | 0.6977 | 0.3781 | 0.2920 | 0.1503 | 0.1022 | 0.0711 | 0.0501 |
| 2 | 3 | 1 | 0.0181 | 0.0098 | 0.0076 | 0.0039 | 0.0027 | 0.0018 | 0.0013 |
| 2 | 3 | 2 | 0.0005 | 0.0003 | 0.0002 | 0.0001 | <0.0001 | <0.0001 | <0.0001 |

**Table A.4.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of a Gaussian and a Uniform. $a$&$b$: support parameters (left endpoint and right endpoint), $q$: power parameter, $\sigma$: angle parameter.

| $\gamma_1$ | $\gamma_2$ | $q$ | $\sigma=\frac{\pi}{12}$ | $\sigma=\frac{\pi}{9}$ | $\sigma=\frac{\pi}{8}$ | $\sigma=\frac{\pi}{6}$ | $\sigma=\frac{7\pi}{36}$ | $\sigma=\frac{2\pi}{9}$ | $\sigma=\frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 0.5 | 24.6104 | 13.3380 | 10.2985 | 5.3008 | 3.6039 | 2.5096 | 1.7669 |
| 2 | 2 | 1 | 6.7474 | 3.6569 | 2.8235 | 1.4533 | 0.9881 | 0.6880 | 0.4844 |
| 2 | 2 | 2 | 2.9252 | 1.5854 | 1.2241 | 0.6301 | 0.4284 | 0.2983 | 0.2100 |
| 2 | 3 | 0.5 | 17.6214 | 9.5502 | 7.3739 | 3.7955 | 2.5804 | 1.7969 | 1.2652 |
| 2 | 3 | 1 | 4.8336 | 2.6197 | 2.0227 | 1.0411 | 0.7078 | 0.4929 | 0.3470 |
| 2 | 3 | 2 | 2.1205 | 1.1492 | 0.8874 | 0.4567 | 0.3105 | 0.2162 | 0.1522 |
| 3 | 2 | 0.5 | 18.5026 | 10.0278 | 7.7426 | 3.9853 | 2.7095 | 1.8867 | 1.3284 |
| 3 | 2 | 1 | 4.9073 | 2.6596 | 2.0535 | 1.0570 | 0.7186 | 0.5004 | 0.3523 |
| 3 | 2 | 2 | 2.0014 | 1.0847 | 0.8375 | 0.4311 | 0.2931 | 0.2041 | 0.1437 |
| 3 | 3 | 0.5 | 12.4455 | 6.7450 | 5.2080 | 2.6806 | 1.8225 | 1.2691 | 0.8935 |
| 3 | 3 | 1 | 2.9966 | 1.6241 | 1.2540 | 0.6454 | 0.4388 | 0.3056 | 0.2151 |
| 3 | 3 | 2 | 0.9642 | 0.5225 | 0.4035 | 0.2077 | 0.1412 | 0.0983 | 0.0692 |

**Table A.5.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of three Gaussians. $\gamma_1$: off-set parameter between component 1 and 2, $\gamma_2$: off-set parameter between component 2 and 3, $q$: power parameter, $\sigma$: angle parameter.

| $r$ | $\eta$ | $q$ | $\sigma = \frac{\pi}{12}$ | $\sigma = \frac{\pi}{9}$ | $\sigma = \frac{\pi}{8}$ | $\sigma = \frac{\pi}{6}$ | $\sigma = \frac{7\pi}{36}$ | $\sigma = \frac{2\pi}{9}$ | $\sigma = \frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 5.5 | 1 | 0.5 | 1.0348 | 0.5608 | 0.4330 | 0.2229 | 0.1515 | 0.1055 | 0.0743 |
| 5.5 | 1 | 1 | 0.0654 | 0.0354 | 0.0274 | 0.0141 | 0.0096 | 0.0067 | 0.0047 |
| 5.5 | 1 | 2 | 0.0095 | 0.0051 | 0.0040 | 0.0020 | 0.0014 | 0.0010 | 0.0007 |
| 4.5 | 1 | 0.5 | 5.5329 | 2.9987 | 2.3153 | 1.1917 | 0.8102 | 0.5642 | 0.3972 |
| 4.5 | 1 | 1 | 0.7821 | 0.4239 | 0.3273 | 0.1685 | 0.1145 | 0.0798 | 0.0562 |
| 4.5 | 1 | 2 | 0.3225 | 0.1748 | 0.1350 | 0.0695 | 0.0472 | 0.0329 | 0.0232 |
| 5.5 | 2 | 0.5 | 8.2842 | 4.4898 | 3.4666 | 1.7843 | 1.2131 | 0.8448 | 0.5948 |
| 5.5 | 2 | 1 | 1.2891 | 0.6986 | 0.5394 | 0.2777 | 0.1888 | 0.1314 | 0.0926 |
| 5.5 | 2 | 2 | 1.1531 | 0.6249 | 0.4825 | 0.2484 | 0.1689 | 0.1176 | 0.0828 |
| 4.5 | 2 | 0.5 | 13.5518 | 7.3446 | 5.6709 | 2.9189 | 1.9845 | 1.3819 | 0.9730 |
| 4.5 | 2 | 1 | 2.5205 | 1.3660 | 1.0547 | 0.5429 | 0.3691 | 0.2570 | 0.1810 |
| 4.5 | 2 | 2 | 2.5701 | 1.3929 | 1.0755 | 0.5536 | 0.3764 | 0.2621 | 0.1845 |

**Table A.6.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of an annulus and a ball inside the annulus. $r$: radius parameter, $\eta$: thickness parameter, $q$: power parameter, $\sigma$: angle parameter.

| $\eta$ | $\rho$ | $q$ | $\sigma = \frac{\pi}{12}$ | $\sigma = \frac{\pi}{9}$ | $\sigma = \frac{\pi}{8}$ | $\sigma = \frac{\pi}{6}$ | $\sigma = \frac{7\pi}{36}$ | $\sigma = \frac{2\pi}{9}$ | $\sigma = \frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.5 | 2.4691 | 1.3382 | 1.0332 | 0.5318 | 0.3616 | 0.2518 | 0.1773 |
| 1 | 0 | 1 | 0.3423 | 0.1855 | 0.1432 | 0.0737 | 0.0501 | 0.0349 | 0.0246 |
| 1 | 0 | 2 | 0.1236 | 0.0670 | 0.0517 | 0.0266 | 0.0181 | 0.0126 | 0.0089 |
| 1 | 0.5 | 0.5 | 2.0050 | 1.0866 | 0.8390 | 0.4319 | 0.2936 | 0.2044 | 0.1440 |
| 1 | 0.5 | 1 | 0.2778 | 0.1505 | 0.1162 | 0.0598 | 0.0407 | 0.0283 | 0.0199 |
| 1 | 0.5 | 2 | 0.1159 | 0.0628 | 0.0485 | 0.0250 | 0.0170 | 0.0118 | 0.0083 |
| 1.1 | 0 | 0.5 | 3.8606 | 2.0923 | 1.6155 | 0.8315 | 0.5653 | 0.3937 | 0.2772 |
| 1.1 | 0 | 1 | 0.8175 | 0.4431 | 0.3421 | 0.1761 | 0.1197 | 0.0834 | 0.0587 |
| 1.1 | 0 | 2 | 0.5019 | 0.2720 | 0.2100 | 0.1081 | 0.0735 | 0.0512 | 0.0360 |
| 1.1 | 0.5 | 0.5 | 3.1077 | 1.6843 | 1.3005 | 0.6694 | 0.4551 | 0.3169 | 0.2231 |
| 1.1 | 0.5 | 1 | 0.6123 | 0.3318 | 0.2562 | 0.1319 | 0.0897 | 0.0624 | 0.0440 |
| 1.1 | 0.5 | 2 | 0.4156 | 0.2252 | 0.1739 | 0.0895 | 0.0609 | 0.0424 | 0.0298 |

**Table A.7.** Behavior of incorrect coverage ratio $\delta^*$ for mixture of two ellipses and the figure shaped as '$\infty$'. $\eta$: variance parameter, $\rho$: correlation parameter, $q$: power parameter, $\sigma$: angle parameter.

| $\gamma$ | $h$ | $\sigma = \frac{\pi}{12}$ | $\sigma = \frac{\pi}{9}$ | $\sigma = \frac{\pi}{8}$ | $\sigma = \frac{\pi}{6}$ | $\sigma = \frac{7\pi}{36}$ | $\sigma = \frac{2\pi}{9}$ | $\sigma = \frac{\pi}{4}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3.3694 | 2.4805 | 2.1796 | 1.5637 | 1.2894 | 1.0759 | 0.9028 |
| 1 | 5 | 3.6354 | 2.6763 | 2.3517 | 1.6872 | 1.3912 | 1.1609 | 0.9741 |
| 1 | 10 | 3.7253 | 2.7425 | 2.4099 | 1.7289 | 1.4256 | 1.1896 | 0.9982 |
| 2 | 2 | 1.8813 | 1.3850 | 1.2170 | 0.8731 | 0.7199 | 0.6008 | 0.5041 |
| 2 | 5 | 2.7412 | 2.0180 | 1.7732 | 1.2722 | 1.0490 | 0.8753 | 0.7345 |
| 2 | 10 | 3.3239 | 2.4470 | 2.1502 | 1.5426 | 1.2720 | 1.0614 | 0.8906 |
| 3 | 2 | 0.6008 | 0.4423 | 0.3887 | 0.2789 | 0.2299 | 0.1919 | 0.1610 |
| 3 | 5 | 1.3502 | 0.9940 | 0.8734 | 0.6266 | 0.5167 | 0.4312 | 0.3618 |
| 3 | 10 | 2.3288 | 1.7144 | 1.5065 | 1.0808 | 0.8912 | 0.7436 | 0.6240 |
| 5 | 2 | 0.0083 | 0.0061 | 0.0054 | 0.0039 | 0.0032 | 0.0027 | 0.0022 |
| 5 | 5 | 0.0717 | 0.0528 | 0.0464 | 0.0333 | 0.0274 | 0.0229 | 0.0192 |
| 5 | 10 | 0.4013 | 0.2954 | 0.2596 | 0.1862 | 0.1536 | 0.1281 | 0.1075 |

**Table A.8.** Behavior of incorrect coverage ratio $\delta^*$ for equal mixture of two Gaussians in the kernel PCA case. $\gamma$: off-set parameter, $h$: bandwidth parameter, $\sigma$: angle parameter.

# Bibliography


[1] S. Agarwal, K. Branson, and S. Belongie, *Higher order learning with graphs*, in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 17–24.

[2] S. ALBEVERIO, S. KUZHEL, and L. P. NIZHNIK, *On the perturbation theory of self-adjoint operators*, Tokyo Journal of Mathematics, 31 (2008), pp. 273–292.

[3] R. G. Antonini, *Subgaussian random variables in hilbert spaces*, Rendiconti del Seminario Matematico della Università di Padova, 98 (1997), pp. 89–99.

[4] B. Aragam, C. Dan, E. P. Xing, and P. Ravikumar, *Identifiability of nonparametric mixture models and bayes optimal clustering*, The Annals of Statistics, 48 (2020), pp. 2277–2302.

[5] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American mathematical society, 68 (1950), pp. 337–404.

[6] A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman, *A limit theorem for scaled eigenvectors of random dot product graphs*, Sankhya A, 78 (2016), pp. 1–18.

[7] A. AZZALINI and M. G. GENTON, *On gauss's characterization of the normal distribution*, Bernoulli, 13 (2007), pp. 169–174.

[8] K. Balasubramanian, T. Li, and M. Yuan, *On the optimality of kernel-embedding based goodness-of-fit tests.*, J. Mach. Learn. Res., 22 (2021), pp. 1–1.

[9] L. Bartholdi, T. Schick, N. Smale, and S. Smale, *Hodge theory on metric spaces*, Foundations of Computational Mathematics, 12 (2012), pp. 1–48.

[10] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, *Networks beyond pairwise interactions: structure and dynamics*, Physics Reports, 874 (2020), pp. 1–92.

[11] M. Belkin and P. Niyogi, *Towards a theoretical foundation for laplacian-based manifold methods.*, in COLT, vol. 3559, Springer, 2005, pp. 486–500.

[12] S. Ben-David, U. Von Luxburg, and D. Pál, *A sober look at clustering stability*, in Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19, Springer, 2006, pp. 5–19.

[13] Y. Bengio, O. Delalleau, N. L. Roux, J.-F. Paiement, P. Vincent, and M. Ouimet, *Learning eigenfunctions links spectral embedding and kernel pca*, Neural computation, 16 (2004), pp. 2197–2219.

[14] Y. Bengio, J.-f. Paiement, P. Vincent, O. Delalleau, N. Roux, and M. Ouimet, *Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering*, Advances in neural information processing systems, 16 (2003).

[15] A. R. Benson, D. F. Gleich, and J. Leskovec, *Tensor spectral clustering for partitioning higher-order network structures*, in Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 118–126.

[16] P. H. Bérard, *Spectral geometry: direct and inverse problems*, vol. 1207, Springer, 2006.

[17] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*, Springer Science & Business Media, 2011.

[18] G. Blanchard, O. Bousquet, and L. Zwald, *Statistical properties of kernel principal component analysis*, Machine Learning, 66 (2007), pp. 259–294.

[19] H. R. Brown, W. Myrvold, and J. Uffink, *Boltzmann's h-theorem, its discontents, and the birth of statistical mechanics*, Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics, 40 (2009), pp. 174–191.

[20] T. Bühler and M. Hein, *Spectral clustering based on the graph p-laplacian*, in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 81–88.

[21] D. Burago, S. Ivanov, and Y. Kurylev, *A graph discretization of the laplace–beltrami operator*, Journal of Spectral Theory, 4 (2015), pp. 675–714.

[22] D. R. Burt, *Spectral methods in Gaussian process approximations*, PhD thesis, Master's thesis, University of Cambridge, 2018.(Cited on pages v, ix, 26, 59 . . . , 2018.

[23] T. T. Cai and A. Zhang, *Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics*, The Annals of Statistics, 46 (2018), pp. 60–89.

[24] J. Calder and N. G. Trillos, *Improved spectral convergence rates for graph laplacians on $\varepsilon$-graphs and k-nn graphs*, Applied and Computational Harmonic Analysis, 60 (2022), pp. 123–175.

[25] Y. Canzani, *Analysis on manifolds via the laplacian*, Lecture Notes available at: http://www. math. harvard. edu/canzani/docs/Laplacian. pdf, (2013).

[26] J. Cape, M. Tang, and C. E. Priebe, *Signal-plus-noise matrix models: eigenvector deviations and fluctuations*, Biometrika, 106 (2019), pp. 243–250.

[27] A. Celisse and M. Wahl, *Analyzing the discrepancy principle for kernelized spectral filter learning algorithms*, Journal of Machine Learning Research, 22 (2021), pp. 1–59.

[28] F. Chatelin, *Spectral approximation of linear operators*, SIAM, 2011.

[29] F. R. Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[30] R. R. Coifman and S. Lafon, *Diffusion maps*, Applied and computational harmonic analysis, 21 (2006), pp. 5–30.

[31] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. iii*, SIAM Journal on Numerical Analysis, 7 (1970), pp. 1–46.

[32] E. De Vito, N. Mücke, and L. Rosasco, *Reproducing kernel hilbert spaces on manifolds: Sobolev and diffusion spaces*, Analysis and Applications, 19 (2021), pp. 363–396.

[33] H. Edelsbrunner and J. L. Harer, *Computational topology: an introduction*, American Mathematical Society, 2022.

[34] G. E. Fasshauer, *Positive definite kernels: past, present and future*, Dolomites Research Notes on Approximation, 4 (2011), pp. 21–63.

[35] J. Ferreira and V. Menegatto, *Eigenvalue decay rates for positive integral operators*, Annali di Matematica Pura ed Applicata, 192 (2013), pp. 1025–1041.

[36] K. Fukumizu, F. R. Bach, and A. Gretton, *Statistical consistency of kernel canonical correlation analysis.*, Journal of Machine Learning Research, 8 (2007).

[37] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator*, Foundations of Computational Mathematics, 20 (2020), pp. 827–887.

[38] R. G. Ghanem and P. D. Spanos, *Stochastic finite elements: a spectral approach*, Courier Corporation, 2003.

[39] M. E. Gier, *Eigenvalue multiplicites of the Hodge Laplacian on coexact 2-forms for generic metrics on 5-manifolds*, University of Kentucky, 2014.

[40] E. Giné and V. Koltchinskii, *Empirical graph laplacian approximation of laplace-beltrami operators: large sample results*, Lecture Notes-Monograph Series, (2006), pp. 238–259.

[41] G. Giorgobiani, V. Kvaratskhelia, and V. Tarieladze, *Notes on sub-gaussian random elements*, in Applications of Mathematics and Informatics in Natural Sciences and Engineering: AMINSE 2019, Tbilisi, Georgia, September 23-26, Springer, 2020, pp. 197–203.

[42] T. E. Goldberg, *Combinatorial laplacians of simplicial complexes*, Senior Thesis, Bard College, 6 (2002).

[43] K. W. Govek, V. S. Yamajala, and P. G. Camara, *Spectral simplicial theory for feature selection and applications to genomics*, arXiv preprint arXiv:1811.03377, (2018).

[44] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, *A kernel view of the dimensionality reduction of manifolds*, in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 47.

[45] M. Hein, J.-Y. Audibert, and U. v. Luxburg, *Graph laplacians and their convergence on random neighborhood graphs.*, Journal of Machine Learning Research, 8 (2007).

[46] M. Hein, J.-Y. Audibert, and U. Von Luxburg, *From graphs to manifolds-weak and strong pointwise consistency of graph laplacians.*, in COLT, vol. 3559, Springer, 2005, pp. 470–485.

[47] F. Hoffmann, B. Hosseini, A. A. Oberai, and A. M. Stuart, *Spectral analysis of weighted laplacians arising in data clustering*, Applied and Computational Harmonic Analysis, 56 (2022), pp. 189–249.

[48] H. Hoffmann, *Kernel pca for novelty detection*, Pattern recognition, 40 (2007), pp. 863–874.

[49] D. Horak and J. Jost, *Spectra of combinatorial laplace operators on simplicial complexes*, Advances in Mathematics, 244 (2013), pp. 303–336.

[50] T. Hsing and R. Eubank, *Theoretical foundations of functional data analysis, with an introduction to linear operators*, vol. 997, John Wiley & Sons, 2015.

[51] A. J. Izenman, *Introduction to manifold learning*, Wiley Interdisciplinary Reviews: Computational Statistics, 4 (2012), pp. 439–446.

[52] M. Jirak and M. Wahl, *Relative perturbation bounds with applications to empirical covariance operators*, arXiv preprint arXiv:1802.02869, (2018).

[53] ———, *Perturbation bounds for eigenspaces under a relative gap condition*, Proceedings of the American Mathematical Society, 148 (2020), pp. 479–494.

[54] I. Jolliffe, *Generalizations and adaptations of principal component analysis*, in Principal Component Analysis, Springer, 1986, pp. 223–234.

[55] T. Kato, *Perturbation theory for linear operators.— springer-verlag: Berlin*, Heidelberg, New York, (1980).

[56] T. Kato, *Perturbation theory for linear operators*, vol. 132, Springer Science & Business Media, 2013.

[57] V. Koltchinskii and E. Giné, *Random matrix approximation of spectra of integral operators*, Bernoulli, (2000), pp. 113–167.

[58] V. Koltchinskii, M. Löffler, and R. Nickl, *Efficient estimation of linear functionals of principal components*, The Annals of Statistics, 48 (2020), pp. 464–490.

[59] V. Koltchinskii and K. Lounici, *Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance*, arXiv preprint arXiv:1408.4643, (2014).

[60] ———, *Concentration inequalities and moment bounds for sample covariance operators*, Bernoulli, 23 (2017), pp. 110–133.

[61] ———, *New asymptotic results in principal component analysis*, Sankhya A, 79 (2017), pp. 254–297.

[62] J. Lafferty, G. Lebanon, and T. Jaakkola, *Diffusion kernels on statistical manifolds.*, Journal of Machine Learning Research, 6 (2005).

[63] Z. Liang, *Eigen-analysis of kernel operators for nonlinear dimension reduction and discrimination*, The Ohio State University, 2014.

[64] L.-H. Lim, *Hodge laplacians on graphs*, Siam Review, 62 (2020), pp. 685–715.

[65] Z. Meng and K. Xia, *Persistent spectral based machine learning (perspect ml) for drug design*, arXiv preprint arXiv:2002.00582, (2020).

[66] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, *Kernel mean embedding of distributions: A review and beyond*, arXiv preprint arXiv:1605.09522, (2016).

[67] A. Muhammad and M. Egerstedt, *Control using higher order laplacians in network topologies*, in Proc. of 17th International Symposium on Mathematical Theory of Networks and Systems, Citeseer, 2006, pp. 1024–1038.

[68] V. I. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*, vol. 152, Cambridge University Press, 2016.

[69] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, *Cluster canonical correlation analysis*, in Artificial intelligence and statistics, PMLR, 2014, pp. 823–831.

[70] C. E. Rasmussen, C. K. Williams, et al., *Gaussian processes for machine learning*, vol. 1, Springer, 2006.

[71] V. C. Raykar, *Spectral clustering and kernel principal component analysis are pursuing good projections*, Project Report, (2004).

[72] M. Reiss and M. Wahl, *Nonasymptotic upper bounds for the reconstruction error of pca*, The Annals of Statistics, 48 (2020), pp. 1098–1123.

[73] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, *A unifying review of deep and shallow anomaly detection*, Proceedings of the IEEE, 109 (2021), pp. 756–795.

[74] M. Scetbon and Z. Harchaoui, *A spectral analysis of dot-product kernels*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3394–3402.

[75] G. Schiebinger, M. J. Wainwright, B. Yu, et al., *The geometry of kernelized spectral clustering*, The Annals of Statistics, 43 (2015), pp. 819–846.

[76] B. Schölkopf, A. Smola, and K.-R. Müller, *Kernel principal component analysis*, in International conference on artificial neural networks, Springer, 1997, pp. 583–588.

[77] ———, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural computation, 10 (1998), pp. 1299–1319.

[78] ———, *Kernel principal component analysis*, in Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings, Springer, 2005, pp. 583–588.

[79] B. Schölkopf, A. J. Smola, F. Bach, et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[80] S. Seol, M. Stiénon, and P. Xu, *Dg manifolds, formal exponential maps and homotopy lie algebras*, Communications in Mathematical Physics, 391 (2022), pp. 33–76.

[81] T. Shi, M. Belkin, B. Yu, et al., *Data spectroscopy: Eigenspaces of convolution operators and clustering*, The Annals of Statistics, 37 (2009), pp. 3960–3984.

[82] D. Slepcev and M. Thorpe, *Analysis of p-laplacian regularization in semisupervised learning*, SIAM Journal on Mathematical Analysis, 51 (2019), pp. 2085–2120.

[83] B. Sriperumbudur and N. Sterge, *Approximate kernel pca using random features: Computational vs. statistical trade-off*, arXiv preprint arXiv:1706.06296, (2017).

[84] I. Steinwart and A. Christmann, *Support vector machines*, Springer Science & Business Media, 2008.

[85] M. Tang and C. E. Priebe, *Limit theorems for eigenvectors of the normalized laplacian for random graphs*, arXiv preprint arXiv:1607.08601, (2016).

[86] H. Teicher, *Identifiability of finite mixtures*, The annals of Mathematical statistics, (1963), pp. 1265–1269.

[87] D. Ting, L. Huang, and M. Jordan, *An analysis of the convergence of graph laplacians*, arXiv preprint arXiv:1101.5435, (2011).

[88] N. G. Trillos, M. Gerlach, M. Hein, and D. Slepcev, *Error estimates for spectral convergence of the graph laplacian on random geometric graphs towards the laplace–beltrami operator*, arXiv preprint arXiv:1801.10108, (2018).

[89] N. G. Trillos, F. Hoffmann, and B. Hosseini, *Geometric structure of graph laplacian embeddings*, arXiv preprint arXiv:1901.10651, (2019).

[90] N. G. Trillos, F. Hoffmann, and B. Hosseini, *Geometric structure of graph laplacian embeddings*, Journal of Machine Learning Research, 22 (2021), pp. 1–55.

[91] N. G. Trillos and D. Slepčev, *On the rate of convergence of empirical measures in $\infty$-transportation distance*, Canadian Journal of Mathematics, 67 (2015), pp. 1358–1383.

[92] N. G. Trillos and D. Slepčev, *A variational approach to the consistency of spectral clustering*, Applied and Computational Harmonic Analysis, 45 (2018), pp. 239–281.

[93] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[94] U. Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing, 17 (2007), pp. 395–416.

[95] U. Von Luxburg, M. Belkin, and O. Bousquet, *Consistency of spectral clustering*, The Annals of Statistics, (2008), pp. 555–586.

[96] M. Wahl, *On the perturbation series for eigenvalues and eigenprojections*, arXiv preprint arXiv:1910.08460, (2019).

[97] M. Welling, *Kernel canonical correlation analysis*, Depertment of Computer Science University of Toronto, Canada, (2005).

[98] C. Williams, *On a connection between kernel pca and metric multidimensional scaling*, Advances in neural information processing systems, 13 (2000).

[99] T. Wu, A. R. Benson, and D. F. Gleich, *General tensor spectral co-clustering for higher-order data*, Advances in Neural Information Processing Systems, 29 (2016).

[100] Z. Yang, K. Balasubramanian, and H. Liu, *On stein's identity and near-optimal estimation in high-dimensional index models*, arXiv preprint arXiv:1709.08795, (2017).

[101] Y. Yu, T. Wang, and R. J. Samworth, *A useful variant of the davis–kahan theorem for statisticians*, Biometrika, 102 (2015), pp. 315–323.

[102] H. ZHU, C. WILLIAMS, R. ROHWER, and M. MORCINIEC, *Gaussian regression and optimal finite dimensional linear models*, NATO ASI series. Series F: computer and system sciences, (1998), pp. 167–184.