Journal of Machine Learning Research 26 (2025) 1-48 Submitted 11/24; Revised 5/25; Published 6/25

On Linear Separation Capacity of Self-Supervised Representation Learning

Shulei Wang SHULEIW@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA

Editor: Kilian Weinberger

Abstract

Recent advances in self-supervised learning have highlighted the efficacy of data augmentation in learning data representation from unlabeled data. Training a linear model atop these enhanced representations can yield an adept classifier. Despite the remarkable empirical performance, the underlying mechanisms that enable data augmentation to unravel nonlinear data structures into linearly separable representations remain elusive. This paper seeks to bridge this gap by investigating under what conditions learned representations can linearly separate manifolds when data is drawn from a multi-manifold model. Our investigation reveals that data augmentation offers additional information beyond observed data and can thus improve the information-theoretic optimal rate of linear separation capacity. In particular, we show that self-supervised learning can linearly separate manifolds with a smaller distance than unsupervised learning, underscoring the additional benefits of data augmentation. Our theoretical analysis further underscores that the performance of downstream linear classifiers primarily hinges on the linear separability of data representations rather than the size of the labeled data set, reaffirming the viability of constructing efficient classifiers with limited labeled data amid an expansive unlabeled data set.

Keywords: self-supervised learning, data augmentation, linear separation capacity

1. Introduction

1.1 Self-Supervised Representation Learning

The recent advance in large-scale machine learning models demonstrates their superior capacity and performance in various fields (Vaswani et al., 2017), but also demands millions of labeled samples for training, which can be inaccessible in some applications Dosovitskiy et al. (2021). Self-supervised pre-training and transfer learning are introduced to address the challenge of scarce labeled data in natural language processing and computer vision. Unlike the classical supervised learning framework, the current self-supervised learning framework usually involves two steps: self-supervised pre-training and fine-tuning in the downstream task. In the pre-training stage, self-supervised learning first learns data representations/features from a large unlabeled data set and pseudo-labels automatically generated from the unlabeled data (Hjelm et al., 2018; Bachman et al., 2019; Chen et al., 2020b; He et al., 2020; Chen and He, 2021; Zbontar et al., 2021; He et al., 2022). The pre-trained representations are then transferred to train a full model via fine-tuning on a labeled data set in the

©2025 Shulei Wang.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/24-2032.html.

downstream task (Kumar et al., 2022). A comprehensive review is available in Balestriero et al. (2023). Recent progress shows that self-supervised pre-training can reduce the need for external supervision and extract information from a large amount of unlabeled data more efficiently than classical unsupervised learning methods (LeCun, 2022). Due to these beneficial properties, self-supervised learning has been widely used in the pre-training of large foundational models (Zhou et al., 2023). For instance, self-supervised pre-training has played a crucial role in the success of recent large language models, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018, 2019).

Generating pseudo labels from unlabeled data is a pivotal component of self-supervised representation learning and a significant departure from traditional supervised and unsupervised learning methods. Various approaches to pseudo label generation are employed in different applications of self-supervised pre-training, including data masking (Brown et al., 2020; He et al., 2022), data augmentation (Chen et al., 2020b; Zbontar et al., 2021), and multi-modal data fusion (Akbari et al., 2021). These generated pseudo labels serve as supervisory signals in self-supervised representation learning, with the trained representations subsequently fine-tuned for diverse downstream tasks. Notably, empirical observations suggest that linear probing of learned representations, achieved by training a linear model on top of frozen representations, can surpass training all parameters of the same model from scratch (Chen et al., 2020b,c; Kumar et al., 2022). This implies that the representations acquired through self-supervised pre-training exhibit linear separability, enhancing downstream task efficiency via a linear model. The improved linear separability brought by self-supervised pre-training made constructing an efficient classifier in scenarios with limited labeled data possible (Tan et al., 2018; Wang et al., 2020). Given these intriguing empirical findings, questions arise about why training with pseudo labels yields linearly separable representations and what underlying structural information self-supervised representation learning captures through pseudo labels. This paper endeavors to address these questions by theoretically investigating the linear separation capacity of self-supervised representation learning when pseudo labels are generated using data augmentation.

Motivated by recent empirical successes, numerous existing studies have embarked on exploring the theoretical foundations of self-supervised pre-training's advantages in downstream tasks (Tsai et al., 2020; Tian et al., 2020b; Wen and Li, 2022; Saunshi et al., 2022; Cabannes et al., 2023). Specifically, a prevalent approach to investigating self-supervised learning is grounded in the context of linear representation or linear prediction functions (Tosh et al., 2021; Lee et al., 2021; Wen and Li, 2021; Wang, 2023; Kumar et al., 2022). While the linear framework facilitates precise characterization and insightful insights in this specific scenario, it may fall short of comprehensively elucidating the triumph of self-supervised learning, given the inherently nonlinear nature of data distributions. In addition to linear structures, existing endeavors have embraced more adaptable models, including conditional distributions of discrete latent variables (Arora et al., 2019), distributions on graphs (Wei et al., 2020; Balestriero and LeCun, 2022), nonlinear representations via kernels (Johnson et al., 2022), and manifolds (HaoChen et al., 2021; Wang, 2025). Despite the remarkable progress, the mechanisms behind how and why pseudo labels facilitate the transformation of nonlinear structures into linearly separable representations remain enigmatic. To bridge this gap, our study delves into self-supervised learning within a multi-manifold

model, demonstrating the pivotal role of pseudo labels in learning linearly separable representations, particularly when manifolds are close to each other.

1.2 Linearly Separable Representation and Unsupervised Learning

In this and the next subsections, we demonstrate the key idea and results using a two-manifold case, with a more generalized case and formal results provided in later sections. Specifically, we consider the scenario where the observed data is drawn from a union of two smooth and compact d-dimensional manifolds:

$$\mathcal{M} = \mathcal{M}_+ \cup \mathcal{M}_-,$$

where $\mathcal{M}_+$ and $\mathcal{M}_-$ correspond to distinct classes and $\mathcal{M} \subset \mathbb{R}^D$. The objective is to develop a classifier that distinguishes data originating from these two manifolds. In representation learning, our aim is to construct a mapping $\Theta : \mathcal{M} \to \mathbb{R}^S$ from the observed data, enabling us to build a more effective classifier using $\Theta(x)$ as opposed to the conventional data representation x. Recent advancements in self-supervised representation learning have indicated that a simple linear classifier can achieve perfect accuracy in downstream tasks when a linearly separable representation $\Theta(x)$ can be learned. Formally, the data representation $\Theta(x)$ is considered linearly separable if there exists a weight vector $w \in \mathbb{R}^S$ satisfying

$$\sup_{x \in \mathcal{M}_+} w^T \Theta(x) < \inf_{x \in \mathcal{M}_-} w^T \Theta(x).$$

Given a data set $(X_1, \ldots, X_n)$ randomly drawn from $\mathcal{M}$, it is natural to inquire about the conditions under which consistent learning of linearly separable representations is possible, as well as the potential benefits of using pseudo labels to enhance the learning process.

To address these questions, we first analyze the performance of the graph Laplacian-based method, a classical unsupervised learning approach widely used in spectral clustering methods (Ng et al., 2001; Belkin and Niyogi, 2003; Coifman and Lafon, 2006; Von Luxburg, 2007; Arias-Castro et al., 2017; García Trillos et al., 2021; Chen and Yang, 2021). Our investigation reveals that the resulting representation converges to a linearly separable one when the manifolds are well-separated, as indicated by the condition

$$\delta(\mathcal{M}) \gg \left(\frac{\log n}{n}\right)^{1/d}, \tag{1}$$

where $\delta(\mathcal{M})$ represents the smallest Euclidean distance between $\mathcal{M}_+$ and $\mathcal{M}_-$, given by

$$\delta(\mathcal{M}) := \inf_{x \in \mathcal{M}_+, y \in \mathcal{M}_-} \|x - y\|.$$

where $\|\cdot\|$ is the Euclidean distance in $\mathbb{R}^D$. When the condition (1) is not met, our analysis demonstrates that there exists an instance in which the graph Laplacian-based method treats $\mathcal{M}$ as a single manifold rather than a union of two separated manifolds. Furthermore, we establish an information-theoretic lower bound, revealing that any method fundamentally struggles to differentiate a single manifold from a union of multiple manifolds with

$$\delta(\mathcal{M}) \leq c \left(\frac{\log n}{n}\right)^{1/d},$$

where c is a sufficiently small constant. These findings indicate that condition (1) serves as a necessary and sufficient requirement for learning linearly separable representation in the context of observing an unlabeled data set. Importantly, the graph Laplacian-based method achieves rate-optimal linear separation capacity in the two-manifold setting. This leads us to the question of whether there are strategies to enhance the linear separation capacity stated in (1) by leveraging additional information.

1.3 Data Augmentation Improves Linear Separation Capacity

The graph Laplacian-based method is efficient at separating two manifolds but often overlooks a rich source of invariant structures present in the observed data across many applications. For instance, an image can represent the same context and be considered equivalent even after applying data augmentation techniques such as flipping, rotation, and cropping (Shorten and Khoshgoftaar, 2019; Chen et al., 2020a). To capture such invariant structures in the data, we can extend our assumption that each $\mathcal{M}_+$ and $\mathcal{M}_-$ is an isometric embedding of a product manifold:

$$\mathcal{M}_+ = T_+(\mathcal{N}_{s,+} \times \mathcal{N}_{v,+}) \qquad \text{and} \qquad \mathcal{M}_- = T_-(\mathcal{N}_{s,-} \times \mathcal{N}_{v,-}),$$

where T_+ and T_- are isometric diffeomorphisms. Here, $\mathcal{N}_{s,+}$ and $\mathcal{N}_{s,-}$ represent the d_s-dimensional manifold capturing the data augmentation invariant structure, while $\mathcal{N}_{v,+}$ and $\mathcal{N}_{v,-}$ correspond to the d_v-dimensional manifolds related to irrelevant structure due to data augmentation. A similar product manifold model is also employed in Lederman and Talmon (2018); Wang (2025). A toy example of a single product manifold is illustrated in Figure 1. This model offers a straightforward approach to modeling the data augmentation process. For instance, when $X_i \in \mathcal{M}_+$, we can express the data point as $X_i = T_+(\phi_i, \psi_i)$, where $\phi_i \in \mathcal{N}_{s,+}$ and $\psi_i \in \mathcal{N}_{v,+}$, and describe its augmented data as $X_i' = T_+(\phi_i, \psi_i')$, where ψ_i' is randomly drawn from $\mathcal{N}_{v,+}$. Similarly, when $X_i \in \mathcal{M}_-$, ψ_i' is drawn randomly from $\mathcal{N}_{v,-}$.

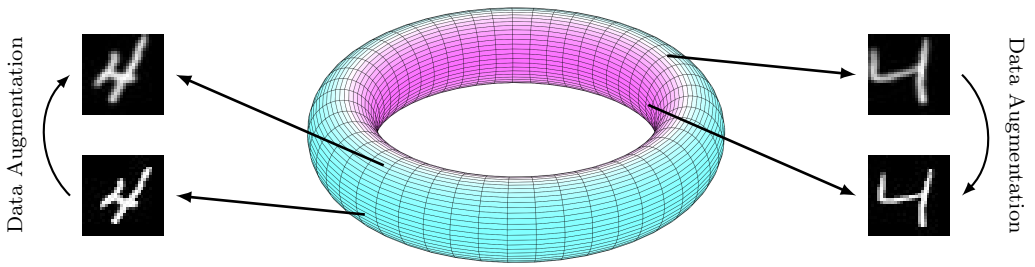

Figure 1: A toy example of a single product manifold: the circle with major radius captures data augmentation invariant structure, and the circle with minor radius captures irrelevant structure due to data augmentation. Data augmentation can help randomly draw samples from each smaller circle.

Unlike classical unsupervised techniques, self-supervised learning explores the aforementioned invariant structure of observed data when pseudo labels are generated by data

augmentation. Specifically, the representations learned by most existing data-augmentation based self-supervised learning methods aim to preserve similarity between augmented data, i.e., $\Theta(X_i) \approx \Theta(X_i')$. Can we better learn a linearly separable representation by exploiting the similarity of augmented data? To address this question, this paper studies the performance of Augmentation Invariant Manifold Learning (AIML), introduced in Wang (2025), as it simultaneously leverages the manifold's low dimensionality and the structure induced by augmented data. Capturing the invariant structure in data doesn't appear directly related to linear separation capacity, bacause it's often believed that data augmentation helps to reduce redundant information (Chen et al., 2020a; Wang, 2023). However, our investigation suggests that augmentation invariant manifold learning can surprisingly yield a linearly separable representation by requiring a smaller distance between manifolds than the graph Laplacian-based method, i.e.,

$$\delta(\mathcal{M}) \gg \left(\frac{\log n}{n}\right)^{1/d_s}. \tag{2}$$

In other words, the linear separation capacity of augmentation invariant manifold learning relies solely on the dimension of the data augmentation invariant structure. These results suggest that selecting a data augmentation method that maintains a concise invariant structure (i.e., smaller d_s) can better facilitate the learning of a linearly separable representation. When data augmentation can be applied, our analysis further demonstrates that no method can consistently discern whether the observed data is drawn from a single manifold or a union of multiple manifolds with

$$\delta(\mathcal{M}) \leq c\left(\frac{\log n}{n}\right)^{1/d_s}$$

for some sufficiently small constant c. Combined with the upper bound in (2), we observe that augmentation invariant manifold learning can achieve rate-optimal linear separation capacity when exploring the invariant structure within observed data through data augmentation.

1.4 Impact on Downstream Classifier

If we compare unsupervised and self-supervised learning, data augmentation can improve the optimal rate of linear separation capacity by

$$\left(\frac{\log n}{n}\right)^{1/d} \qquad \Rightarrow \qquad \left(\frac{\log n}{n}\right)^{1/d_s}.$$

In other words, through data augmentation, self-supervised learning can achieve separation between manifolds with a finer resolution, detecting differences between manifolds in greater detail compared to unsupervised learning. However, how does this improved linear separation capacity impact downstream analysis?

To investigate the impact on downstream analysis, we consider training a logistic regression, one of the most widely used linear classifiers, on top of learned representations to predict whether data originates from $\mathcal{M}_+$ or $\mathcal{M}_-$. The logistic regression is trained

using gradient descent on a limited labeled data set. Our investigation reveals that the misclassification rate of the resulting logistic classifier primarily hinges on how effectively the learned data representation linearly separates the data, rather than the size of the labeled data set in downstream tasks. In other words, training on a modest number of labeled samples can lead to a precise linear classifier, as long as the nonlinear structure is disentangled by the data representation learned from unlabeled samples. Consequently, with the aid of data augmentation, the data representation learned through self-supervised learning can result in a superior linear classifier for downstream analysis compared to unsupervised learning. These theoretical findings are also validated through numerical examples in Section 6. Specifically, the numerical experiments suggest that 1) the data augmentation can improve the linear separability of learned representation and 2) the performance of the downstream linear model highly relies on the linear separability of representation rather than the number of labeled data.

2. Model and Linear Separable Representation

In this section, we generalize the two-manifold setting in the Introduction section to a multi-manifold model.

2.1 A Multi-Manifold Model for Data Augmentation

The conventional representation of data is often in a high-dimensional space, but empirical observations suggest that the data distribution in various applications, such as natural image data (Pope et al., 2021), can be characterized by latent low-dimensional variables. In particular, the manifold assumption is commonly used to model the underlying low-dimensional structure in high-dimensional space. Alongside the low-dimensional structure, data can exhibit grouping into small clusters (Martinez and Saar, 2001; Basri and Jacobs, 2003; Fu et al., 2005; Vidal and Ma, 2006). Motivated by these insights, we consider data drawn from a union of multiple manifolds

$$\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \ldots \cup \mathcal{M}_K,$$

where each $\mathcal{M}_k \subset \mathbb{R}^D$ is a smooth and compact manifold of dimension d_k representing a distinct subclass of data. Each manifold $\mathcal{M}_k$ may correspond to a subset of a category or class, such as images of blue shirts, blue dresses, red shirts, and red dresses. We assume that the manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_K$ are well-separated, meaning that

$$\delta(\mathcal{M}) := \min_{1 \le k_1 < k_2 \le K} \left(\inf_{x \in \mathcal{M}_{k_1}, y \in \mathcal{M}_{k_2}} \|x - y\| \right) > 0.$$

Here, $\delta(\mathcal{M})$ represents the smallest Euclidean distance between any pair of manifolds. Since data augmentation transformations are believed not to change the subclass of data (Chen et al., 2020a), we adopt a similar approach as in Wang (2025) by assuming each $\mathcal{M}_k$ is an isometric embedding of a product manifold (Figure 1)

$$\mathcal{M}_k = T_k(\mathcal{N}_{s,k} \times \mathcal{N}_{v,k}),$$

where T_k is an isometric diffeomorphism, and $\mathcal{N}_{s,k}, \mathcal{N}_{v,k}$ are two manifolds with dimensions $d_{s,k}$ and $d_{v,k}$, such that $d_k = d_{s,k} + d_{v,k}$. Here, $\mathcal{N}_{s,k}$ and $\mathcal{N}_{v,k}$ represent $\mathcal{M}_k$'s the invariant

structure of interest and irrelevant nuisance structures resulting from data augmentation, respectively. For simplicity, we assume $d_{s,1} = \ldots = d_{s,K} = d_s$, $d_{v,1} = \ldots = d_{v,K} = d_v$, and $d_1 = \ldots = d_K = d$.

Given the above multi-manifold assumption, we need to define how an observed data point X is sampled from $\mathcal{M}$ and how the augmented data X' is sampled given a data point X. We assume that the observed data point X is drawn from a mixture distribution π defined on $\mathcal{M}$:

$$\pi(x) = \sum_{k=1}^{K} w_k \pi_k(x), \quad \text{where} \quad w_k > 0 \quad \text{and} \quad \sum_{k=1}^{K} w_k = 1.$$

Here, $\pi(x)$ and $\pi_k(x)$ are probability density functions defined on $\mathcal{M}$ (see the formal definition in Appendix), with $\pi_k(x)$ having support only on $\mathcal{M}_k$. In other words, X is drawn from $\mathcal{M}_k$ with probability w_k. Next, we discuss how data are sampled from π_k on each $\mathcal{M}_k$. The product manifold assumption suggests that all points on $\mathcal{M}_k$ can be represented as $X = T_k(\phi, \psi)$, where $\phi \in \mathcal{N}_{s,k}$ and $\psi \in \mathcal{N}_{v,k}$. Thus, it suffices to consider the probability distribution of ϕ and ψ on $\mathcal{N}_{s,k}$ and $\mathcal{N}_{v,k}$. We sample independent random variables $\phi \in \mathcal{N}_{s,k}$ and $\psi \in \mathcal{N}_{v,k}$ as follows:

$$\phi \sim \pi_k^s(\phi) \quad \text{and} \quad \psi \sim \pi_k^v(\psi).$$

Then, we set $X = T_k(\phi, \psi)$. For simplicity, we assume that π_k^v is a uniform distribution on $\mathcal{N}_{v,k}$. The product manifold assumption also provides a straightforward way to model the sampling process of augmented data X' given a data point X. Given $X = T_k(\phi, \psi) \in \mathcal{M}_k$, we assume that the augmented data $X' = T_k(\phi, \psi')$, where ψ' is another independent random variable drawn from π_k^v on $\mathcal{N}_{v,k}$. In other words, given $X = T_k(\phi, \psi) \in \mathcal{M}_k$, we randomly sample the augmented data point from the fiber $\mathcal{M}(\phi) = \{x \in \mathcal{M}_k : x = T_k(\phi, \psi), \psi \in \mathcal{N}_{v,k}\}$.

2.2 Linearly Separable Representation

Representation learning aims to learn a mapping $\Theta : \mathcal{M} \to \mathbb{R}^S$ that enhances the performance of downstream analyses, including clustering, classification, and regression. In this paper, we focus solely on classification as the downstream task. In a classification task, our goal is to train a classifier $H : \mathcal{M} \to \{-1, 1\}$ that predicts the true label $H^* : \mathcal{M} \to \{-1, 1\}$, where -1 and 1 denote the two possible binary label values. Among the simplest and most commonly used classifiers is the linear classifier:

$$H_w(x) = \begin{cases} 1 & w^T x > \zeta \\ -1 & w^T x \leq \zeta \end{cases},$$

where $w \in \mathbb{R}^D$ represents the weight vector and $\zeta \in \mathbb{R}$ is a scalar threshold. Various algorithms have been proposed in the literature to train a linear classifier, such as logistic regression, linear discriminant analysis, and support vector machines (Hastie et al., 2009). A linear classifier can achieve 100% accuracy when the sets $\mathcal{M}_+(H^*) = \{x \in \mathcal{M} : H^*(x) = 1\}$ and $\mathcal{M}_-(H^*) = \{x \in \mathcal{M} : H^*(x) = -1\}$ are linearly separable in terms of their x values.

However, in practice, $\mathcal{M}_+(H^*)$ and $\mathcal{M}_-(H^*)$ are often not linearly separable due to their complex shapes in high-dimensional space.

Through representation learning, our aim is to enhance the classifier trained by $\Theta(x)$, making it more accurate than the classifier trained using x. Additionally, we aspire for the data representation $\Theta(x)$ to be versatile, applicable across multiple classification labels. Specifically, we address a collection of classification labels within the multi-manifold model

$$\mathcal{H}^* = \{H^* : H^*(x) \in \{-1, 1\}, H^*(x) = H^*(y), \quad \forall\ x, y \in \mathcal{M}_k, \quad 1 \le k \le K\}.$$

In this definition, we enforce that the labels of data points within each subclass, i.e., manifold $\mathcal{M}_k$, are identical. As a result, $\mathcal{H}^*$ encompasses 2^K distinct possible classification labels. Our objective is to develop a data representation that enhances the classifier's performance across all classification tasks within $\mathcal{H}^*$. For the purpose of this study, we concentrate on a linear classifier and endeavor to construct a linearly separable representation. In particular, a data representation $\Theta(x)$ is deemed linearly separable if, for any classification task $H^* \in \mathcal{H}^*$, a weight vector $w_{H^*} \in \mathbb{R}^S$ can be found such that

$$\sup_{x \in \mathcal{M}_+(H^*)} w_{H^*}^T \Theta(x) < \inf_{x \in \mathcal{M}_-(H^*)} w_{H^*}^T \Theta(x).$$

Having introduced this definition, the natural inquiry arises: does a linearly separable representation exist, and if so, how can we learn it from our observed data?

3. Classical graph Laplacian-based Method

To learn data representation on the multi-manifold, it is sufficient to find a way to separate each individual manifold. A commonly used unsupervised method to learn individual manifold structure is based on the spectral analysis of the graph Laplacian on a neighborhood graph (Belkin and Niyogi, 2003; Coifman and Lafon, 2006). Let $X_1, \ldots, X_n$ be independent identically distributed samples drawn from the distribution π, which is introduced in Section 2. Given the observed samples, we can construct a neighborhood graph by connecting an edge between two sample points X_i and X_j if and only if $\|X_i - X_j\| \le r$ where $\|\cdot\|$ is the Euclidean distance in $\mathbb{R}^D$, and assigning weights $W_{i,j} = \mathbf{I}(\|X_i - X_j\| \le r)$, where $\mathbf{I}(\cdot)$ is the indicator function. With the neighborhood graph introduced, we define the graph Laplacian matrix L as

$$L_{i,j} = \begin{cases} \sum_{i' \neq i} W_{i,i'} & \text{if } i = j \\ -W_{i,j} & \text{if } i \neq j \end{cases}.$$

Then, the first S eigenvectors (corresponding to the S smallest eigenvalues) of L (denoted $U_1, \ldots, U_S$) can form the data representations for $X_1, \ldots, X_n$. Since these data representations can be used to separate manifolds, they have already been employed in various versions of spectral clustering algorithms (Ng et al., 2001; Von Luxburg, 2007; Arias-Castro et al., 2017; García Trillos et al., 2021; Chen and Yang, 2021).

To understand why and when the graph Laplacian can help recover the multi-manifold structure, we need to study the asymptotic behavior of L's eigenvectors and introduce the continuum level of the Laplacian operator. Specifically, we define the Laplacian operator on $\mathcal{M}_k$ as

$$\Delta_{\mathcal{M}_k} \theta_k = -\frac{1}{\pi_k} \mathrm{div}_{\mathcal{M}_k}(\pi_k^2 \nabla_{\mathcal{M}_k} \theta_k),$$

where $\theta_k : \mathcal{M}_k \to \mathbb{R}$ is a function defined on $\mathcal{M}_k$. After introducing Laplacian operator on each $\mathcal{M}_k$, we now define the tensorized Laplacian operator $\Delta_{\mathcal{M}}$

$$\Delta_{\mathcal{M}}\theta = (w_1\Delta_{\mathcal{M}_1}\theta_1, \ldots, w_K\Delta_{\mathcal{M}_K}\theta_K),$$

where $\theta : \mathcal{M} \to \mathbb{R}$ is a function defined on $\mathcal{M}$ and can be written as $\theta = (\theta_1, \ldots, \theta_K)$ where each θ_k is defined on $\mathcal{M}_k$. Since the operator $\Delta_{\mathcal{M}}$ is defined in a manifold-wise fashion, the eigenfunctions of $\Delta_{\mathcal{M}}$ only has support on one manifold, i.e., the eigenfunctions $\theta_{k,l_k}(x) = \theta^k_{l_k}(x)/\sqrt{w_k}$ if $x \in \mathcal{M}_k$ and $\theta_{k,l_k}(x) = 0$ if $x \notin \mathcal{M}_k$, where $\theta^k_{l_k}(x)$ is the l_kth eigenfunction of $\Delta_{\mathcal{M}_k}$. In particular, the first K eigenfunctions of $\Delta_{\mathcal{M}}$ has the form $\theta(x) = c_k\mathbf{I}(x \in \mathcal{M}_k)$, where c_k is some normalization constant and $1 \le k \le K$. In other words, the structure information of different manifolds is mapped to different coordinates with the help of these eigenfunctions. Because of this special property, the representation based on these eigenfunctions can localize each manifold and thus lead to linearly separable representation. It is also interesting to note that these eigenfunctions can capture the geometric information within each manifold.

The primary reason behind the capability of the eigenvectors of L to recover the multi-manifold structure stems from the fact that the graph Laplacian provides a reliable approximation of the Laplacian operator $\Delta_{\mathcal{M}}$, and the eigenvectors of L converge to the eigenfunctions of $\Delta_{\mathcal{M}}$. To establish this convergence, we need to introduce the following assumption:

Assumption 1 *We assume that the following conditions hold:*

1. *There exists a constant $C_\pi > 1$ such that*

$$\frac{1}{C_\pi} \le \pi^s_k(\phi) \le C_\pi, \quad 1 \le k \le K;$$

2. *The parameter r is chosen such that*

$$2r < \min\{1, i_o, \Gamma^{-1/2}, R/2\},$$

where R and Γ denote the upper bounds of the reach and absolute values of sectional curvatures, respectively, and i_0 is a lower bound on the injectivity radius.

These assumptions are commonly used in the analysis of the convergence of the spectrum of the graph Laplacian (Calder and García Trillos, 2022; García Trillos et al., 2021). Given Assumption 1, we can establish the convergence of the eigenvectors of L.

Theorem 1 *If Assumption 1 holds and we assume $r \to 0$,*

$$r \gg \left(\frac{\log n}{n}\right)^{1/d} \qquad \text{and} \qquad \delta(\mathcal{M}) > r, \tag{3}$$

then with probability at least $1 - 4Kn^{-\alpha}$, if U_s is normalized eigenvector of L with eigenvalue $\lambda_s(L)$, there is a normalized eigenfunction θ_s of $\Delta_{\mathcal{M}}$ with eigenvalue $\lambda_s(\mathcal{M})$ such that

$$\|U_s - \vec{\theta}_s\|_{L^2(\pi_n)} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((U_s(X_i) - \theta_s(X_i)\right)^2} \to 0, \qquad n \to \infty,$$

where $\vec{\theta}_s = (\theta_s(X_1), \ldots, \theta_s(X_n)) \in \mathbb{R}^n$.

The proof of Theorem 1 adopts the same variational method as in Burago et al. (2015); García Trillos et al. (2020); Calder and García Trillos (2022); García Trillos et al. (2021). The detailed convergence rate of eigenfunctions can be found in Appendix. Theorem 1 shows that under certain conditions, including the main condition (3), the eigenvectors of L can approximate the eigenfunctions of $\Delta_{\mathcal{M}}$ very well and thus lead to linearly separable representation. Can these conditions in (3) be relaxed? The lower bound condition for r in (3) is sharp since the connectivity threshold of the random geometric graph is at order $(\log n/n)^{1/d}$ (Penrose, 2003). The other condition $\delta(\mathcal{M}) > r$ is also a necessary condition for separating different manifolds. To show this, we present the following example and theorem to argue that the eigenvectors of L cannot separate multi-manifold when $\delta(\mathcal{M}) \ll r$.

Consider a simple example of two manifolds $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$. $\mathcal{M}_1$ and $\mathcal{M}_2$ are defined in the following way:

$$\mathcal{M}_1 = \left\{(y, 0) : y \in \tilde{\mathcal{M}}\right\} \qquad \text{and} \qquad \mathcal{M}_2 = \left\{(y, z^o) : y \in \tilde{\mathcal{M}}\right\},$$

where $\tilde{\mathcal{M}}$ is a d-dimensional smooth and compact manifold embedded in $\mathbb{R}^{D-1}$, and $z^o \in \mathbb{R}$ is a positive constant. This construction clearly shows that $\delta(\mathcal{M}) = z^o$. The following theorem shows that eigenvectors of L converge to eigenfunctions of $\Delta_{\tilde{\mathcal{M}}}$ instead of $\Delta_{\mathcal{M}}$.

Theorem 2 *Suppose $w_1 = w_2 = 1/2$, and $\pi_1(x)$ and $\pi_2(x)$ are the uniform distribution on $\mathcal{M}_1$ and $\mathcal{M}_2$ in above example. So we can write our observed data as $X_i = (Y_i, Z_i)$ for $1 \le i \le n$, where $Y_i \in \mathbb{R}^{D-1}$ is drawn from a uniform distribution on $\tilde{\mathcal{M}}$ and $\mathbb{P}(Z_i = z^o) = \mathbb{P}(Z_i = 0) = 1/2$. If Assumption 1 holds and we assume $r \to 0$,*

$$r \gg \left(\frac{\log n}{n}\right)^{1/d} \qquad \text{and} \qquad \delta(\mathcal{M}) \ll r,$$

then with probability at least $1 - Cn^{-2}$ for some constant C, if U_s is normalized eigenvector of L with sth eigenvalue, there is a normalized eigenfunction θ_s of $\Delta_{\tilde{\mathcal{M}}}$ with sth eigenvalue such that

$$\|U_s - \vec{\theta}_s\|_{L^2(\pi_n)} \to 0, \qquad n \to \infty,$$

where $\vec{\theta}_s = (\theta_s(Y_1), \ldots, \theta_s(Y_n)) \in \mathbb{R}^n$.

As $\vec{\theta}_s$ depends solely on $Y_1, \ldots, Y_n$ and not on $Z_1, \ldots, Z_n$, Theorem 2 suggests that the eigenvectors of L cannot effectively distinguish between $\mathcal{M}_1$ and $\mathcal{M}_2$, thereby failing to provide a linearly separable representation when $\delta(\mathcal{M}) \ll r$. The combined implications of Theorem 1 and 2 imply that the classical Laplacian-based method necessitates well-separated manifolds in order to learn a linearly separable representation, meaning that the minimum distance between manifolds must be sufficiently large:

$$\delta(\mathcal{M}) \gg \left(\frac{\log n}{n}\right)^{1/d}.$$

While this minimum distance requirement optimally applies to the classical graph Laplacian-based method, it naturally raises the question of whether alternative methods can effectively separate closely situated manifolds.

We investigate the information-theoretic lower bound for separating manifolds to answer this question. Following a similar strategy in Mossel et al. (2015), we address a simpler question: what is the smallest distance between manifolds that enables us to distinguish whether our observed data is drawn from one or two manifolds? Specifically, we consider the following hypothesis testing problem:

$$\mathbb{H}_0 : \mathcal{M} = \mathcal{M}_0 \qquad \text{v.s.} \qquad \mathbb{H}_1 : \mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2. \tag{4}$$

Here $\mathcal{M}_0$, $\mathcal{M}_1$, and $\mathcal{M}_2$ are smooth and compact d-dimensional manifolds such that the distance between $\mathcal{M}_1$ and $\mathcal{M}_2$ is $\delta(\mathcal{M}) > 0$. In other words, under the null hypothesis, the data is drawn from a single manifold, while under the alternative hypothesis, it is drawn from a union of two manifolds. Testing hypothesis in (4) is a simpler problem than constructing linear separable representation because we can distinguish $\mathbb{H}_0$ from $\mathbb{H}_1$ if we can linearly separate $\mathcal{M}_1$ and $\mathcal{M}_2$. To detect the above hypothesis, we can consider a test $T : (X_1, \ldots, X_n) \to \{0, 1\}$ such that we reject the null hypothesis if $T = 1$. A test T is called α-level test if $\mathbb{P}_0(T = 1) \leq \alpha$ where $\mathbb{P}_0$ is the probability measure under $\mathbb{H}_0$. The following theorem characterizes the lower bound for separating manifolds.

Theorem 3 *For any α-level test $T(X_1, \ldots, X_n)$, there exists an instance where*

$$\delta(\mathcal{M}) \geq \left(\frac{b \log n}{n}\right)^{1/d},$$

for some small enough constant $b > 0$, such that the type II error of this test converges to $1 - \alpha$.

This theorem demonstrates that any method is fundamentally limited in its ability to separate manifolds when the minimum distance requirement is not met. Combining this result with Theorem 1, we observe that the optimal rate of linear separation capacity is given by

$$\delta(\mathcal{M}) \asymp \left(\frac{\log n}{n}\right)^{1/d}.$$

While the classical graph Laplacian-based method can achieve this optimal rate, a natural question arises: Can we uncover additional data structure and relax this minimum distance requirement by extracting more information from the data?

4. Augmentation Invariant Manifold Learning

Besides the observed samples, we can also generate additional augmented data from the observed data in various applications. For instance, cropping, rotation, colorization, and scaling can produce new images from the original images (Shorten and Khoshgoftaar, 2019). The augmented data is treated as an equivalent yet distinct version of the original data, offering supplementary information beyond the observed data. In recent years, self-supervised representation learning methods, encompassing both contrastive and non-contrastive approaches, have been introduced to learn data representations by harnessing this equivalent relationship among augmented data (Chen et al., 2020b; Grill et al., 2020; Tian et al., 2020a;

Chen and He, 2021; Zbontar et al., 2021; Wang, 2025). The data augmentation invariant representations learned through these methods can enhance downstream analysis.

To leverage augmented data, existing self-supervised representation learning methods aim to learn a data representation that is invariant to augmented data, i.e.,

$$\Theta(X_i) \approx \Theta(X_i'),$$

where X_i' is a data point randomly generated by applying data augmentation techniques to X_i. Specifically, augmentation invariant manifold learning has been recently introduced to learn data representation by capturing the geometric structure of the manifold and the invariance property of augmented data (Wang, 2025). To elaborate further, augmentation invariant manifold learning can be formulated as an stochastic optimization algorithm and the loss function on a small batch of samples $\mathcal{S} \subset \{1, \ldots, n\}$ is defined as

$$\hat{\ell}(\beta) = \sum_{i,j \in \mathcal{S}} W_{i,j} \|\Theta_\beta(X_i') - \Theta_\beta(X_j'')\|^2 + \lambda_1 \sum_{i \in \mathcal{S}} \|\Theta_\beta(X_i') - \Theta_\beta(X_i'')\|^2 + \lambda_2 \mathcal{R}(\Theta_\beta), \quad (5)$$

where X_i' and X_i'' are independent augmented copies of X_i, $W_{i,j}$ represents the weights between X_i' and X_j'', and $\mathcal{R}(\Theta_\beta)$ is a regularization term enforcing an orthonormal representation. In the above loss function, the first term corresponds to a computationally efficient version of the graph Laplacian, while the second term aims to capture the similarity between augmented data. The detailed algorithm of augmentation invariant manifold learning can be found in Algorithm 1 in Appendix.

Wang (2025) demonstrated that preserving the similarity of augmented data is equivalent to incorporating weights between augmented data of two samples. From this perspective, an equivalent form of augmentation invariant manifold learning in Algorithm 1 follows a similar procedure to the classical Laplacian-based method, but assesses the similarity between two samples using the average of weights between their augmented data:

$$\bar{W}_{i,j} = \mathbb{E}\left(\mathbf{I}(\|X_i' - X_j'\| \leq r)\right),$$

where X_i' and X_j' are the independent augmented data of X_i and X_j, and the expectation is taken over the data augmentation process. These new weights $\bar{W}_{i,j}$ lead to the construction of the graph Laplacian matrix $\bar{L}$ in the same manner as the classical Laplacian-based method. Consequently, the first S eigenvectors of $\bar{L}$ serve as the new data representations. In scenarios involving a single manifold, Wang (2025) shows that augmentation invariant manifold learning can more effectively reduce dimensions nonlinearly compared to the classical Laplacian-based method. Given the resemblance between augmentation invariant manifold learning and the classical Laplacian-based method, a natural question emerges: Can augmentation invariant manifold learning also facilitate the learning of linearly separable representations? If so, how can data augmentation contribute to representation learning?

To study the properties of augmentation invariant manifold learning, it is necessary to introduce the Laplacian operator on the augmentation invariant manifolds $\mathcal{N}_s := \mathcal{N}_{s,1} \cup \ldots \cup \mathcal{N}_{s,K}$. Given the multi-manifold $\mathcal{N}_s$, we define a tensorized Laplacian operator $\Delta_{\mathcal{N}_s}$ as follows:

$$\Delta_{\mathcal{N}_s} \bar{\theta} = \left(\frac{w_1}{\mathrm{Vol}\mathcal{N}_{v,1}} \Delta_{\mathcal{N}_{s,1}} \bar{\theta}_1, \ldots, \frac{w_K}{\mathrm{Vol}\mathcal{N}_{v,K}} \Delta_{\mathcal{N}_{s,K}} \bar{\theta}_K \right),$$

where $\bar{\theta} : \mathcal{N}_s \to \mathbb{R}$ is a function defined on $\mathcal{N}_s$, $\bar{\theta}_k : \mathcal{N}_{s,k} \to \mathbb{R}$ is a function defined on $\mathcal{N}_{s,k}$, $\mathrm{Vol}\mathcal{N}_{v,k}$ represents the volume of $\mathcal{N}_{v,k}$, and $\Delta_{\mathcal{N}_{s,k}}$ denotes the Laplacian operator on $\mathcal{N}_{s,k}$ given by

$$\Delta_{\mathcal{N}_{s,k}}\bar{\theta}_k = -\frac{1}{\pi_k^s}\mathrm{div}_{\mathcal{N}_{s,k}}(\pi_k^{s\,2}\nabla_{\mathcal{N}_{s,k}}\bar{\theta}_k).$$

Similar to $\Delta_{\mathcal{M}}$, the eigenfunctions of $\Delta_{\mathcal{N}_s}$ also have support on each individual manifold. Thus, a representation based on the eigenfunctions of $\Delta_{\mathcal{N}_s}$ can effectively separate manifolds and lead to a linearly separable representation. In the following, we demonstrate that the eigenvectors of $\bar{L}$ converge to eigenfunctions of $\Delta_{\mathcal{N}_s}$, enabling the detection of the multi-manifold structure.

Theorem 4 *If Assumption 1 holds and we assume $r \to 0$,*

$$r \gg \left(\frac{\log n}{n}\right)^{1/d_s} \qquad \text{and} \qquad \delta(\mathcal{M}) > r, \tag{6}$$

then with probability at least $1 - 4Kn^{-\alpha}$, if $\bar{U}_s$ is normalized eigenvector of $\bar{L}$ with sth eigenvalue, there is a normalized eigenfunction $\bar{\theta}_s$ of $\Delta_{\mathcal{N}_s}$ with sth eigenvalue such that

$$\|\bar{U}_s - \vec{\bar{\theta}}_s\|_{L^2(\pi_n)} \to 0, \qquad n \to \infty,$$

where $\vec{\bar{\theta}}_s = (\bar{\theta}_s(\phi_1), \ldots, \bar{\theta}_s(\phi_n)) \in \mathbb{R}^n$.

Theorem 4 suggests that the graph Laplacian of $\bar{L}$ is a good approximation of the Laplacian operator $\Delta_{\mathcal{N}_s}$ rather than $\Delta_{\mathcal{M}}$. Thus, the eigenvectors of $\bar{L}$ can also detect each individual manifold and be used to construct a linearly separable representation when

$$\delta(\mathcal{M}) \gg \left(\frac{\log n}{n}\right)^{1/d_s}.$$

In other words, the weight $\bar{W}_{i,j}$ defined by augmented data can separate manifolds with a smaller $\delta(\mathcal{M})$ than the classical Laplacian-based method. The intuition behind this improvement is that $\bar{W}_{i,j}$ measures the similarity between the fibers $\mathcal{M}(\phi_i)$ and $\mathcal{M}(\phi_j)$ rather than X_i and X_j, leading a kernel between latent variables ϕ_i and ϕ_j even though we cannot observe ϕ_i and ϕ_j directly. More specifically, $\bar{W}_{i,j}$ can be approximately written as the following kernel

$$\frac{1}{r^{d_v}}\bar{W}_{i,j} \approx \frac{V_{d_v}}{\mathrm{Vol}\mathcal{N}_{v,k}}\left(1 - \frac{d^2_{\mathcal{N}_s}(\phi_i, \phi_j)}{r^2}\right)_+^{d_v/2},$$

where $X_i, X_j \in \mathcal{M}_k$, V_{d_v} is the volume of a d_v-dimensional unit ball, $(x)_+ = \max(x, 0)$ and $d_{\mathcal{N}_s}(\phi_i, \phi_j)$ is the geodesic distance on $\mathcal{N}_s$. Therefore, capturing the invariant data structure can help reduce dimension nonlinearly and separate different manifolds.

We show that the multiple manifolds can be better separated via capturing the invariant structure of data, but is the above minimum distance requirement sharp? To address this, we establish an information-theoretic lower bound for the task of separating manifolds by analyzing the hypothesis testing problem introduced in (4). Due to the presence of data augmentation, we are not limited to observing individual samples $X_1, \ldots, X_n$; instead, we

have access to a collection of fiber sets $\mathcal{M}(\phi_1), \ldots, \mathcal{M}(\phi_n)$. When data augmentation is available, we can consider a test defined by these fibers, that is, $T : (\mathcal{M}(\phi_1), \ldots, \mathcal{M}(\phi_n)) \to \{0, 1\}$ such that we reject the null hypothesis if $T = 1$. The ensuing theorem establishes a lower bound for the task of separating manifolds by leveraging the invariant structure of augmented data.

Theorem 5 *For any α-level test $T(\mathcal{M}(\phi_1), \ldots, \mathcal{M}(\phi_n))$, there exists an instance where*

$$\delta(\mathcal{M}) \geq \left(\frac{b \log n}{n}\right)^{1/d_s},$$

for some small enough constant $b > 0$, such that the type II error of this test converges to $1 - \alpha$.

The results presented in Theorems 4 and 5 demonstrate that augmentation invariant manifold learning can attain the optimal rate of linear separation capacity. By leveraging the additional information gained through data augmentation, augmentation invariant manifold learning enhances the rate of linear separation capacity to the following level:

$$\delta(\mathcal{M}) \asymp \left(\frac{\log n}{n}\right)^{1/d_s}.$$

A comparison between Theorem 3 and 5 underscores the indispensability of augmented data in acquiring linearly separable data representations, particularly when the manifolds exhibit proximity to one another.

5. Impact on Downstream Linear Classifier

The previous sections characterize the required conditions for learning linearly separable representations, but it is still unclear how these representations can lead to an efficient linear classifier. To demystify the effectiveness of linearly separable representations, we consider a binary classification downstream task and study the performance of logistic regression. To be specific, suppose we observe a collection of labeled samples $(\tilde{X}_1, Y_1), \ldots, (\tilde{X}_m, Y_m)$ where the label $Y_i = H^*(\tilde{X}_i)$ for some $H^* \in \mathcal{H}^*$. In the logistic regression, our goal is to minimize the following empirical logistic risk

$$\min_{\beta \in \mathbb{R}^S} \quad \mathcal{L}_{\log}(\beta) = \frac{1}{m} \sum_{i=1}^{m} \ln\left(1 + \exp\left[-Y_i \beta^T \hat{\Theta}(\tilde{X}_i)\right]\right),$$

where $\hat{\Theta} : \mathcal{M} \to \mathbb{R}^S$ is some data representations resulted from previous sections. A commonly used way to minimize $\mathcal{L}_{\log}(\beta)$ is the gradient descent method, i.e.,

$$\beta_{t+1} = \beta_t - \eta_t \nabla \mathcal{L}_{\log}(\beta_t),$$

where η_t is the step size in the tth iteration. After stopping the iterations, the resulting classifier is

$$\hat{H}_{\hat{\Theta}, \hat{\beta}}(x) = \begin{cases} 1, & \hat{\beta}^T \hat{\Theta}(x) > 0 \\ -1, & \hat{\beta}^T \hat{\Theta}(x) \leq 0 \end{cases},$$

where $\hat{\beta}$ is the resulting weighted vector in the above gradient descent iterates. To study the performance of $\hat{H}_{\hat{\Theta},\hat{\beta}}$, we consider the misclassification rate as our measure

$$\xi(\hat{\Theta}) = \mathbb{P}\left(\hat{H}_{\hat{\Theta},\hat{\beta}}(X) \neq H^*(X)\right).$$

To characterize the theoretical properties of $\hat{H}_{\hat{\Theta},\hat{\beta}}$, we consider the following assumptions:

Assumption 2 *It holds that*

1. *$\tilde{X}_1, \ldots, \tilde{X}_m$ is independent of the data in unsupervised/self-supervised learning ($X_1, \ldots, X_n$), so $\tilde{X}_1, \ldots, \tilde{X}_m$ is independent of $\hat{\Theta}$.*

2. *We choose the data representation as the first K eigenfunctions resulting from the classical Laplacian-based method or augmentation invariant manifold learning, i.e., $\hat{\Theta}(x) = (\hat{\theta}_1(x), \ldots, \hat{\theta}_K(x))$. We also assume $\hat{\theta}_s(x)$ is close to $\theta_s(x)$ in the following sense*
$$\sum_{k=1}^{K} w_k \int_{\mathcal{M}_k} (\hat{\theta}_s(x) - \theta_s(x))^2 \pi_k(x) dx \leq \chi_n, \qquad 1 \leq s \leq K,$$
where $\theta_s(x)$ is the sth eigenfunction of $\Delta_{\mathcal{M}}$ or $\Delta_{\mathcal{N}_s}$.

3. *For each $1 \leq s \leq K$, we have at least one sample $\tilde{X}_i \in \mathcal{M}_s$.*

4. *The step size of the gradient descent is chosen as $\eta_t = 1$ and the number of steps is large enough.*

The above assumptions are mild in practice. In particular, the second assumption is just a population version of the results in Theorem 1, 2 and 4, and the last assumption is used for the convergence of the gradient descent method when the data is linearly separable Soudry et al. (2018); Ji and Telgarsky (2018). With these assumptions, the following theorem shows the convergence of the misclassification rate.

Theorem 6 *If Assumption 2 holds, and m and K are upper bounded by some constant, then, with probability $1 - C_{m,K}\chi_n$, we have*

$$\xi(\hat{\Theta}) \leq C_K \chi_n.$$

Here, $C_{m,K}$ is some constant relying on m and K and C_K is some constant relying on K.

Theorem 6 demonstrates that the performance of the downstream linear classifier hinges on the quality of our representation in approximating a linearly separable one, rather than the sample size in the downstream analysis. This result implies that a linear classifier, learned from a limited number of labeled samples, can achieve a low misclassification rate when self-supervised learning effectively captures an accurate linearly separable representation through a large quantity of unlabeled samples. By combining Theorem 6 with the findings from preceding sections, we can further compare the impacts of supervised and self-supervised learning on the downstream linear classifier. Let $\hat{\Theta}_L$ and $\hat{\Theta}_{\bar{L}}$ be the data representations acquired through the classical Laplacian-based method and augmentation

invariant manifold learning, respectively. When $(\log n/n)^{1/d_s} \ll \delta(\mathcal{M}) \ll (\log n/n)^{1/d}$, with a probability tending to 1,

$$\lim_{n\to\infty} \xi(\hat{\Theta}_{\bar{L}}) = 0,$$

and there exists an instance (corresponding to the scenario in Theorem 2) such that

$$\lim_{n\to\infty} \xi(\hat{\Theta}_L) \geq 1/2.$$

6. Numerical Illustration

In this section, we present a series of numerical experiments to validate the theoretical results from the previous sections and to compare the performance of unsupervised and self-supervised learning methods. Specifically, we utilize the MNIST data set (LeCun et al., 1998), which consists of 60,000 training and 10,000 testing images of 28×28 gray-scale handwritten digits. For our self-supervised learning approach, we adopt the transformation introduced by Wang (2025) to generate augmented data. This involves random resizing, cropping, and rotation of the images.

In all the numerical experiments, we compare the performance of Augmentation Invariant Manifold Learning (AIML) as defined in (5), with that of a continuous version of the classical graph Laplacian-based method (CML). To comprehensively study the impact of data augmentation, we formulate the unsupervised graph Laplacian-based method as a similar optimization problem to (5), but with the removal of all components related to data augmentation:

$$\min_{\beta\in\mathcal{B}} \quad \sum_{i,j=1}^{n} W_{i,j}\|\Theta_\beta(X_i) - \Theta_\beta(X_j)\|^2 + \lambda_2\mathcal{R}(\Theta_\beta), \tag{7}$$

where $W_{i,j}$ represents the weights between X_i and X_j, and Θ_β is the encoder. This optimization problem can be regarded as a computationally efficient and continuous version of the classical graph Laplacian-based method. For a fair comparison, we employ the same convolutional neural network encoder with two convolution+ReLU layers, two pooling layers, and a fully connected layer. Additionally, we use identical tuning parameters and optimization algorithms for both AIML and CML.

We present t-SNE plots (Hinton and Roweis, 2002) depicting the learned representations in Figure 2 when applying these two methods to the training images. It is important to note that t-SNE plots can distort the distances between data points, but they offer a visual representation of the cluster structure in the learned representations. As shown in Figure 2, each digit corresponds to multiple clusters, and leveraging data augmentation enhances the separation of distinct digits.

To quantitatively assess linear separability, we train a linear classifier on top of the learned representations and measure the resulting classifier's accuracy. We recode the digit labels as 1 when the digit is smaller than 5 and as -1 when the digit is equal to or larger than 5. Thus, the classification task aims to predict whether a digit is smaller than 5. We conduct two sets of numerical experiments to evaluate the influence of sample size on representation learning (via unsupervised or self-supervised methods) and classifier training in the downstream task. In the first set of experiments, we use all training samples (without labels) to learn data representations, and the linear classifier is trained using 20%, 40%, 60%,

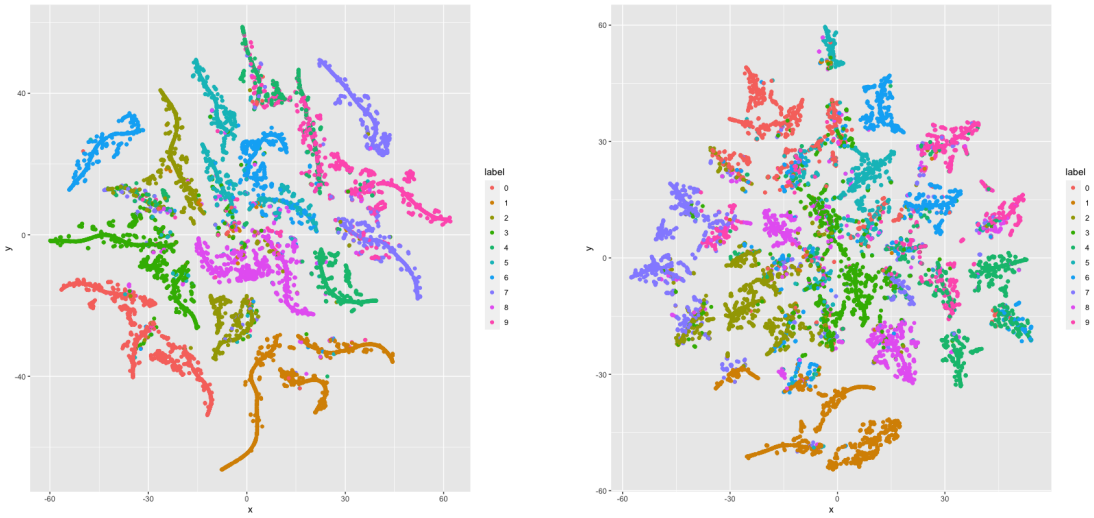

Figure 2: t-SNE plots of representation learned by augmentation invariant manifold learning (left) and graph Laplacian-based method (right).

80%, and 100% of labeled samples. The left figure in Figure 3 presents the misclassification rates for AIML and CML representations. It's evident that the performance of AIML and CML remains stable as the sample size increases from 40% to 100%, confirming the findings in Theorem 6. In the second set of experiments, we learn the representation from 40%, 60%, 80%, and 100% of unlabeled training samples, and the downstream classifier is trained with 40% of labeled samples. The resulting linear classifier's misclassification rates are reported in the right figure of Figure 3. Clearly, a larger sample size in representation learning leads to a more accurate classifier. Combining the outcomes of both experiment sets, we observe that data augmentation aids in learning better linearly separable representations, and the classifier's accuracy primarily depends on the separability of learned representations, rather than the downstream analysis's sample size. These experimental results align with the theoretical findings presented in previous sections.

7. Concluding Remarks

In this paper, we delve into the significance of data augmentation in learning linearly separable representations. Our investigation underscores the pivotal role played by the invariant structures introduced through data augmentation, facilitating the better transformation of intricate nonlinear data structures into linearly separable representations via unlabeled data sets. Coupled with the insights gleaned from the findings in Wang (2025), we conclude that data augmentation yields two principal advantages within self-supervised representation learning: enhanced separation of closely situated manifolds and nonlinear dimension reduction within each distinct manifold. Furthermore, exploring the implications for downstream analysis reveals that the crux of effective classification lies in the successful segregation of data originating from distinct manifolds. Remarkably, even with a limited number of la-

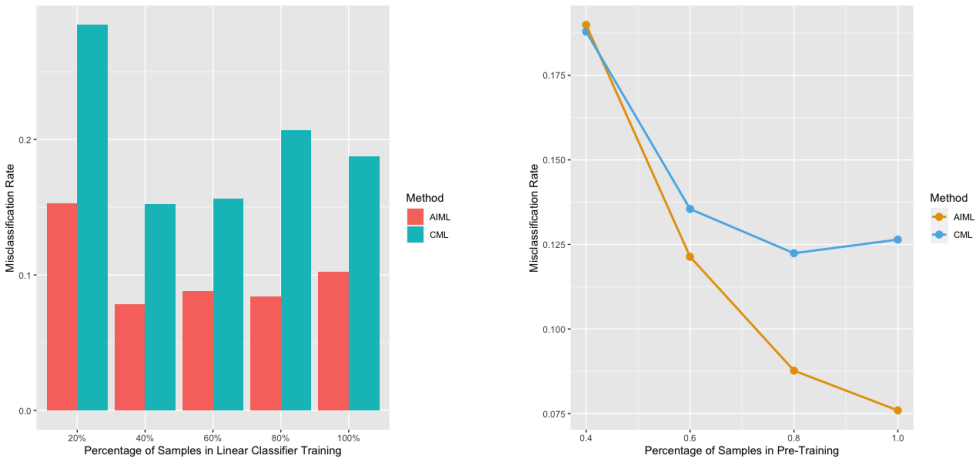

Figure 3: Misclassification rate of AIML and CML: the left figure shows the result when the sample size in representation learning is fixed and in downstream task varies; the right figure shows the result when the sample size in representation learning varies and in downstream task is fixed.

beled samples, the availability of an efficient linearly separable representation empowers the construction of accurate classifiers. This observation elucidates the feasibility of developing efficient algorithms in the context of few-shot learning (Wang et al., 2020), particularly in scenarios where an extensive unlabeled data set coexists with a smaller labeled counterpart.

To simplify the analysis, we assume the sampling distribution in data augmentation π_k^v is uniform on $\mathcal{N}_{v,k}$. Although it seems restrictive, this assumption holds for many applications. For example, the rotation angle is usually uniformly sampled when the image is rotated randomly. If the π_k^v is not a uniform distribution anymore, the form of $\bar{W}_{i,j}$ becomes more complicated than the current one, but the conclusion of Theorem 4 can hold similarly because we can still consider $\bar{W}_{i,j}$ as a kernel between ϕ_i and ϕ_j. Another essential assumption in our investigation is that the data is drawn from a union of multiple manifolds where each manifold represents a subset of a class. This assumption is reasonable in many data sets with well-defined categories, including MNIST and CIFAR-10 data sets. However, this assumption might be limited when it comes to complex data sets such as the ImageNet data set. Therefore, it is interesting to study whether we can study the linear separation capacity under a broader setting beyond the multiple manifolds assumption.

Acknowledgments

This project is supported by grants from the National Science Foundation (DMS-2113458).

Appendix A. Algorithm for Augmentation Invariant Manifold Learning

This section presents the stochastic optimization algorithm for augmentation invariant manifold learning introduced in Wang (2025), summarized in Algorithm 1.

Algorithm 1 Augmentation Invariant Manifold Learning

Input: A set of data $\{X_1, \ldots, X_n\}$, batch size n', encoder Θ_β, stochastic data augmentation transformation $\mathcal{T}$, tuning parameters $(r, \lambda_1, \lambda_2)$.

1: **for** sampled minibatch $\{X_i : i \in \mathcal{S}\}$ **do**

2: Generate two independent augmented copies of each sample $X_i' = \mathcal{T}(X_i)$ and $X_i'' = \mathcal{T}(X_i)$ for $i \in \mathcal{S}$.

3: Evaluate representation of each augmented sample $Z' = \{\Theta_\beta(X_i')\}_{i\in\mathcal{S}}$ and $Z'' = \{\Theta_\beta(X_i'')\}_{i\in\mathcal{S}}$, where $Z', Z'' \in \mathbb{R}^{n'\times S}$.

4: Evaluate the kernel matrix $W = \{\mathbf{I}(\|X_i' - X_j''\| \le r)\}_{i,j\in\mathcal{S}} \in \mathbb{R}^{n'\times n'}$ and corresponding Laplacian matrix L.

5: Evaluate the loss

$$\hat{\ell}(\beta) = \mathrm{tr}({Z'}^T L Z'') + \lambda_1 \|Z' - Z''\|_F^2 + \lambda_2 \|{Z'}^T Z'' - I_S\|_F^2,$$

where I_S is a $S \times S$ identify matrix and $\|\cdot\|_F$ is Frobenius norm of a matrix.

6: Update Θ_β to minimize $\hat{\ell}(\beta)$.

7: **end for**

Output: Encoder Θ_β

Appendix B. Proof

In this proof, $\|\cdot\|$ represents the Euclidean distance and $d_{\mathcal{M}}(\cdot,\cdot)$ represents the geodesic distance on the manifold $\mathcal{M}$. C refers to be a constant which can be different in different places.

B.1 Proof Sketches

Proof sketches for Theorem 1, 2, and 4 The proof strategy in Theorem 1, 2, and 4 adopts the same variational method as in Burago et al. (2015); García Trillos et al. (2020); Calder and García Trillos (2022); García Trillos et al. (2021). The proof can be divided into eight steps.

Step 1: *Dirichlet energy* In this step, we define the Dirichlet energy for continuous function θ. The eigenfunctions of Dirichlet energy are eigenfunctions of the Laplacian operator of interest because $D(\theta) = \langle \theta, \Delta\theta \rangle$. Different theorems define the function θ and the Laplacian operator Δ differently.

Step 2: *Discrete Dirichlet energy* In this step, we introduce the discrete version of Dirichlet energy $b(U)$ for a discrete function $U \in \mathbb{R}^n$, which is the discrete counterpart of Dirichlet energy $D(\theta)$. The discrete Dirichlet energy is defined using the graph Laplacian derived from the observed data. In some theorems, we also need to introduce

intermediate discrete Dirichlet energy that can help establish the connection between continuous and discrete Dirichlet energy in step 4.

Step 3: *Discretization and interpolation maps* This step connects discrete and continuous functions via discretization and interpolation maps. The discretization map $\tilde{P}$ converts a continuous function to a discrete one, while the interpolation map $\tilde{\mathcal{I}}$ transforms a discrete to a continuous one. We can show that the construction of discretization and interpolation maps can approximately preserve the norm of functions

$$\left|\|\theta\|^2 - \|\tilde{P}\theta\|^2\right| = o(\|\theta\|\sqrt{D(\theta)} + \|\theta\|^2) \text{ and } \left|\|U\|^2 - \|\tilde{\mathcal{I}}U\|^2\right| = o(\|U\|\sqrt{b(U)} + \|U\|^2).$$

Step 4: *Connection between continuous and discrete Dirichlet energy* This step aims to establish the connection between the Dirichlet energy $D(\theta)$ and its discrete counterpart $b(U)$. Specifically, we aim to establish the inequalities with the following form:

$$b(\tilde{P}\theta) \leq (1 + o(1))\, D(\theta) + o(1) \quad \text{and} \quad D(\tilde{\mathcal{I}}U) \leq (1 + o(1))\, b(U).$$

Different strategies are needed when $D(\theta)$ and $b(U)$ are defined differently.

Step 5: *Upper bound of eigenvalues* In this step, we characterize the upper bound for eigenvalues of discrete Dirichlet energy, $\lambda_s(L)$, via the connection between continuous and discrete Dirichlet energy. Specifically, we define an s-dimensional subspace $S^o = \left\{\sum_{j=1}^s a_j \tilde{P}\theta^j : a_j \in \mathbb{R}\right\}$, where $\theta^1, \ldots, \theta^s$ is an orthonormal set of eigenfunctions of $D(\theta)$. By the Courant-Fisher minimax principle, we have

$$\lambda_s(L) \leq \max_{U \in S^o, \|U\| \leq 1} b(U) \leq \max_{\|a\| \leq 1 + o(1)} b\left(\tilde{P}\theta(a)\right),$$

where $\theta(a) = \sum_{j=1}^s a_j \theta^j$. If we apply the connection between continuous and discrete Dirichlet energy, we can easily get

$$\lambda_s(L) \leq (1 + o(1)) \max_{\|a\| \leq 1 + o(1)} D\left(\theta(a)\right) + o(1) \leq (1 + o(1))\, \lambda_s(\mathcal{M}) + o(1).$$

Here, $\lambda_s(\mathcal{M})$ is the eigenvalue of $D(\theta)$.

Step 6: *Lower bound of eigenvalues* In this step, we characterize the lower bound for eigenvalues of discrete Dirichlet energy with a strategy similar to the last step. Define an s-dimensional subspace $S^o = \left\{\sum_{j=1}^s a_j \tilde{\mathcal{I}}U^j : a_j \in \mathbb{R}\right\}$, where $U^1, \ldots, U^s$ is an orthonormal set of eigenfunctions of $b(U)$. We can again apply the Courant-Fisher minimax principle to show

$$\lambda_s(\mathcal{M}) \leq \max_{\theta \in S^o, \|\theta\| \leq 1} D(\theta) \leq \max_{\|a\| \leq 1 + o(1)} D\left(\tilde{\mathcal{I}}U(a)\right),$$

where $U(a) = \sum_{j=1}^s a_j U^j$. An application of the connection between continuous and discrete Dirichlet energy suggests

$$\lambda_s(\mathcal{M}) \leq (1 + o(1)) \max_{\|a\| \leq 1 + o(1)} b\left(U(a)\right) \leq (1 + o(1))\, \lambda_s(L).$$

Step 7: *Convergence of the first K eigenvectors* In this step, we study the convergence of the first K eigenvectors of $b(U)$. The convergence of eigenvalues in steps 5 and 6 suggests $\lambda_s(L) = o(1)$ for $s = 1, \ldots, K$ and $\lambda_{K+1}(L) = (1 + o(1))\lambda_{K+1}(\mathcal{M})$. Let S be the subspace spanned by the first K eigenvectors of $b(U)$, $\mathcal{P}_S$ be the projection operator on S, and $\theta^1, \ldots, \theta^K$ be a set of orthonormal basis for the eigenspace of eigenfunctions of $D(\theta)$. We can first show that $\tilde{P}\theta^j$ almost belongs to the subspace S

$$\|\tilde{P}\theta^j - \mathcal{P}_S\tilde{P}\theta^j\| \leq \frac{b(\tilde{P}\theta^j)}{\lambda_{K+1}(L)} \leq \frac{(1 + o(1))D(\theta^j) + o(1)}{(1 + o(1))\lambda_{K+1}(\mathcal{M})} = o(1).$$

Since discretization map $\tilde{P}$ can preserve the norm of the orthonormal basis, we can show that $\tilde{P}\theta^j$ is approximately orthonormal, i.e., $|\langle \tilde{P}\theta^{j_1}, \tilde{P}\theta^{j_2}\rangle - \delta_{j_1,j_2}| = o(1)$ where δ_{j_1,j_2} is Kronecker delta and $1 \leq j_1 < j_2 \leq K$. Because $\tilde{P}\theta^j$ almost belongs to the subspace S and is approximately orthonormal, an orthonormal basis of S, named $V^1, \ldots, V^K$ exists, such that $\|\tilde{P}\theta^j - V^j\| = o(1)$. If we write $\vec{\theta} = (\theta(X_1), \ldots, \theta(X_n))$, we can also show that $\vec{\theta^j}$ and $\tilde{P}\theta^j$ are close, i.e., $\|\vec{\theta^j} - \tilde{P}\theta^j\| = o(1)$. Therefore, we show that $\|\vec{\theta^j} - V^j\| = o(1)$ for $j = 1, \ldots, K$. Equivalently, if we apply some rotation matrix, we can show the convergence of the first K eigenvectors.

Step 8: *Convergence of the rest eigenvectors* In this step, we study the convergence of the rest eigenvectors. The analysis is similar to the last step, but we need more involved arguments to show

$$\|\tilde{P}\theta^j - \mathcal{P}_S\tilde{P}\theta^j\| = o(1),$$

where S is the corresponding subspace spanned by eigenvectors of $b(U)$, $\mathcal{P}_S$ is the projection operator on S, and θ^j is the normalized eigenfunctions of $D(\theta)$ with eigenvalue equal to λ.

Since the continuous and discrete Dirichlet energy is defined differently in steps 1 and 2, we need to adopt different strategies to build a connection between continuous and discrete Dirichlet energy in step 4 of different theorems. Once the connection is established, the analyses in steps 3, 5, 6, 7, and 8 are almost the same, so we only present them in the proof for Theorem 1.

Proof sketches for Theorem 3 and 5 The information-theoretic lower bound is proved in a standard way. See more discussions in Tsybakov (2008). The proof is divided into three steps.

Step 1: *Hypothesis construction* In this step, we construct the least favorable null and alternative hypothesis. The construction is slightly different in the proof of Theorem 3 and 5. Under the null hypothesis, we assume the observed data is drawn from a single manifold and write the probability distribution of observed data as $\mathbb{P}_0$. Under the alternative hypothesis, we assume the observed data is drawn from two disjoint manifolds that are very close to each other, and the union of these two manifolds is quite similar to the single manifold under the null hypotheses. Since there are many ways to construct the two manifolds, we consider a mixture distribution of these alternative hypothesis and the probability distribution of observed data as $\mathbb{P}_1$.

Step 2: *Bounding χ^2 divergence* In this step, we bound the χ^2 divergence between the distributions under null and alternative hypotheses. Specifically, we aim to find a setting such that

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \to 0.$$

Step 3: *Applying theorem of fuzzy hypothesis* Once the bound of χ^2 divergence is established, we can apply the theorem of the fuzzy hypothesis (Tsybakov, 2008).

The construction of the least favorable hypothesis and the way to bound χ^2 divergence differ in different theorems.

B.2 Proof for Theorem 1

Instead of proving Theorem 1 directly, we prove a more general version of Theorem 1.

Theorem 7 *Suppose Assumption 1 holds and for any $1 \le k \le K$*

$$\pi_k\left(B(x,r)\bigcap \mathcal{M}_k\right) \le \beta, \qquad x \in \mathcal{M}\setminus \mathcal{M}_k,$$

where $B(x,r)$ is a ball centered at x with radius r (in Euclidean distance). If we assume $r \to 0$,

$$r \gg \left(\frac{\log n}{n}\right)^{1/d} \qquad \text{and} \qquad \frac{\beta}{r^{d+2}} + \frac{\sqrt{\beta}}{nr^{d+2}} \to 0,$$

then with probability at least $1 - K^2/t^2 - 4Kn^{-\alpha}$ for some $\alpha > 0$, we have

$$|\lambda_s(L) - \lambda_s(\mathcal{M})| \le e_1\lambda_2(\mathcal{M}) + e_2,$$

where $\lambda_s(L)$ is sth eigenvalue of normalized matrix L in $b(U)$ and $\lambda_s(\mathcal{M})$ is sth eigenvalue of $\Delta_{\mathcal{M}}$. Here, e_1 and e_2 are defined as

$$e_1 = C\left(r\sqrt{\lambda_s(\mathcal{M})} + \gamma + r + \frac{\delta}{r}\right) \quad \text{and} \quad e_2 = C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda_s^{d/2+2}(\mathcal{M})\right).$$

Here, we choose $\delta = \sqrt{r(c\log n/n)^{1/d}}$ and $\gamma = \sqrt{(\alpha+1)\log n/n\delta^d}$ for some large enough α so $\delta, \gamma, \delta/r \to 0$. With probability at least $1 - K^2/t^2 - 4Kn^{-\alpha}$, if U_s is normalized eigenvector of L with eigenvalue $\lambda_s(L)$, there is a normalized eigenfunction θ_s of $\Delta_{\mathcal{M}}$ with eigenvalue $\lambda_s(\mathcal{M})$ such that

$$\|U_s - \vec{\theta}_s\|_{L^2(\pi_n)} \le \begin{cases} C\left(\gamma + \delta + \frac{1}{\sqrt{\gamma_\lambda}}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)^{1/2}\right), & s \le K \\ C\left(\frac{\lambda}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right) + \left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2}\right)^{1/2} + C\delta, & s > K \end{cases},$$

where $\lambda = \lambda_s(\mathcal{M})$, $\vec{\theta}_s = (\theta_s(X_1), \ldots, \theta_s(X_n))$ and γ_λ is the eigengap.

This theorem can lead to Theorem 1 if we choose $\beta = 0$. This theorem also suggests that the first K eigenvectors converge faster than the rest of eigenvectors.

Step 1: Dirichlet energy To study the convergence of graph Laplacian, we need to introduce some notations. For a function $\theta : \mathcal{M} \to \mathbb{R}$ defined on $\mathcal{M}$, we also write it in the following form

$$\theta(x) = (\theta_1(x), \ldots, \theta_K(x)),$$

where each $\theta_k : \mathcal{M}_k \to \mathbb{R}$ is a function defined on each manifold $\mathcal{M}_k$. Given two functions θ^A and θ^B defined on $\mathcal{M}$, we define their inner product as

$$\langle \theta^A, \theta^B \rangle_{L^2(\pi)} = \sum_{k=1}^{K} w_k \int_{\mathcal{M}_k} \theta_k^A(x)\theta_k^B(x)\pi_k(x)dx.$$

We also define the weighted Dirichlet energy for a function θ as

$$D(\theta) = \sum_{k=1}^{K} w_k^2 D_k(\theta_k), \qquad \text{where} \quad D_k(\theta_k) = \int_{\mathcal{M}_k} \|\nabla\theta_k(x)\|^2 \pi_k^2(x)dx.$$

From this definition, it is straightforward to see $D(\theta) = \langle \theta, \Delta_{\mathcal{M}}\theta \rangle$.

Step 2: Discrete Dirichlet energy To study graph Laplacian, we now introduce a more general neighborhood graph and the corresponding graph Laplacian. Specifically, when $\|X_i - X_j\| \le r$, we assign weights

$$W_{i,j} = h\left(\frac{\|X_i - X_j\|}{r}\right).$$

where h is a function supported on $[0, 1]$. Clearly, $h(x) = \mathbf{I}(\|x\| \le 1)$ used in Section 3 is a special case. Given the general weights, we can define discrete Dirichlet energy as

$$b(U) = \frac{1}{n^2}\sum_{i,j} \frac{1}{\sigma_h r^d} W_{i,j} \left(\frac{U(X_i) - U(X_j)}{r}\right)^2 = \frac{1}{\sigma_h n^2 r^{d+2}} U^T L U,$$

where $U(X_i)$ is the ith component of vector U (corresponds to X_i) and σ_h is the surface tension of h

$$\sigma_h = \int |y_1|^2 h(|y|)dy.$$

In particular, when $h(x) = \mathbf{I}(\|x\| \le 1)$, $\sigma_h = V_d/(d+2)$, where V_d is the volume of d-dimensional Euclidean unit ball. Although $U \in \mathbb{R}^n$, it can also be considered as a function defined on discrete point $\{X_1, \ldots, X_n\}$. If we write the empirical measure of $\{X_1, \ldots, X_n\}$ as

$$\pi_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{X_i},$$

then we can define the inner product between two discrete functions

$$\langle U^A, U^B \rangle_{L^2(\pi_n)} = \frac{1}{n}\sum_{i=1}^{n} U^A(X_i)U^B(X_i).$$

Similarly, we can define within and cross manifold Dirichlet energy

$$b_k(U) = \frac{1}{n_k^2} \sum_{X_i, X_j \in \mathcal{M}_k} \frac{1}{\sigma_h r^d} W_{i,j} \left(\frac{U(X_i) - U(X_j)}{r} \right)^2, \qquad b_W(U) = \sum_{k=1}^{K} \left(\frac{n_k}{n} \right)^2 b_k(U),$$

and

$$b_C(U) = \frac{1}{n^2} \sum_{k_1 \neq k_2} \sum_{X_i \in \mathcal{M}_{k_1}, X_j \in \mathcal{M}_{k_2}} \frac{1}{\sigma_h r^d} W_{i,j} \left(\frac{U(X_i) - U(X_j)}{r} \right)^2.$$

Here, n_k is the number of point on $\mathcal{M}_k$, i.e., $n_k = |\{i : X_i \in \mathcal{M}_k\}|$.

Step 3: Discretization and interpolation maps After introducing the notations, we now construct the relationship between Dirichlet energy $D(\theta)$ and $b(U)$ via discretization and interpolation maps. The Proposition 2.12 in Calder and García Trillos (2022) or Corollary A.3 in García Trillos et al. (2021) suggests the following Proposition directly. We omit the proof.

Proposition 8 *With probability at least* $1 - Kn \exp(-Cn\gamma^2\delta^d) - 2K \exp(-cn)$*, there exists a probability measure* $\tilde{\mu}_{n,k}$ *with probability density function* $\tilde{\pi}_{n,k}$ *for* $k = 1, \ldots, K$ *such that*

$$\|\pi_k - \tilde{\pi}_{n,k}\|_{L^\infty(\mathcal{M}_k)} \leq C(\gamma + \delta)$$

and there exist transportation maps $\tilde{R}_1, \ldots, \tilde{R}_K$ *such that*

$$\sup_{x \in \mathcal{M}_k} d_{\mathcal{M}_k}(x, \tilde{R}_k(x)) \leq \delta.$$

In Proposition 8, we can choose $\delta = \sqrt{r(c \log n/n)^{1/d}}$ and $\gamma = \sqrt{(\alpha+1)\log n/n\delta^d}$ for some large enough α. Based on the transportation maps $\tilde{R}_1, \ldots, \tilde{R}_K$ in Proposition 8, we can also introduce the discretization and interpolation maps. First, we can define a partition of $\mathcal{M}$ by $\tilde{R}_1, \ldots, \tilde{R}_K$. Specifically, if $X_i \in \mathcal{M}_k$, then

$$\tilde{U}_i = \tilde{R}_k^{-1}(X_i) \subset \mathcal{M}_k.$$

After defining the partition, we can introduce the discretization map $\tilde{P} : L^2(\pi) \to L^2(\pi_n)$

$$\tilde{P}\theta(X_i) = n_k \int_{\tilde{U}_i} \theta(x) \tilde{\pi}_{n,k}(x) dx,$$

where $X_i \in \mathcal{M}_k$. Similarly, we can introduce the associated extension map $\tilde{P}^* : L^2(\pi_n) \to L^2(\pi)$

$$\tilde{P}^* U(x) = \sum_{i=1}^{n} u(X_i) \mathbf{I}(x \in \tilde{U}_i).$$

We then define the interpolation map $\tilde{\mathcal{I}}$

$$\tilde{\mathcal{I}} u = \Lambda_{r-2\delta} \tilde{P}^* u$$

where $\Lambda_{r-2\delta}$ is a convolution operator defined on $\mathcal{M}$. More concretely, we define

$$\Lambda_r\theta(x) = \frac{\int K_r(x,y)\theta_k(y)dy}{\int K_r(x,y)dy}, \qquad x \in \mathcal{M}_k,$$

where the kernel $K_r(x,y)$ is defined as

$$K_r(x,y) = \frac{1}{r^d}\psi\left(\frac{d_{\mathcal{M}}(x,y)}{r}\right) \qquad \text{and} \qquad \psi(t) = \frac{1}{\sigma_\eta}\int_t^\infty h(s)ds.$$

Step 4: Connection between Dirichlet energy $D(\theta)$ and $b(U)$ With the newly introduced discretization and interpolation maps, we can connect the Dirichlet energy $D(\theta)$ and $b(U)$. We can directly apply the results from Proposition 4.1 and 4.2 in Calder and García Trillos (2022) or Proposition A.4 and A.5 in García Trillos et al. (2021) to show the following Proposition.

Proposition 9 *Suppose γ and δ are two parameters in Proposition 8. With probability at least $1 - Kn\exp(-Cn\gamma^2\delta^d) - 2K\exp(-cn)$, we have*

$$b_W(\tilde{P}\theta) \le \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right) D(\theta) \quad \text{and} \quad D(\tilde{\mathcal{I}}U) \le \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right) b(U). \tag{8}$$

In addition, we also have

$$\left|\|\theta\|^2_{L^2(\pi)} - \|\tilde{P}\theta\|^2_{L^2(\pi_n)}\right| \le C\delta\|\theta\|_{L^2(\pi)}\sqrt{D(\theta)} + C(\gamma+\delta)\|\theta\|^2_{L^2(\pi)} \tag{9}$$

and

$$\left|\|U\|^2_{L^2(\pi_n)} - \|\tilde{\mathcal{I}}U\|^2_{L^2(\pi)}\right| \le Cr\|U\|_{L^2(\pi_n)}\sqrt{b(U)} + C(\gamma+\delta)\|U\|^2_{L^2(\pi_n)}. \tag{10}$$

Here, $\theta \in L^2(\pi)$ and $U \in L^2(\pi_n)$.

To connect Dirichlet energy $D(\theta)$ and $b(U)$, we also need to find an upper bound for the cross manifold Dirichlet energy.

Proposition 10 *Suppose Assumption 1 holds and for any $1 \le k \le K$*

$$\pi_k\left(B(x,r)\bigcap \mathcal{M}_k\right) \le \beta, \qquad \forall\ x \in \mathcal{M}\setminus\mathcal{M}_k,$$

where $B(x,r)$ is a ball centered at x with radius r (in Euclidean distance). When θ is in the span of $\Delta_{\mathcal{M}}$'s eigenfunctions with corresponding eigenvalue less than λ, then we have

$$b_C(\tilde{P}\theta) \le C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda^{d/2+2}\right)\|\theta\|^2_{L^2(\pi)},$$

with probability at least $1 - K^2/t^2 - 2K\exp(-cn)$.

Step 5: Upper bound of $\lambda_s(L)$ To show the convergence of eigenvalues, we now construct an upper bound for $\lambda_s(L)$ in terms of $\lambda_s(\mathcal{M})$. Let $\theta^1, \ldots, \theta^s$ be an orthonormal set of eigenfunctions of $\Delta_{\mathcal{M}}$. Then, we can define

$$V^j = \tilde{P}\theta^j, \qquad j = 1, \ldots, s.$$

We can apply (9) to $\theta^{j_1} - \theta^{j_2}$, so we have

$$\begin{aligned}&\left|\|\theta^{j_1}\|^2_{L^2(\pi)} + \|\theta^{j_2}\|^2_{L^2(\pi)} - 2\langle\theta^{j_1},\theta^{j_2}\rangle_{L^2(\pi)} - \|V^{j_1}\|^2_{L^2(\pi_n)} - \|V^{j_2}\|^2_{L^2(\pi_n)} + 2\langle V^{j_1},V^{j_2}\rangle_{L^2(\pi_n)}\right| \\ \leq& C\delta\sqrt{\lambda_s(\mathcal{M})} + C(\gamma+\delta)\end{aligned}$$

So we can conclude that for some large enough n, we have

$$\left|\langle\theta^{j_1},\theta^{j_2}\rangle_{L^2(\pi)} - \langle V^{j_1},V^{j_2}\rangle_{L^2(\pi_n)}\right| \leq C\delta\sqrt{\lambda_s(\mathcal{M})} + C(\gamma+\delta) < \frac{1}{s}.$$

Because $\langle\theta^{j_1},\theta^{j_2}\rangle_{L^2(\pi)} = 1$ when $j_1 = j_2$ and $\langle\theta^{j_1},\theta^{j_2}\rangle = 0$ when $j_1 \neq j_2$, we can know $V^1,\ldots,V^s$ are linearly independent. Based on $V^1,\ldots,V^s$, we define a s-dimensional subspace

$$S^o = \left\{\sum_{j=1}^s a_j V^j : a_j \in \mathbb{R}\right\}.$$

By the Courant-Fisher minimax principle, we know

$$\lambda_s(L) = \min_{S\in\mathcal{S}_s}\max_{U\in S\setminus\{0\}}\frac{b(U)}{\|U\|^2_{L^2(\pi_n)}},$$

where $\mathcal{S}_s$ is the collection of all possible s-dimensional subspaces. This immediately suggests

$$\lambda_s(L) \leq \max_{U\in S^o, \|U\|_{L^2(\pi_n)}=1} b(U).$$

For any $U \in S^o$, we can apply (8) and Proposition 10 to obtain

$$\begin{aligned}b(U) = b(\tilde{P}\theta) &\leq \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right)D(\theta) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1+\lambda^{d/2+2}\right)\|\theta\|^2_{L^2(\pi)} \\ &\leq \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right)\lambda_s(\mathcal{M})\|\theta\|^2_{L^2(\pi)} + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1+\lambda^{d/2+2}\right)\|\theta\|^2_{L^2(\pi)}.\end{aligned}$$

where $\theta = \sum_{j=1}^s a_j\theta^j$ if $U = \sum_{j=1}^s a_j V^j$ and $\lambda = \lambda_s(\mathcal{M})$. An application of (9) suggests

$$\|\theta\|^2_{L^2(\pi)} \leq \|U\|^2_{L^2(\pi_n)} + C(\delta\sqrt{\lambda_s(\mathcal{M})} + \gamma + \delta)\|\theta\|^2_{L^2(\pi)}.$$

If $\|U\|_{L^2(\pi_n)} = 1$, then

$$\|\theta\|^2_{L^2(\pi)} \leq 1 + C(\delta\sqrt{\lambda_s(\mathcal{M})} + \gamma + \delta).$$

This leads to

$$\lambda_s(L) \leq \left(1 + C\left(\delta\sqrt{\lambda_s(\mathcal{M})} + \gamma + r + \frac{\delta}{r}\right)\right)\lambda_s(\mathcal{M}) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda_s^{d/2+2}(\mathcal{M})\right).$$

Step 6: Lower bound of $\lambda_s(L)$ We can find a lower bound for $\lambda_s(L)$ in terms of $\lambda_s(\mathcal{M})$ with a similar idea. Let $U^1,\ldots,U^s$ be an orthonormal set of eigenvectors of L and

$$\theta^j = \tilde{\mathcal{I}}U^j, \qquad j = 1,\ldots,s.$$

We can apply (10) to $U^{j_1} - U^{j_2}$

$$\begin{aligned}&\left|\|\theta^{j_1}\|^2_{L^2(\pi)} + \|\theta^{j_2}\|^2_{L^2(\pi)} - 2\langle\theta^{j_1},\theta^{j_2}\rangle_{L^2(\pi)} - \|U^{j_1}\|^2_{L^2(\pi_n)} - \|U^{j_2}\|^2_{L^2(\pi_n)} + 2\langle U^{j_1},U^{j_2}\rangle_{L^2(\pi_n)}\right| \\ \leq& Cr\sqrt{\lambda_s(L)} + C(\gamma+\delta)\end{aligned}$$

So we can conclude that for some large enough n, we have

$$\left|\langle\theta^{j_1},\theta^{j_2}\rangle_{L^2(\pi)} - \langle U^{j_1},U^{j_2}\rangle_{L^2(\pi_n)}\right| \leq Cr\sqrt{\lambda_s(L)} + C(\gamma+\delta) < \frac{1}{s}.$$

Because $\langle U^{j_1},U^{j_2}\rangle_{L^2(\pi_n)} = 1$ when $j_1 = j_2$ and $\langle U^{j_1},U^{j_2}\rangle_{L^2(\pi_n)} = 0$ when $j_1 \neq j_2$, we can know $\theta^1,\ldots,\theta^s$ are linearly independent. With $\theta^1,\ldots,\theta^s$, we define

$$S^o = \left\{\sum_{j=1}^s a_j\theta^j : a_j \in \mathbb{R}\right\}.$$

Again, by the Courant-Fisher minimax principle,

$$\lambda_s(\mathcal{M}) = \min_{S\in\mathcal{S}_s}\max_{\theta\in S\setminus\{0\}}\frac{D(\theta)}{\|\theta\|^2_{L^2(\pi)}} \leq \max_{\theta\in S^o,\|\theta\|_{L^2(\pi)}=1} D(\theta).$$

For $\theta \in S^o$, there exist a U such that $\theta = \tilde{\mathcal{I}}U$ where $U = \sum_{j=1}^s a_jU^j$. If we apply (8), then we have

$$D(\theta) = b(\tilde{\mathcal{I}}U) \leq \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right) b(U) \leq \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right)\lambda_s(L)\|U\|^2_{L^2(\pi_n)}.$$

An application of (8) again suggests that if $\|\theta\|_{L^2(\pi)} = 1$, then

$$\|U\|^2_{L^2(\pi_n)} \leq 1 + C(r\sqrt{\lambda_s(L)} + \gamma + \delta).$$

Now we can conclude

$$\lambda_s(\mathcal{M}) \leq \left(1 + C\left(r\sqrt{\lambda_s(L)} + \gamma + r + \frac{\delta}{r}\right)\right)\lambda_s(L).$$

Putting upper and lower bounds together yields

$$|\lambda_s(\mathcal{M}) - \lambda_s(L)| \leq C\left(r\sqrt{\lambda_s(\mathcal{M})} + \gamma + r + \frac{\delta}{r}\right)\lambda_s(\mathcal{M}) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda_s^{d/2+2}(\mathcal{M})\right).$$

In the rest of proof, we use the following notations

$$e_1 = C\left(r\sqrt{\lambda_s(\mathcal{M})} + \gamma + r + \frac{\delta}{r}\right) \qquad \text{and} \qquad e_2 = C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda_s^{d/2+2}(\mathcal{M})\right).$$

Step 7: Convergence of the first K eigenvectors After showing the convergence of eigenvalues, we now study the convergence of eigenvectors. We first show the convergence of the

eigenvectors corresponding to the first K eigenvalues ($\lambda_1(\mathcal{M}) = \ldots = \lambda_K(\mathcal{M}) = 0$). Let γ_λ be the eigengap, i.e., $\gamma_\lambda = \lambda_{K+1}(\mathcal{M})$. When n is large enough, we have

$$|\lambda_s(\mathcal{M}) - \lambda_s(L)| \leq \frac{\gamma_\lambda}{4}, \qquad \text{when} \quad s = 1, \ldots, K+1.$$

Let S be the subspace spanned by the eigenvectors of L with eigenvalues $\lambda_1(L), \ldots, \lambda_K(L)$ and $\mathcal{P}_S$ be the projection operator on S. Let θ be a normalized eigenvector of $\Delta_\mathcal{M}$ with eigenvalue 0 and $U = \tilde{P}\theta$. By Proposition 9 and 10, we can know that

$$b(U) \leq C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right).$$

The definition of eigenvector suggests

$$b(U) = \frac{1}{\sigma_h n^2 r^{d+2}} U^T L U \geq \lambda_1(L)\|\mathcal{P}_S U\|^2_{L^2(\pi_n)} + \lambda_{K+1}(L)\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} = \lambda_{K+1}(L)\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)}.$$

Putting above two inequalities together yields

$$\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} \leq \frac{C}{\gamma_\lambda}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right). \tag{11}$$

By the definition of $\tilde{P}$, we can know that if $X_i \in \mathcal{M}_k$, then

$$|U(X_i) - \theta(X_i)| = \left| n_k \int_{\tilde{U}_i} (\theta(x) - \theta(X_i))\, \tilde{\pi}_{n,k}(x) dx \right| \leq \|\nabla \theta_k\|_\infty \delta.$$

Since θ is the eigenvector of $\Delta_\mathcal{M}$ with eigenvalue 0, then we can know $\|\nabla\theta_k\|_\infty = 0$ and $U = \vec{\theta}$, where $\vec{\theta} = (\theta(X_1), \ldots, \theta(X_n)) \in \mathbb{R}^n$ is a n dimensional vector. So we can know

$$\|\vec{\theta} - \mathcal{P}_S \tilde{P}\theta\|^2_{L^2(\pi_n)} \leq \frac{C}{\gamma_\lambda}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right). \tag{12}$$

Let $\theta^1, \ldots, \theta^K$ be a set of orthonormal basis for the eigenspace of eigenfunctions of $\Delta_\mathcal{M}$ with eigenvalue 0, $\tilde{U}^j = \tilde{P}\theta^j$ and $\tilde{V}^j = \mathcal{P}_S \tilde{P}\theta^j$. By (12), we have

$$\|\vec{\theta^j} - \tilde{V}^j\|^2_{L^2(\pi_n)} = \|\tilde{U}^j - \tilde{V}^j\|^2_{L^2(\pi_n)} \leq \frac{C}{\gamma_\lambda}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right).$$

An application of (9) on $\theta^{j_1} - \theta^{j_2}$ suggests

$$\left|\langle \tilde{U}^{j_1}, \tilde{U}^{j_2}\rangle_{L^2(\pi_n)} - \langle \theta^{j_1}, \theta^{j_2}\rangle_{L^2(\pi)}\right| \leq C(\gamma + \delta).$$

We combine (11) and above inequality to obtain

$$\left|\langle \tilde{V}^{j_1}, \tilde{V}^{j_2}\rangle_{L^2(\pi_n)} - \langle \theta^{j_1}, \theta^{j_2}\rangle_{L^2(\pi)}\right| \leq C\left(\gamma + \delta + \frac{1}{\sqrt{\gamma_\lambda}}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)^{1/2}\right).$$

Let $V^1, \ldots, V^K$ be the Gram-Schmidt orthogonalization of $\tilde{V}^1, \ldots, \tilde{V}^K$. Then we can know that

$$\|\tilde{V}^j - V^j\|_{L^2(\pi_n)} \leq C\left(\gamma + \delta + \frac{1}{\sqrt{\gamma_\lambda}}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)^{1/2}\right).$$

Therefore, we can conclude that

$$\|V^j - \vec{\theta^j}\|_{L^2(\pi_n)} \leq C\left(\gamma + \delta + \frac{1}{\sqrt{\gamma_\lambda}}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)^{1/2}\right).$$

Equivalently, if we apply some rotation matrix, we can find an orthonormal set $\theta'_1, \ldots, \theta'_K$ of eigenfunctions of $\Delta_{\mathcal{M}}$ with eigenvalues 0 such that

$$\|U_j - \vec{\theta'_j}\|_{L^2(\pi_n)} \leq C\left(\gamma + \delta + \frac{1}{\sqrt{\gamma_\lambda}}\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)^{1/2}\right).$$

Step 8: Convergence of the rest eigenvectors We next study the convergence of the other eigenvectors by induction. In particular, let $\lambda_{s-1}(\mathcal{M}) < \lambda = \lambda_s(\mathcal{M}) = \ldots = \lambda_{s+l-1}(\mathcal{M}) < \lambda_{s+l}(\mathcal{M})$. Suppose $\theta^1, \ldots, \theta^{s-1}$ are a set of orthonormal basis for the eigenspace of eigenfunctions of $\Delta_{\mathcal{M}}$ with eigenvalue smaller than λ and there exists a set of orthonormal basis for subspace spanned by the eigenvectors of L with eigenvalues smaller than λ such that

$$\|V^j - \tilde{P}\theta^j\|_{L^2(\pi_n)} \leq C\left(\frac{1}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda^{d/2+2}\right)\right)^{1/2} + C\delta \tag{13}$$

for $j = 1, \ldots, s-1$. Let S be the subspace spanned by the eigenvectors of L with eigenvalues $\lambda_s(L), \ldots, \lambda_{s+l-1}(L)$ and $\mathcal{P}_S$ be the projection operator on S. Similarly, we define S_+ as the subspace spanned by the eigenvectors of L with eigenvalues larger than $\lambda_{s+l-1}(L)$ and S_- as the subspace spanned by the eigenvectors of L with eigenvalues smaller than $\lambda_s(L)$. $\mathcal{P}_{S_+}$ and $\mathcal{P}_{S_-}$ are the projection operator on S_+ and S_-. Let θ be an eigenvector of $\Delta_{\mathcal{M}}$ with eigenvalue λ and let $U = \tilde{P}\theta$. By Proposition 9 and 10, we have

$$b(U) \leq C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda^{d/2+2}\right) + \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right)\lambda.$$

By the spectrum decomposition, we have

$$\begin{aligned}
b(U) &= \frac{1}{\sigma_h n^2 r^{d+2}} U^T L U \\
&\geq \lambda_s(L)\|\mathcal{P}_S U\|^2_{L^2(\pi_n)} + \lambda_{s+l}(L)\|\mathcal{P}_{S_+} U\|^2_{L^2(\pi_n)} \\
&\geq \lambda_s(L)\left(\|U\|^2_{L^2(\pi_n)} - \|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)}\right) + \lambda_{s+l}(L)\left(\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} - \|\mathcal{P}_{S_-} U\|^2_{L^2(\pi_n)}\right) \\
&\geq \lambda_s(L)\|U\|^2_{L^2(\pi_n)} + (\lambda_{s+l}(L) - \lambda_s(L))\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} - \lambda_{s+l}(L)\|\mathcal{P}_{S_-} U\|^2_{L^2(\pi_n)}.
\end{aligned}$$

By (9), we can know

$$\left|\|U\|^2_{L^2(\pi_n)} - 1\right| \leq C\delta\sqrt{\lambda} + C(\gamma + \delta).$$

The results in step 6 suggests

$$|\lambda - \lambda_k(L)| \leq C\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right)\lambda + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda^{d/2+2}\right).$$

Therefore, we can have

$$\frac{\gamma_\lambda}{2}\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} \leq C\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right)\lambda + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2} + \lambda_{k+l}(L)\|\mathcal{P}_{S_-} U\|^2_{L^2(\pi_n)}.$$

By the choice of orthonormal basis $V^1, \ldots, V^{s-1}$, we have

$$\mathcal{P}_{S_-} u = \sum_{j=1}^{s-1} \langle V^j, U\rangle_{L^2(\pi_n)} V^j.$$

Since (13) and

$$\left|\langle \tilde{P}\theta^j, U\rangle_{L^2(\pi_n)}\right| \leq C\delta\sqrt{\lambda} + C(\gamma + \delta),$$

which is implied by (9), we can know

$$\left|\langle V^j, U\rangle_{L^2(\pi_n)}\right| \leq C\delta\sqrt{\lambda} + C\left(\frac{1}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2}\right)^{1/2}.$$

This leads to

$$\|\mathcal{P}_{S_-} U\|_{L^2(\pi_n)} \leq C\delta\sqrt{\lambda} + C\left(\frac{1}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2}\right)^{1/2}.$$

The convergence of eigenvalue suggests

$$|\lambda_{s+l}(\mathcal{M}) - \lambda_{s+l}(L)| \leq C\left(r\sqrt{\lambda_{k+l}(\mathcal{M})} + \gamma + r + \frac{\delta}{r}\right)\lambda_{s+l}(\mathcal{M}) + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2}_{s+l}(\mathcal{M}).$$

This immediately leads to

$$\frac{\gamma_\lambda}{2}\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} \leq C\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right)\lambda + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2} + C\delta^2.$$

Therefore, we can conclude that

$$\|U - \mathcal{P}_S U\|^2_{L^2(\pi_n)} \leq C\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right)\frac{\lambda}{\gamma_\lambda} + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2} + C\delta^2.$$

By the definition of $\tilde{P}$, we can know that if $X_i \in \mathcal{M}_k$, then

$$|U(X_i) - \theta(X_i)| = \left|n_k \int_{\tilde{U}_i} (\theta(x) - \theta(X_i))\,\tilde{\pi}_{n,k}(x)dx\right| \leq \|\nabla\theta_k\|_\infty \delta.$$

Since θ is the eigenvector of $\Delta_\mathcal{M}$, then we can know $\|\nabla\theta_k\|_\infty$ is finite. So we can know

$$\|\vec{\theta} - \mathcal{P}_S\tilde{P}\theta\|^2_{L^2(\pi_n)} \leq C\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right)\frac{\lambda}{\gamma_\lambda} + C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2} + C\delta^2.$$

Then, we can follow the same argument in step 7 to show that if $U_1, \ldots, U_{s+l-1}$ is an orthonormal basis for the eigenspace of L with eigenvalues smaller than $\lambda_{s+l}(L)$, then there exists an orthonormal basis for the eigenspace of $\Delta_{\mathcal{M}}$ with eigenvalues smaller than $\lambda_{s+l}(\mathcal{M})$ such that

$$\|U_j - \vec{\theta}_j\|_{L^2(\pi_n)} \leq C\left(\frac{\lambda}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r}\right) + \left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\lambda^{d/2+2}\right)^{1/2} + C\delta.$$

B.3 Proof for Theorem 2

In this proof, we adopt the same notations in proof of Theorem 1. Besides the Dirichlet energy $D(\theta)$ and $b(U)$, we define a new discrete Dirichlet energy based on $Y_1, \ldots, Y_n$

$$b_Y(U) = \frac{1}{n^2}\sum_{i,j}\frac{1}{\sigma_h \tilde{r}^d}\tilde{W}_{i,j}\left(\frac{U(Y_i) - U(Y_j)}{\tilde{r}}\right)^2,$$

where $\tilde{r} = \sqrt{r^2 - z^{o2}}$ and the weight $\tilde{W}_{i,j}$ is defined by Y_i and Y_j

$$\tilde{W}_{i,j} = \mathbf{I}\left(\|Y_i - Y_j\| \leq \tilde{r}\right).$$

We also define a Dirichlet energy on $\tilde{\mathcal{M}}$

$$D_Y(\theta) = \int_{\tilde{\mathcal{M}}} \|\nabla\theta(y)\|\pi_Y(y)dy,$$

where π_Y is a uniform distribution on $\tilde{\mathcal{M}}$. Similarly, we can also introduce the discretization map $\tilde{P}_Y$ and interpolation map $\tilde{\mathcal{I}}_Y$ by $Y_1, \ldots, Y_n$ on $\tilde{\mathcal{M}}$. We choose $\delta = \sqrt{r(c\log n/n)^{1/d}}$ and $\gamma = \sqrt{(\alpha+1)\log n/n\delta^d}$ for some large enough α in $\tilde{P}_Y$ and $\tilde{\mathcal{I}}_Y$. Instead of finding the connection between $D(\theta)$ and $b(U)$, we will build the connection between $D_Y(\theta)$ and $b(U)$.

By the construction of $\mathcal{M}_1$ and $\mathcal{M}_2$, we have

$$\|Y_i - Y_j\| \leq \tilde{r} \qquad \Rightarrow \qquad \|X_i - X_j\| \leq r.$$

Equivalently, we have

$$\mathbf{I}\left(\|Y_i - Y_j\| \leq \tilde{r}\right) \leq \mathbf{I}\left(\|X_i - X_j\| \leq r\right) \qquad \text{or} \qquad \tilde{W}_{i,j} \leq W_{i,j}.$$

This suggests

$$\begin{aligned} b_Y(U) &= \frac{1}{n^2}\sum_{i,j}\frac{1}{\sigma_h\tilde{r}^d}\tilde{W}_{i,j}\left(\frac{U(Y_i) - U(Y_j)}{\tilde{r}}\right)^2 \\ &\leq \frac{1}{n^2}\sum_{i,j}\frac{1}{\sigma_h\tilde{r}^d}W_{i,j}\left(\frac{U(Y_i) - U(Y_j)}{\tilde{r}}\right)^2 \\ &\leq \left(\frac{r}{\tilde{r}}\right)^{d+2} b(U). \end{aligned}$$

By Proposition 9, we have

$$\begin{aligned} D_Y(\tilde{\mathcal{I}}_Y U) &\le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma\right)\right) b_Y(U) \\ &\le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma\right)\right)\left(\frac{r}{\tilde{r}}\right)^{d+2} b(U) \\ &\le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma + \left(\frac{z^o}{\tilde{r}}\right)^2\right)\right) b(U). \end{aligned}$$

On the other hand, if we write $U = \tilde{P}_Y \theta$ where θ is some eigenfucntion of $\Delta_{\tilde{\mathcal{M}}}$, then

$$\begin{aligned} &\left(\frac{r}{\tilde{r}}\right)^{d+2} b(U) - b_Y(U) \\ =&\frac{1}{\sigma_h n^2 \tilde{r}^{d+2}} \sum_{Z_i = Z_j} \left(\mathbf{I}(\|Y_i - Y_j\| \le r) - \mathbf{I}(\|Y_i - Y_j\| \le \tilde{r})\right)(U(Y_i) - U(Y_j))^2 \\ =&\frac{1}{\sigma_h n^2 \tilde{r}^{d}} \sum_{Z_i = Z_j} \mathbf{I}(\tilde{r} < \|Y_i - Y_j\| \le r)\left(\frac{U(Y_i) - U(Y_j)}{\tilde{r}}\right)^2 \\ \le&\frac{1}{\sigma_h n^2 \tilde{r}^{d}} \sum_{Z_i = Z_j} \mathbf{I}(\tilde{r} < \|Y_i - Y_j\| \le r)\left(\frac{3r\|\nabla\theta\|_\infty}{\tilde{r}}\right)^2 \\ \le&\frac{9r^2\|\nabla\theta\|_\infty^2}{\sigma_h n^2 \tilde{r}^{d+2}} \sum_{i,j} \mathbf{I}(\tilde{r} < \|Y_i - Y_j\| \le r) \end{aligned}$$

Define a U-statistics

$$Q = \frac{2}{n(n-1)} \sum_{i<j} \mathbf{I}(\tilde{r} < \|Y_i - Y_j\| \le r).$$

Since

$$\Sigma^2 := \mathbb{E}\left(\mathbf{I}^2(\tilde{r} < \|Y_i - Y_j\| \le r)\right) = \mathbb{E}\left(\mathbf{I}(\tilde{r} < \|Y_i - Y_j\| \le r)\right) \le C(r^d - \tilde{r}^d),$$

we can apply a Bernstein-type inequality for U-statistics (Arcones, 1995) to have

$$\mathbb{P}\left(\sqrt{n}|Q - \mathbb{E}(Q)| > t\right) \le 2\exp\left(-\frac{2t^2}{2\Sigma^2 + (2/3)tn^{-1/2}}\right).$$

This suggest

$$\mathbb{P}\left(Q > C(r^d - \tilde{r}^d) + \sqrt{\frac{C(r^d - \tilde{r}^d)\log n}{n}} + \frac{C\log n}{n}\right) \le 2n^{-2}$$

With probability at least $1 - 2n^{-2}$,

$$\left(\frac{r}{\tilde{r}}\right)^{d+2} b(U) - b_Y(U) \le \frac{Cr^2\|\nabla\theta\|_\infty^2}{\sigma_h \tilde{r}^{d+2}}\left((r^d - \tilde{r}^d) + \frac{\log n}{n}\right) \le \frac{C\|\nabla\theta\|_\infty^2}{\sigma_h}\left(\frac{z^{o2}}{r^2} + \frac{\log n}{nr^d}\right).$$

Therefore, we can conclude that

$$b(\tilde{P}_Y\theta) \le b_Y(\tilde{P}_Y\theta) + \frac{C\|\nabla\theta\|_\infty^2}{\sigma_h}\left(\frac{z^{o2}}{r^2} + \frac{\log n}{nr^d}\right)$$
$$\le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma\right)\right) D_Y(\theta) + \frac{C\|\nabla\theta\|_\infty^2}{\sigma_h}\left(\frac{z^{o2}}{r^2} + \frac{\log n}{nr^d}\right).$$

After building the connection between $D_Y(\theta)$ and $b(U)$, we can apply the same arguments in step 5-8 of proof for Theorem 1 to show that, with probability at least $1 - Cn^{-2}$, if U_s is normalized eigenvector of L with sth eigenvalue, there is a normalized eigenfunction θ_s of $\Delta_{\tilde{\mathcal{M}}}$ with sth eigenvalue such that

$$\|U_s - \vec{\theta}_s\|_{L^2(\pi_n)} \le C\left(\frac{\lambda}{\gamma_\lambda}\left(r\sqrt{\lambda} + \gamma + r + \frac{\delta}{r} + \left(\frac{z^o}{\tilde{r}}\right)^2\right) + \frac{C}{\sigma_h}\left(\left(\frac{z^o}{r}\right)^2 + \frac{\log n}{nr^d}\right)\right)^{1/2} + C\delta,$$

where $\lambda = \lambda_s(\tilde{\mathcal{M}})$, $\vec{\theta}_s = (\theta_s(Y_1), \ldots, \theta_s(Y_n))$ and γ_λ is the eigengap. The choices of δ, γ, z^o and r suggest

$$\|U_s - \vec{\theta}_s\|_{L^2(\pi_n)} \to 0.$$

B.4 Proof for Theorem 4

We adopt a similar strategy in Theorem 1 to prove results.

Step 1: Dirichlet energy We need to introduce some notations parallel to the notations in the proof of Theorem 1. Given two functions $\bar{\theta}^A$ and $\bar{\theta}^B$ defined on $\mathcal{N}_s$, we define their inner product as

$$\langle \bar{\theta}^A, \bar{\theta}^B \rangle_\phi = \sum_{k=1}^K w_k \int_{\mathcal{N}_{s,k}} \bar{\theta}_k^A(\phi)\bar{\theta}_k^B(\phi)\pi_k^s(\phi)d\phi.$$

Similarly, we define the weighted Dirichlet energy for a function $\bar{\theta} : \mathcal{N}_s \to \mathbb{R}$ as

$$D_\phi(\bar{\theta}) = \sum_{k=1}^K \frac{w_k^2}{\mathrm{Vol}\mathcal{N}_{v,k}} D_{\phi,k}(\bar{\theta}_k), \qquad \text{where} \quad D_{\phi,k}(\bar{\theta}_k) = \int_{\mathcal{N}_{s,k}} \|\nabla\bar{\theta}_k(\phi)\|^2 \pi_k^{s\,2}(\phi)d\phi.$$

Step 2: Discrete Dirichlet energy Given the new weights, we can define corresponding discrete Dirichlet energy as

$$b_\phi(\bar{U}) = \frac{1}{n^2}\sum_{i,j}\frac{1}{\sigma_h r^d}\bar{W}_{i,j}\left(\frac{\bar{U}(\phi_i) - \bar{U}(\phi_j)}{r}\right)^2,$$

where $\bar{U}(\phi_i)$ is the ith component of vector $\bar{U}$ and σ_h is the surface tension when

$$h(x) = (1 - x^2)_+^{d_v/2} \qquad \text{where} \quad (y)_+ = \max(y, 0).$$

We can define within and cross manifold Dirichlet energy

$$b_{\phi,k}(\bar{U}) = \frac{1}{n_k^2}\sum_{\phi_i,\phi_j\in\mathcal{N}_{s,k}}\frac{1}{\sigma_h r^d}\bar{W}_{i,j}\left(\frac{\bar{U}(\phi_i) - \bar{U}(\phi_j)}{r}\right)^2, \qquad b_{\phi,W}(\bar{U}) = \sum_{k=1}^K\left(\frac{n_k}{n}\right)^2 b_{\phi,k}(\bar{U}),$$

and

$$b_{\phi,C}(\bar{U}) = \frac{1}{n^2} \sum_{k_1 \neq k_2} \sum_{\phi_i \in \mathcal{N}_{s,k_1}, \phi_j \in \mathcal{N}_{s,k_2}} \frac{1}{\sigma_h r^d} \bar{W}_{i,j} \left(\frac{\bar{U}(\phi_i) - \bar{U}(\phi_j)}{r} \right)^2.$$

We also need to define an intermediate discrete Dirichlet energy

$$\tilde{b}_{\phi}(\bar{U}) = \frac{1}{n^2} \sum_{i,j} \frac{1}{\sigma_h r^{d_s}} h \left(\frac{d_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)}{r} \right) \left(\frac{\bar{U}(\phi_i) - \bar{U}(\phi_j)}{r} \right)^2,$$

and corresponding within manifold Dirichlet energy

$$\tilde{b}_{\phi,W}(\bar{U}) = \sum_{k=1}^{K} \left(\frac{n_k}{n} \right)^2 \frac{1}{\mathrm{Vol}\mathcal{N}_{v,k}} b_{\phi,k}(\bar{U}),$$

where

$$\tilde{b}_{\phi,k}(\bar{U}) = \frac{1}{n_k^2} \sum_{\phi_i, \phi_j \in \mathcal{N}_{s,k}} \frac{1}{\sigma_h r^{d_s}} h \left(\frac{d_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)}{r} \right) \left(\frac{\bar{U}(\phi_i) - \bar{U}(\phi_j)}{r} \right)^2.$$

Step 3: Discretization and interpolation maps By applying Proposition 8, we can know that with probability at least $1 - Kn \exp(-Cn\gamma^2\delta^{d_s}) - 2K\exp(-cn)$, there exists a probability measure $\tilde{\mu}^s_{n,k}$ with probability density function $\tilde{\pi}^s_{n,k}$ for $k = 1, \ldots, K$ such that

$$\|\pi^s_k - \tilde{\pi}^s_{n,k}\|_{L^\infty(\mathcal{N}_{s,k})} \leq C(\gamma + \delta)$$

and there exist transportation maps $\tilde{R}_1, \ldots, \tilde{R}_K$ such that

$$\sup_{x \in \mathcal{N}_{s,k}} d_{\mathcal{N}_{s,k}}(x, \tilde{R}_k(x)) \leq \delta.$$

Similarly, we can choose $\delta = \sqrt{r(c \log n/n)^{1/d_s}}$ and $\gamma = \sqrt{(\alpha+1)\log n/n\delta^{d_s}}$ for some large enough α. Based on the transportation maps $\tilde{R}_1, \ldots, \tilde{R}_K$, we can introduce the discretization map $\tilde{P}_\phi : L^2(\pi^s) \to \mathbb{R}^n$ and the interpolation map $\tilde{\mathcal{I}}_\phi : \mathbb{R}^n \to L^2(\pi^s)$ in the same way as the proof of Theorem 1.

Step 4: Connection between Dirichlet energy $D_\phi(\bar{\theta})$ and $b_\phi(\bar{U})$ To build the connection between $D_\phi(\bar{\theta})$ and $b_\phi(\bar{U})$, we need to find more explicit upper and lower bound for $\bar{W}_{i,j}$.

If $\phi_i, \phi_j \in \mathcal{N}_{s,k}$ and $d_{\mathcal{N}_{s,k}}(\phi_i, \phi_j) \leq r$, then

$$\begin{aligned}
\frac{1}{r^d}\bar{W}_{i,j} &= \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{M}(\phi_i)} \int_{\mathcal{M}(\phi_j)} \mathbf{I}(\|x-y\| \leq r) dx dy \\
&\geq \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{M}(\phi_i)} \int_{\mathcal{M}(\phi_j)} \mathbf{I}\left(d_{\mathcal{M}_k}(x,y) \leq r\right) dx dy \\
&\geq \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \mathbf{I}\left(d_{\mathcal{N}_{v,k}}(\psi_i, \psi_j) \leq \sqrt{r^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)}\right) d\psi_i d\psi_j \\
&\geq \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \left(1 - C(r^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j))\right) V_{d_v} \left(r^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)\right)^{d_v/2} d\psi_1 \\
&\geq \frac{V_{d_v}}{r^d \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 - C(r^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j))\right) \left(r^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)\right)^{d_v/2} \\
&\geq \frac{V_{d_v}}{r^{d_s} \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 - Cr^2\right) \left(1 - \frac{d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)}{r^2}\right)^{d_v/2}.
\end{aligned}$$

Here we use the following facts

1. $\|x - y\| \leq d_{\mathcal{M}_k}(x, y)$;
2. $d^2_{\mathcal{M}_k}(x, y) = d^2_{\mathcal{N}_{s,k}}(\phi_i, \phi_j) + d^2_{\mathcal{N}_{v,k}}(\psi_i, \psi_j)$ when $x = T_k(\phi_i, \psi_i)$ and $y = T_k(\phi_j, \psi_j)$;
3. $|\mathrm{Vol}(B_{\mathcal{N}_{v,k}}(\psi, r)) - V_{d_v} r^{d_v}| \leq C r^{d_v+2}$, where $B_{\mathcal{N}_{v,k}}(\psi, r)$ is a geodesic ball with center at ψ and radius r;
4. $\mathcal{N}_{v,k}$ has no boundary.

This suggests

$$\frac{1}{r^d}\bar{W}_{i,j} \geq \frac{V_{d_v}}{r^{d_s} \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 - Cr^2\right) h\left(\frac{d_{\mathcal{N}_{s,k}}(\phi_i, \phi_j)}{r}\right)$$

and

$$b_{\phi,k}(\bar{U}) \geq \left(1 - Cr^2\right) \frac{V_{d_v}}{\mathrm{Vol} \mathcal{N}_{v,k}} \tilde{b}_{\phi,k}(\bar{U}).$$

If we apply Proposition 9, we have

$$\begin{aligned}
b_{\phi,W}(\bar{U}) &\geq \left(1 - Cr^2\right) V_{d_v} \tilde{b}_{\phi,W}(\bar{U}) \\
&\geq \left(1 - Cr^2\right) V_{d_v} D_\phi(\tilde{\mathcal{I}}_\phi \bar{U}) / \left(1 + C\left(\frac{\delta}{r} + r + \gamma\right)\right).
\end{aligned}$$

Because $b_{\phi,C}(\bar{U}) \geq 0$,

$$\begin{aligned}
D_\phi(\tilde{\mathcal{I}}_\phi \bar{U}) &\leq \frac{1}{V_{d_v}} \left(1 + C\left(\frac{\delta}{r} + r + \gamma + r^2\right)\right) b_{\phi,W}(\bar{U}) \\
&\leq \frac{1}{V_{d_v}} \left(1 + C\left(\frac{\delta}{r} + r + \gamma + r^2\right)\right) b_\phi(\bar{U}).
\end{aligned}$$

Next, we work on the upper bound of $\bar{W}_{i,j}$. If $\phi_i, \phi_j \in \mathcal{N}_{s,k}$, then

$$\begin{aligned}
\frac{1}{r^d}\bar{W}_{i,j} &= \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{M}(\phi_i)} \int_{\mathcal{M}(\phi_j)} \mathbf{I}(\|x-y\| \le r) dx dy \\
&\le \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{M}(\phi_i)} \int_{\mathcal{M}(\phi_j)} \mathbf{I}\left(d_{\mathcal{M}}(x,y) \le \tilde{r}\right) dx dy \\
&\le \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \mathbf{I}\left(d_{\mathcal{N}_{v,k}}(\psi_i,\psi_j) \le \sqrt{\tilde{r}^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i,\phi_j)}\right) d\psi_i d\psi_j \\
&\le \frac{1}{r^d \mathrm{Vol}^2 \mathcal{N}_{v,k}} \int_{\mathcal{N}_{v,k}} \left(1 + c\tilde{r}^2\right) V_{d_v} (\tilde{r}^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i,\phi_j))^{d_v/2} d\psi_i \\
&\le \frac{1}{r^d \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 + c\tilde{r}^2\right) V_{d_v} \left(\tilde{r}^2 - d^2_{\mathcal{N}_{s,k}}(\phi_i,\phi_j)\right)^{d_v/2} \\
&\le \frac{(\tilde{r}/r)^{d_v}}{r^{d_s} \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 + c\tilde{r}^2\right) V_{d_v} \left(1 - \frac{d^2_{\mathcal{N}_{s,k}}(\phi_i,\phi_j)}{\tilde{r}^2}\right)^{d_v/2}
\end{aligned}$$

Here $\tilde{r} = r + 8r^3/R^2$ and we use the following facts

1. $d_{\mathcal{M}_k}(x,y) \le \|x-y\| + 8\|x-y\|^3/R^2$, where R is the reach of the manifold;
2. $d^2_{\mathcal{M}_k}(x,y) = d^2_{\mathcal{N}_{s,k}}(\phi_i,\phi_j) + d^2_{\mathcal{N}_{v,k}}(\psi_i,\psi_j)$ when $x = T_k(\phi_i,\psi_i)$ and $y = T_k(\phi_j,\psi_j)$;
3. $|\mathrm{Vol}(B_{\mathcal{N}_{v,k}}(\psi, r)) - V_{d_v} r^{d_v}| \le C r^{d_v+2}$, where $B_{\mathcal{N}_{v,k}}(\psi, r)$ is a geodesic ball with center at ψ and radius r;
4. $\mathcal{N}_{v,k}$ has no boundary.

So we can have

$$\frac{1}{r^d}\bar{W}_{i,j} \le \left(\frac{\tilde{r}}{r}\right)^d \frac{V_{d_v}}{\tilde{r}^{d_s} \mathrm{Vol} \mathcal{N}_{v,k}} \left(1 + C\tilde{r}^2\right) h\left(\frac{d_{\mathcal{N}_{s,k}}(\phi_i,\phi_j)}{\tilde{r}}\right)$$

and

$$b_{\phi,k}(\bar{U}) \le \left(\frac{\tilde{r}}{r}\right)^d \left(1 + Cr^2\right) \frac{V_{d_v}}{\mathrm{Vol} \mathcal{N}_{v,k}} \tilde{b}_{\phi,k}(\bar{U}).$$

Note that the size of neighborhood in $\tilde{b}_{\phi,k}(\bar{U})$ of the above inequality is $\tilde{r}$. An application of Proposition 9 leads to

$$\begin{aligned}
b_{\phi,W}(\tilde{P}_\phi \bar{\theta}) &\le \left(1 + Cr^2\right) V_{d_v} \tilde{b}_{\phi,W}(\tilde{P}_\phi \bar{\theta}) \\
&\le \left(1 + Cr^2\right) \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma\right)\right) V_{d_v} D_\phi(\bar{\theta}) \\
&\le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma + r^2\right)\right) V_{d_v} D_\phi(\bar{\theta}).
\end{aligned}$$

When $\delta(\mathcal{M}) > r$, we can know $b_{\phi,C}(\bar{U}) = 0$. This immediately suggests

$$b_\phi(\tilde{P}_\phi \bar{\theta}) \le \left(1 + C\left(\frac{\delta}{\tilde{r}} + \tilde{r} + \gamma + r^2\right)\right) V_{d_v} D_\phi(\bar{\theta}).$$

After building the connection between $D_\phi(\bar{\theta})$ and $b_\phi(\bar{U})$, we can apply the same arguments in step 5-8 of proof for Theorem 1 to complete the proof.

B.5 Proof for Theorem 3

The proof is divided into three steps.

Step 1: Hypothesis construction In this step, we construct the least favorable hypothesis $\mathbb{H}_0$ and $\mathbb{H}_1$. Under $\mathbb{H}_0$, we assume the observed data is drawn from a single manifold

$$\mathcal{M} = \mathcal{M}_0 = \{(x, 0_{D-d}) : x \in [0,1]^d\},$$

where 0_{D-d} is a $D-d$ dimensional vector of zeros. We also assume that the probability distribution of our observed data is uniform distribution on $\mathcal{M}_0$. Given the manifold, we write $\mathbb{P}_0$ as the probability distribution of observed data $X_1, \ldots, X_n$ under $\mathbb{H}_0$. Next, we construct the other hypothesis $\mathbb{H}_1$. We divide the space $[0,1]^d$ into $L = M^d$ cubes of size $M^{-1} \times \ldots \times M^{-1}$ for some positive integer M and name these disjoint cubes $Q_1, \ldots, Q_L$. M will be specified in the third step. For each Q_l, we also define $\tilde{Q}_l$ as a cube with the same center of Q_l but a smaller size $(3M)^{-1} \times \ldots \times (3M)^{-1}$. Given Q_l, we consider the following two manifolds

$$\mathcal{M}_1(Q_l) = \{(x, 0_{D-d}) : x \in [0,1]^d \setminus Q_l\} \qquad \text{and} \qquad \mathcal{M}_2(Q_l) = \{(x, 0_{D-d}) : x \in \tilde{Q}_l\}.$$

Clearly, $\mathcal{M}_1(Q_l)$ and $\mathcal{M}_2(Q_l)$ are two disjoint manifolds and the distance between them is $(3M)^{-1}$. So $\delta(\mathcal{M}(Q_l)) = (3M)^{-1}$ if we define $\mathcal{M}(Q_l) = \mathcal{M}_1(Q_l) \cup \mathcal{M}_2(Q_l)$. Given these two manifolds, we consider a uniform distribution over $\mathcal{M}(Q_l)$ and define it as hypothesis $\mathbb{H}_1(Q_l)$. Write $\mathbb{P}_1(Q_l)$ as the probability distribution of observed data $X_1, \ldots, X_n$ under $\mathbb{H}_1(Q_l)$. The hypothesis $\mathbb{H}_1$ is a mixture of hypotheses $\mathbb{H}_1(Q_l)$ for $l = 1, \ldots, L$ and we write

$$\mathbb{P}_1 = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}_1(Q_l).$$

By the construction, the observed data is drawn from single manifold under $\mathbb{H}_0$ and two disjoint manifolds under $\mathbb{H}_1$.

Step 2: Bounding χ^2 divergence The goal of this step is to bound χ^2 divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$. To the end, we write the likelihood between $\mathbb{P}_1(Q_l)$ and $\mathbb{P}_0$ as

$$F_l(X_1, \ldots X_n) = \frac{d\mathbb{P}_1(Q_l)}{d\mathbb{P}_0} = \begin{cases} \left(\frac{1}{1-\Delta}\right)^n, & \forall X_i \in ([0,1]^d \setminus Q_l) \cup \tilde{Q}_l \\ 0, & \exists X_i \in Q_l \setminus \tilde{Q}_l \end{cases},$$

where $\Delta = M^{-d} - (3M)^{-d}$ is the volume of set $Q_l \setminus \tilde{Q}_l$. Given the likelihood $F_1, \ldots, F_L$, we can write the χ^2 divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$ as

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) = \mathbb{E}_0 \left(\frac{d\mathbb{P}_1}{d\mathbb{P}_0} - 1\right)^2 = \mathrm{Var}_0 \left(\frac{1}{L} \sum_{l=1}^{L} F_l\right),$$

where $\mathbb{E}_0$ and Var_0 are the expectation and variance under the probability distribution $\mathbb{P}_0$. Therefore, it is sufficient to bound the variance and covariance of $F_1, \ldots, F_L$. For any $l_1 \neq l_2$, we have

$$F_{l_1} F_{l_2} = \begin{cases} \left(\frac{1}{1-\Delta}\right)^{2n}, & \forall X_i \in \left([0,1]^d \setminus (Q_{l_1} \cup Q_{l_2})\right) \cup \tilde{Q}_{l_1} \cup \tilde{Q}_{l_2} \\ 0, & \exists X_i \in (Q_{l_1} \setminus \tilde{Q}_{l_1}) \cup (Q_{l_2} \setminus \tilde{Q}_{l_2}) \end{cases}.$$

Because $\mathbb{E}_0(F_l) = 1$, we can bound the covariance of F_{l_1} and F_{l_2} in the following way

$$\begin{aligned}
|\mathrm{Cov}_0(F_{l_1}, F_{l_2})| &= |\mathbb{E}_0(F_{l_1}F_{l_2}) - 1| \\
&= \left| \left(\frac{1}{1-\Delta}\right)^{2n} (1-2\Delta)^n - 1 \right| \\
&= \left| \left(1 - \frac{\Delta^2}{(1-\Delta)^2}\right)^n - 1 \right| \\
&\le 1 - \exp\left(-n\Delta^2/(1-\Delta)\sqrt{1-2\Delta}\right) \\
&\le \frac{n\Delta^2}{(1-\Delta)\sqrt{1-2\Delta}}.
\end{aligned}$$

Here we the fact that $\log(1-x) \ge -x/\sqrt{1-x}$ when $0 < x < 1$ and $1 - x \le e^{-x}$. For the variance term, we have

$$\begin{aligned}
\mathrm{Var}_0(F_l) \le \mathbb{E}_0(F_l^2) &= \left(\frac{1}{1-\Delta}\right)^n \\
&= \exp\left(n \log\left(1 + \frac{\Delta}{1-\Delta}\right)\right) \\
&\le \exp\left(\frac{n\Delta}{1-\Delta}\right),
\end{aligned}$$

where we use the fact $\log(1+x) \le x$. Putting the bound of covariance and variance together yields

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \le \frac{1}{L} \exp\left(\frac{n\Delta}{1-\Delta}\right) + \frac{n\Delta^2}{(1-\Delta)\sqrt{1-2\Delta}}.$$

Step 3: Wrap up the proof Let T be any test. We still need to specify the choice of M in the hypotheses constructed in step 1. Specifically, we can choose $M = \lceil (2n/\log n)^{1/d} \rceil$, so

$$L \ge 2n/\log n \qquad \text{and} \qquad \Delta \le \frac{\log n}{2n}.$$

Therefore, we have

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \le \frac{\log n}{n^{1/3}} + \frac{\log^2 n}{n} \to 0.$$

Therefore, we have

$$\mathbb{P}_0(T = 1) + \mathbb{P}_1(T = 0) \ge 1 - \chi^2(\mathbb{P}_1, \mathbb{P}_0) \to 1.$$

The construction suggests that under the hypothesis $\mathbb{H}_1$, we have

$$\delta(\mathcal{M}) \ge \frac{1}{6}\left(\frac{\log n}{n}\right)^{1/d}.$$

Therefore, we can conclude that when $c = 1/6$,

$$\mathbb{P}_0(T = 1) + \mathbb{P}_1(T = 0) \to 1.$$

B.6 Proof for Theorem 5

We follow a similar strategy and the same notations in proof for Theorem 3, but need a different way to construct hypotheses.

Step 1: Hypothesis construction Same to the proof for Theorem 3, we assume the observed data is drawn from a uniform distribution on the single manifold $\mathcal{M}_0$ under $\mathbb{H}_0$. The hypothesis $\mathbb{H}_1$ is constructed in a different way from the proof for Theorem 3. Specifically, we divide the space $[0,1]^{d_s}$ into $L = M^{d_s}$ cubes of size $M^{-1} \times \ldots \times M^{-1}$ for some positive integer M, which will be specified later. We still name these disjoint cubes $Q_1, \ldots, Q_L$ and define $\tilde{Q}_l$ as a cube with the same center of Q_l but a smaller size $(3M)^{-1} \times \ldots \times (3M)^{-1}$. Under $\mathbb{H}_1$, we consider the following manifolds

$$\mathcal{M}_1(Q_l) = \{(x, y, 0_{D-d}) : x \in [0,1]^{d_s} \setminus Q_l, y \in [0,1]^{d_v}\}$$

and

$$\mathcal{M}_2(Q_l) = \{(x, y, 0_{D-d}) : x \in \tilde{Q}_l, y \in [0,1]^{d_v}\}.$$

If we define $\mathcal{M}(Q_l) = \mathcal{M}_1(Q_l) \cup \mathcal{M}_2(Q_l)$, we have $\delta(\mathcal{M}(Q_l)) = (3M)^{-1}$. Given each pair of Q_l and $\tilde{Q}_l$, we define a uniform distribution over $\mathcal{M}(Q_l)$ as hypothesis $\mathbb{H}_1(Q_l)$ and write $\mathbb{P}_1(Q_l)$ as the corresponding probability distribution of observed data $X_1, \ldots, X_n$. $\mathbb{P}_1$ is defined as a mixture distribution of $\mathbb{P}_1(Q_l)$

$$\mathbb{P}_1 = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}_1(Q_l).$$

Step 2: Bounding χ^2 divergence We can bound the χ^2 divergence in the same way as the proof for Theorem 3. Specifically, we can show

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \leq \frac{1}{L} \exp\left(\frac{n\Delta}{1-\Delta}\right) + \frac{n\Delta^2}{(1-\Delta)\sqrt{1-2\Delta}},$$

where $\Delta = M^{-d_s} - (3M)^{-d_s}$.

Step 3: Applying theorem of fuzzy hypothesis We can choose $M = \lceil (2n/\log n)^{1/d_s} \rceil$ and still show

$$\mathbb{P}_0(T = 1) + \mathbb{P}_1(T = 0) \to 1.$$

B.7 Proof for Theorem 6

Without loss of generality, we can assume $Y_i = 1$ when $\tilde{X}_i \in \mathcal{M}_s$ for $1 \leq s \leq K'$ and $Y_i = -1$ if $\tilde{X}_i \in \mathcal{M}_s$ for $K' < s \leq K$. Furthermore, we assume the first K eigenfunctions have the following forms

$$\theta_s(x) = \begin{cases} 1, & x \in \mathcal{M}_s \\ 0, & x \in \mathcal{M} \setminus \mathcal{M}_s \end{cases}.$$

Step 1 In this step, we aim to show that the learned representation $\hat{\Theta}(\tilde{X}_1), \ldots, \hat{\Theta}(\tilde{X}_m)$ are linearly separable with respect to Y in a large probability. An application of Markov's inequality suggests

$$\mathbb{P}\left(|\hat{\theta}_s(\tilde{X}_i) - \theta_s(\tilde{X}_i)| > \frac{1}{3K}\right) \leq \frac{\mathbb{E}(\hat{\theta}_s(\tilde{X}_i) - \theta_s(\tilde{X}_i))^2}{(1/3K)^2} \leq 9K^2\chi_n.$$

Therefore, with probability $1 - 9mK^3\chi_n$, we have

$$|\hat{\theta}_s(\tilde{X}_i) - \theta_s(\tilde{X}_i)| \leq \frac{1}{3K}, \qquad 1 \leq s \leq K,\ 1 \leq i \leq m. \tag{14}$$

Let $\beta^o = (\beta_s^o)$ such that $\beta_s^o = 1$ when $1 \leq s \leq K'$ and $\beta_s^o = -1$ when $K' < s \leq K$. When $Y_i = 1$, then

$$\beta^{oT}\hat{\Theta}(\tilde{X}_i) \geq 1 - \frac{1}{3K} - (K'-1)\frac{1}{3K} - (K-K')\frac{1}{3K} \geq \frac{2}{3} > 0.$$

On the other hand, if $Y_i = -1$, then

$$\beta^{oT}\hat{\Theta}(\tilde{X}_i) \leq K'\frac{1}{3K} + (K-K'-1)\frac{1}{3K} - \left(1 - \frac{1}{3K}\right) \leq -\frac{2}{3} < 0.$$

Therefore, $\{\hat{\Theta}(\tilde{X}_i) : Y_i = 1\}$ and $\{\hat{\Theta}(\tilde{X}_i) : Y_i = -1\}$ are linearly separable.

Step 2 The goal of this step is to identify a set $\mathcal{B}$ which covers all converged weight vector β in the logistic regression. According to Soudry et al. (2018); Ji and Telgarsky (2018), the weight vector β of logistic regression converges to the direction of the max-margin solution. Specifically, Theorem 1.1 in Ji and Telgarsky (2018) suggests that

$$\lim_{t\to\infty} \frac{\beta_t}{\|\beta_t\|} = \frac{\beta^*}{\|\beta^*\|},$$

where β^* is the optimal solution of the following optimization problem

$$\min_{\beta} \|\beta\|^2, \qquad \text{s.t.} \quad Y_i\beta^T\hat{\Theta}(\tilde{X}_i) \geq 1, \quad 1 \leq i \leq n. \tag{15}$$

Now we study the property of β^*. Clearly, $3\beta^o/2$ satisfy the constraint in (15) and therefore we have

$$\|\beta^*\|^2 \leq \|3\beta^o/2\|^2 = 9K/4.$$

When $\tilde{X}_i \in \mathcal{M}_s$ for some $1 \leq s \leq K'$, (14) implies

$$\beta^T\hat{\Theta}(\tilde{X}_i) \leq \beta_s + \frac{1}{3K}\sum_{s=1}^{K}|\beta_s| \leq \beta_s + \frac{1}{3K}\|\beta\|\sqrt{K}.$$

An application of $\|\beta^*\|^2 \leq 9K/4$ yields

$$1 \leq \beta^{*T}\hat{\Theta}(\tilde{X}_i) \leq \beta_s^* + \frac{1}{3K}\|\beta^*\|\sqrt{K} \leq \beta_s^* + \frac{1}{2}.$$

Therefore, we can show that $\beta_s^* \geq 1/2$ when $1 \leq s \leq K'$. Through the same way, we can show that $\beta_s^* \leq -1/2$ when $K' < s \leq K$. Overall, we can conclude

$$\beta^* \in \mathcal{B} := \left\{\beta \in \mathbb{R}^K : \|\beta\|^2 \leq \frac{9K}{4}, \quad \beta_s \geq \frac{1}{2} \text{ if } 1 \leq s \leq K' \quad \text{and} \quad \beta_s \leq -\frac{1}{2} \text{ if } K' < s \leq K\right\}.$$

Step 3 Because $\beta^* \in \mathcal{B}$, it is sufficient to study the misclassification rate of the following classifier

$$\hat{H}_{\hat{\Theta},\beta}(x) = \begin{cases} 1, & \beta^T \hat{\Theta}(x) > 0 \\ -1, & \beta^T \hat{\Theta}(x) \leq 0 \end{cases}$$

for any $\beta \in \mathcal{B}$. Recall the true label is

$$H^*(x) = \begin{cases} 1, & x \in \mathcal{M}_s \text{ when } 1 \leq s \leq K' \\ -1, & x \in \mathcal{M}_s \text{ when } K' < s \leq K \end{cases}.$$

When $\beta \in \mathcal{B}$, we can have

$$\mathbb{P}(\hat{H}_{\hat{\Theta},\beta}(X) \neq H^*(X)) \leq \mathbb{P}\left(|\hat{\theta}_s(X) - \theta_s(X)| > \frac{1}{6K},\ 1 \leq s \leq K\right)$$

because when $|\hat{\theta}_s(X) - \theta_s(X)| \leq 1/6K$,

$$\beta^T \hat{\Theta}(X) \geq \beta_s - \frac{1}{6K}\sum_s |\beta_s| \geq \frac{1}{2} - \frac{1}{6K}\frac{3K}{2} \geq \frac{1}{4} \qquad \text{if } X \in \mathcal{M}_s \text{ for some } 1 \leq s \leq K'$$

and

$$\beta^T \hat{\Theta}(X) \leq \beta_s + \frac{1}{6K}\sum_s |\beta_s| \leq -\frac{1}{2} + \frac{1}{6K}\frac{3K}{2} \leq -\frac{1}{4} \qquad \text{if } X \in \mathcal{M}_s \text{ for some } K' < s \leq K.$$

By union bound and Markov's inequality, we have

$$\begin{aligned} \mathbb{P}\left(|\hat{\theta}_s(X) - \theta_s(X)| > \frac{1}{6K},\ 1 \leq s \leq K\right) &\leq \sum_{s=1}^{K} \mathbb{P}\left(|\hat{\theta}_s(X) - \theta_s(X)| > \frac{1}{6K}\right) \\ &\leq K\frac{\mathbb{E}(\hat{\theta}_s(X) - \theta_s(X))^2}{(1/6K)^2} \\ &\leq 36K^3\chi_n. \end{aligned}$$

Therefore, we can conclude that with probability $1 - 9mK^3\chi_n$,

$$\xi(\hat{\Theta}) \leq 36K^3\chi_n.$$

B.8 Proof for Proposition 10

Let $k(X_i)$ be the manifold that X_i belongs to, i.e., $k(X_i) = k$ if $X_i \in \mathcal{M}_k$. The following analysis is conducted by conditioning on $k(X_1), \ldots, k(X_n)$. We write the number of edges connecting points in $\mathcal{M}_{k_1}$ and $\mathcal{M}_{k_2}$ as

$$Z_{k_1,k_2} = \sum_{X_i \in \mathcal{M}_{k_1}, X_j \in \mathcal{M}_{k_2}} \mathbf{I}(\|X_i - X_j\| \leq r).$$

The expectation of Z_{k_1,k_2} can be written as

$$\begin{aligned}\mathbb{E}(Z_{k_1,k_2}) &= \mathbb{E}\left(\sum_{X_i\in\mathcal{M}_{k_1},X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)\middle|k(X_1),\ldots,k(X_n)\right)\\ &\le n_{k_1}\sup_{x\in\mathcal{M}_{k_1}}\mathbb{E}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|x-X_j\|\le r)\middle|k(X_1),\ldots,k(X_n)\right)\\ &\le \beta n_{k_1}n_{k_2}\end{aligned}$$

To bound the variance of Z_{k_1,k_2}, we apply Efron-Stein inequality (See Theorem 3.1 in Boucheron et al., 2013)

$$\begin{aligned}\mathrm{Var}(Z_{k_1,k_2}) \le& \frac{1}{2}\sum_{i=1}^{k_1}\mathbb{E}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)-\mathbf{I}(\|X_i'-X_j\|\le r)\right)^2\\ &+\frac{1}{2}\sum_{j=1}^{k_2}\mathbb{E}\left(\sum_{X_i\in\mathcal{M}_{k_1}}\mathbf{I}(\|X_i-X_j\|\le r)-\mathbf{I}(\|X_i-X_j'\|\le r)\right)^2,\end{aligned}$$

where X_i' and X_j' are independent copies of X_i and X_j. The first term of the right hand side in above inequality can be written as

$$\begin{aligned}&\mathbb{E}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)-\mathbf{I}(\|X_i'-X_j\|\le r)\right)^2\\ \le&4\mathrm{Var}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)\right)\\ \le&4\sup_{x\in\mathcal{M}_{k_1}}\mathrm{Var}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|x-X_j\|\le r)\right)+4\mathbb{E}\left(\mathbb{E}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)\middle|X_i\right)\right)^2\\ \le&4\sup_{x\in\mathcal{M}_{k_1}}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathrm{Var}\left(\mathbf{I}(\|x-X_j\|\le r)\right)\right)+4\beta^2n_{k_2}^2\\ \le&4\beta n_{k_2}+4\beta^2n_{k_2}^2.\end{aligned}$$

Here we use the fact

$$0\le\mathbb{E}\left(\sum_{X_j\in\mathcal{M}_{k_2}}\mathbf{I}(\|X_i-X_j\|\le r)\middle|X_i\right)\le\beta n_{k_2}.$$

We can apply the same strategy to bound the other term of the right hand side to obtain

$$\mathbb{E}\left(\sum_{X_i\in\mathcal{M}_{k_1}}\mathbf{I}(\|X_i-X_j\|\le r)-\mathbf{I}(\|X_i-X_j'\|\le r)\right)^2\le 4\beta n_{k_1}+4\beta^2n_{k_1}^2.$$

Putting two terms together yields

$$\mathrm{Var}(Z_{k_1,k_2}) \leq 4\beta n_{k_1} n_{k_2} + 2\beta^2 n_{k_1} n_{k_2}(n_{k_1} + n_{k_2}).$$

By Chebyshev's inequality, we can conclude that

$$\mathbb{P}\left(Z_{k_1,k_2} \geq 2\left(\beta n_{k_1} n_{k_2} + t\sqrt{\beta n_{k_1} n_{k_2}} + t\beta\sqrt{n_{k_1} n_{k_2}(n_{k_1} + n_{k_2})}\right)\right) \leq \frac{1}{t^2}.$$

By union bound and the fact

$$\mathbb{P}(nw_k/2 \leq n_k \leq 2nw_k, \forall 1 \leq k \leq K) \geq 1 - 2K\exp(-cn),$$

we have

$$\mathbb{P}\left(\max_{1\leq k_1,k_2\leq K} Z_{k_1,k_2} \geq c\beta n^2 + t\left(\sqrt{\beta}n + \beta n^{3/2}\right)\right) \leq 2K\exp(-cn) + \frac{K^2}{t^2}$$

If we apply Proposition A.4 in García Trillos et al. (2021), we can conclude that

$$b_C(\tilde{P}\theta) \leq C\left(\frac{\beta n + t\sqrt{\beta} + t\beta\sqrt{n}}{nr^{d+2}}\right)\left(1 + \lambda^{d/2+2}\right)\|\theta\|^2_{L^2(\pi)}.$$

Appendix C. Notations

In this section, we formally define the probability density function on the manifold. The volume of a Riemannian manifold $(\mathcal{M}, g)$ is defined in the following way:

$$\mathrm{Vol}\mathcal{M} = \int_{\mathcal{M}} \sqrt{\det(g_{ij})}dx_1 \wedge \ldots \wedge dx_d,$$

where $(x_1, \ldots, x_d)$ is a set of local coordinates. If X is a random variable defined on the manifold $(\mathcal{M}, g)$, the probability density function π of X is a function $\mathcal{M} \to \mathbb{R}$ that satisfies

$$\mathbb{P}(X \in \mathcal{A}) = \int_{\mathcal{A}} \pi(x)\sqrt{\det(g_{ij})}dx_1 \wedge \ldots \wedge dx_d,$$

for any measurable set $\mathcal{A} \subset \mathcal{M}$. See also Hendriks (1990).

References

H. Akbari, L. Yuan, R. Qian, W. Chuang, S. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

M. A. Arcones. A bernstein-type inequality for u-statistics and u-processes. *Statistics & probability letters*, 22(3):239–247, 1995.

E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.

S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

R. Balestriero and Y. LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022.

R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, A. Schwarzschild, A.G. Wilson, J. Geiping, Q. Garrido, P. Fernandez, A. Bar, H. Pirsiavash, Y. LeCun, and M. Goldblum. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the laplace–beltrami operator. *Journal of Spectral Theory*, 4(4):675–714, 2015.

V. Cabannes, B. T. Kiani, R. Balestriero, Y. LeCun, and A. Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. *arXiv preprint arXiv:2302.02774*, 2023.

J. Calder and N. García Trillos. Improved spectral convergence rates for graph laplacians on ε-graphs and k-nn graphs. *Applied and Computational Harmonic Analysis*, 60:123–175, 2022.

S. Chen, E. Dobriban, and J. H. Lee. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1):9885–9955, 2020a.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020b.

X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

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

X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Z. Fu, W. Hu, and T. Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–602. Ieee, 2005.

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(4):827–887, 2020.

N. García Trillos, P. He, and C. Li. Large sample spectral analysis of graph-based multi-manifold clustering. *arXiv preprint arXiv:2107.13610*, 2021.

J. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

H. Hendriks. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *The Annals of Statistics*, pages 832–849, 1990.

G. E. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, 15, 2002.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

D. D. Johnson, A. E. Hanchi, and C. J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions. *arXiv preprint arXiv:2210.01883*, 2022.

A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Y. LeCun. A path towards autonomous machine intelligence. *Open Review*, 62, 2022.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

R. R. Lederman and R. Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

V. J. Martinez and E. Saar. *Statistics of the galaxy distribution*. Chapman and Hall/CRC, 2001.

E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

M. Penrose. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *OpenAI*, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020a.

Y. Tian, L. Yu, X. Chen, and S. Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020b.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

Y. Tsai, Y. Wu, R. Salakhutdinov, and L. Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation and estimation. *Journal of Mathematical Imaging and Vision*, 25(3):403–421, 2006.

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

S. Wang. Self-supervised metric learning in multi-view data: A downstream task perspective. *Journal of the American Statistical Association*, 118(544):2454–2467, 2023.

S. Wang. Augmentation invariant manifold learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf003, 2025.

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

C. Wei, K. Shen, Y. Chen, and T. Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Z. Wen and Y. Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

Z. Wen and Y. Li. The mechanism of prediction head in non-contrastive self-supervised learning. *arXiv preprint arXiv:2205.06226*, 2022.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.