VECTOR DIFFUSION MAPS AND THE CONNECTION LAPLACIAN

A. SINGER[*] AND H.-T. WU[†]

Dedicated to the Memory of Partha Niyogi

Abstract. We introduce *vector diffusion maps* (VDM), a new mathematical framework for organizing and analyzing massive high dimensional data sets, images and shapes. VDM is a mathematical and algorithmic generalization of diffusion maps and other non-linear dimensionality reduction methods, such as LLE, ISOMAP and Laplacian eigenmaps. While existing methods are either directly or indirectly related to the heat kernel for functions over the data, VDM is based on the heat kernel for vector fields. VDM provides tools for organizing complex data sets, embedding them in a low dimensional space, and interpolating and regressing vector fields over the data. In particular, it equips the data with a metric, which we refer to as the *vector diffusion distance.* In the manifold learning setup, where the data set is distributed on (or near) a low dimensional manifold $\mathcal{M}^d$ embedded in $\mathbb{R}^p$, we prove the relation between VDM and the connection-Laplacian operator for vector fields over the manifold.

Key words. Dimensionality reduction, vector fields, heat kernel, parallel transport, local principal component analysis, alignment.

1. Introduction. A popular way to describe the affinities between data points is using a weighted graph, whose vertices correspond to the data points, edges that connect data points with large enough affinities and weights that quantify the affinities. In the past decade we have been witnessed to the emergence of non-linear dimensionality reduction methods, such as locally linear embedding (LLE) [33], ISOMAP [39], Hessian LLE [12], Laplacian eigenmaps [2] and diffusion maps [9]. These methods use the local affinities in the weighted graph to learn its global features. They provide invaluable tools for organizing complex networks and data sets, embedding them in a low dimensional space, and studying and regressing functions over graphs. Inspired by recent developments in the mathematical theory of cryo-electron microscopy [36, 18] and synchronization [34, 10], in this paper we demonstrate that in many applications, the representation of the data set can be vastly improved by attaching to every edge of the graph not only a weight but also a linear orthogonal transformation (see Figure 1.1).

Consider, for example, a data set of images, or small patches extracted from images (see, e.g., [27, 8]). While weights are usually derived from the pairwise comparison of the images in their original representation, we instead associate the weight w_{ij} to the similarity between image i and image j when they are optimally rotationally aligned. The dissimilarity between images when they are optimally rotationally aligned is sometimes called the rotationally invariant distance [31]. We further define the linear transformation O_{ij} as the 2×2 orthogonal transformation that registers the two images (see Figure 1.2). Similarly, for data sets consisting of three-dimensional shapes, O_{ij} encodes the optimal 3×3 orthogonal registration transformation. In the case of manifold learning, the linear transformations can be constructed using local principal component analysis (PCA) and alignment, as discussed in Section 2.

While diffusion maps and other non-linear dimensionality reduction methods are either directly or indirectly related to the heat kernel for functions over the data,

[*]Department of Mathematics and PACM, Princeton University, Fine Hall, Washington Road, Princeton NJ 08544-1000 USA, email: amits@math.princeton.edu
[†]Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton NJ 08544-1000 USA, email: hauwu@math.princeton.edu
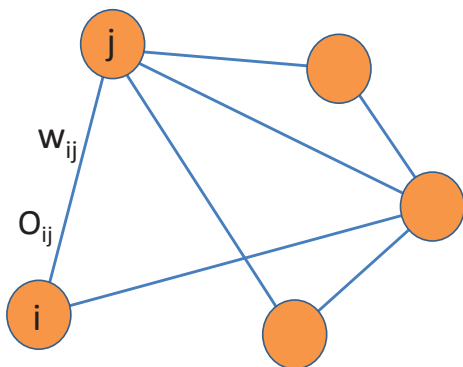
FIG. 1.1. *In VDM, the relationships between data points are represented as a weighted graph, where the weights w_{ij} are accompanied by linear orthogonal transformations O_{ij}.*
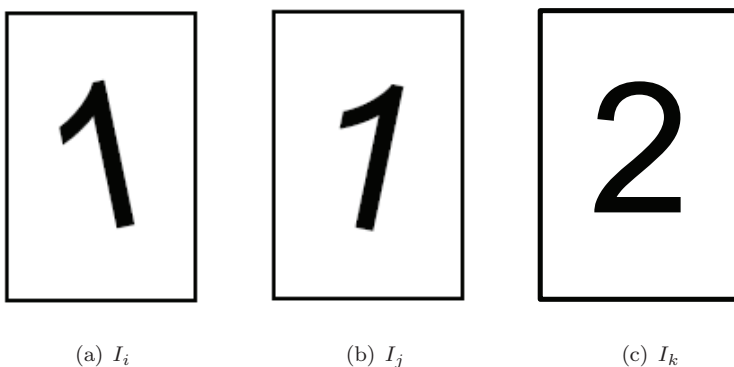

(a) I_i (b) I_j (c) I_k

FIG. 1.2. *An example of a weighted graph with orthogonal transformations: I_i and I_j are two different images of the digit one, corresponding to nodes i and j in the graph. O_{ij} is the 2×2 rotation matrix that rotationally aligns I_j with I_i and w_{ij} is some measure for the affinity between the two images when they are optimally aligned. The affinity w_{ij} is large, because the images I_i and $O_{ij}I_j$ are actually the same. On the other hand, I_k is an image of the digit two, and the discrepancy between I_k and I_i is large even when these images are optimally aligned. As a result, the affinity w_{ik} would be small, perhaps so small that there is no edge in the graph connecting nodes i and k. The matrix O_{ik} is clearly not as meaningful as O_{ij}. If there is no edge between i and k, then O_{ik} is not represented in the weighted graph.*

our VDM framework is based on the heat kernel for vector fields. We construct this kernel from the weighted graph and the orthogonal transformations. Through the spectral decomposition of this kernel, VDM defines an embedding of the data in a Hilbert space. In particular, it defines a metric for the data, that is, distances between data points that we call vector diffusion distances. For some applications, the vector diffusion metric is more meaningful than currently used metrics, since it takes into account the linear transformations, and as a result, it provides a better organization of the data. In the manifold learning setup, we prove a convergence theorem illuminating the relation between VDM and the connection-Laplacian operator for vector fields over the manifold.

The paper is organized in the following way: In Section 2 we describe the manifold learning setup and a procedure to extract the orthogonal transformations from a point cloud scattered in a high dimensional Euclidean space using local PCA and alignment. In Section 3 we specify the vector diffusion mapping of the data set into a finite dimensional Hilbert space. At the heart of the vector diffusion mapping construction lies a certain symmetric matrix that can be normalized in slightly different ways. Different normalizations lead to different embeddings, as discussed in Section 4. These normalizations resemble the normalizations of the graph Laplacian in spectral graph theory and spectral clustering algorithms. In the manifold learning setup, it is known that when the point cloud is uniformly sampled from a low dimensional Riemannian manifold, then the normalized graph Laplacian approximates the Laplace-Beltrami operator for scalar functions. In Section 5 we formulate a similar result, stated as Theorem 5.1, for the convergence of the appropriately normalized vector diffusion mapping matrix to the connection-Laplacian operator for vector fields [1]. The proof of Theorem 5.1 appears in Appendix B. We verified Theorem 5.1 numerically for spheres of different dimensions, as reported in Section 6 and Appendix C. We also used other surfaces to perform numerical comparisons between the vector diffusion distance, the diffusion distance, and the geodesic distance. In Section 7 we briefly discuss out-of-sample extrapolation of vector fields via the Nyström extension scheme. The role played by the heat kernel of the connection-Laplacian is discussed in Section 8. We use the well known short time asymptotic expansion of the heat kernel to show the relationship between vector diffusion distances and geodesic distances for nearby points. In Section 9 we briefly discuss the application of VDM to cryo-electron microscopy, as a prototypical multi-reference rotational alignment problem. We conclude in Section 10 with a summary followed by a discussion of some other possible applications and extensions of the mathematical framework.

2. Data sampled from a Riemannian manifold. One of the main objectives in the analysis of a high dimensional large data set is to learn its geometric and topological structure. Even though the data itself is parameterized as a point cloud in a high dimensional ambient space $\mathbb{R}^p$, the correlation between parameters often suggests the popular "manifold assumption" that the data points are distributed on (or near) a low dimensional Riemannian manifold $\mathcal{M}^d$ embedded in $\mathbb{R}^p$, where d is the dimension of the manifold and $d \ll p$. Suppose that the point cloud consists of n data points $x_1, x_2, \ldots, x_n$ that are viewed as points in $\mathbb{R}^p$ but are restricted to the manifold. We now describe how the orthogonal transformations O_{ij} can be constructed from the point cloud using local PCA and alignment.

Local PCA. For every data point x_i we suggest to estimate a basis to the tangent plane $T_{x_i}\mathcal{M}$ to the manifold at x_i using the following procedure which we refer to as local PCA. We fix a scale parameter $\epsilon_{\mathrm{PCA}} > 0$ and define $\mathcal{N}_{x_i,\epsilon_{\mathrm{PCA}}}$ as the neighbors of x_i inside a ball of radius $\sqrt{\epsilon_{\mathrm{PCA}}}$ centered at x_i:

$$\mathcal{N}_{x_i,\epsilon_{\mathrm{PCA}}} = \{x_j : 0 < \|x_j - x_i\|_{\mathbb{R}^p} < \sqrt{\epsilon_{\mathrm{PCA}}}\}.$$

---

[1] One of the main considerations in the way this paper is presented was to make it as accessible as possible, also to readers who are not familiar with differential geometry. Although the connection-Laplacian is essential to the understanding of the mathematical framework that underlies VDM, and differential geometry is extensively used in Appendix B for the proof of Theorem 5.1, we do not assume knowledge of differential geometry in Sections 2-10 (except for some parts of Section 8) that detail the algorithmic framework. The concepts of differential geometry that are required for achieving basic familiarity with the connection-Laplacian are explained in Appendix A.

Denote the number of neighboring points of x_i by[2] N_i, that is, $N_i = |\mathcal{N}_{x_i,\epsilon_{\mathrm{PCA}}}|$, and denote the neighbors of x_i by $x_{i_1}, x_{i_2}, \ldots, x_{i_{N_i}}$. We assume that ϵ_{PCA} is large enough so that $N_i \geq d$, but at the same time ϵ_{PCA} is small enough such that $N_i \ll n$. In Theorem B.1 we show that a satisfactory choice for ϵ_{PCA} is given by $\epsilon_{\mathrm{PCA}} = O(n^{-\frac{2}{d+1}})$, so that $N_i = O(n^{\frac{1}{d+1}})$. In fact, it is even possible to choose $\epsilon_{\mathrm{PCA}} = O(n^{-\frac{2}{d+2}})$ if the manifold does not have a boundary.

Observe that the neighboring points are located near $T_{x_i}\mathcal{M}$, where deviations are possible either due to curvature or due to neighboring data points that lie slightly off the manifold. Define X_i to be a $p \times N_i$ matrix whose j'th column is the vector $x_{i_j} - x_i$, that is,

$$X_i = \begin{bmatrix} x_{i_1} - x_i & x_{i_2} - x_i & \ldots & x_{i_{N_i}} - x_i \end{bmatrix}.$$

In other words, X_i is the data matrix of the neighbors shifted to be centered at the point x_i. Notice, that while it is more common to shift the data for PCA by the mean $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{i_j}$, here we shift the data by x_i. Shifting the data by μ_i is also possible for all practical purposes, but has the slight disadvantage of complicating the proof for the convergence of the local PCA step (see Appendix B.1).

Let K be a C^2 positive monotonic decreasing function with support on the interval $[0,1]$, for example, the Epanechnikov kernel $K(u) = (1 - u^2)\chi_{[0,1]}$, where χ is the indicator function [3] Let D_i be an $N_i \times N_i$ diagonal matrix with

$$D_i(j,j) = \sqrt{K\left(\frac{\|x_i - x_{i_j}\|_{\mathbb{R}^p}}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right)}, \quad j = 1, 2, \ldots, N_i.$$

Define the $p \times N_i$ matrix B_i as

$$B_i = X_i D_i.$$

That is, the j'th column of B_i is the vector $(x_{i_j} - x_i)$ scaled by $D_i(j,j)$. The purpose of the scaling is to give more emphasis to nearby points over far away points (recall that K is monotonic decreasing). We denote the singular values of B_i by $\sigma_{i,1} \geq \sigma_{i,2} \geq \cdots \geq \sigma_{i,N_i}$.

In many cases, the intrinsic dimension d is not known in advance and needs to be estimated directly from the data. If the neighboring points in $\mathcal{N}_{x_i,\epsilon_{\mathrm{PCA}}}$ are located exactly on $T_{x_i}\mathcal{M}$, then $\mathrm{rank}\, X_i = \mathrm{rank}\, B_i = d$, and there are only d non-vanishing singular values (i.e., $\sigma_{i,d+1} = \ldots = \sigma_{i,N_i} = 0$). In such a case, the dimension can be estimated as the number of non-zero singular values. In practice, however, due to the curvature effect, there may be more than d non-zero singular values. A common practice is to estimate the dimension as the number of singular values that account for high enough percentage of the variability of the data. That is, one sets a threshold γ between 0 and 1 (usually closer to 1 than to 0), and estimates the dimension as the smallest integer d_i for which

$$\frac{\sum_{j=1}^{d_i} \sigma_{i,j}^2}{\sum_{j=1}^{N_i} \sigma_{i,j}^2} > \gamma.$$

---

[2] Since N_i depends on ϵ_{PCA}, it should be denoted as $N_{i,\epsilon_{\mathrm{PCA}}}$, but since ϵ_{PCA} is kept fixed it is suppressed from the notation, a convention that we use except for cases in which confusion may arise.

[3] In fact, K can be chosen in a more general fashion, for example, monotonicity is not required for all theoretical purposes. However, in practice, a monotonic decreasing K leads to a better behavior of the PCA step.

For example, setting $\gamma = 0.9$ means that d_i singular values account for at least 90% variability of the data, while $d_i - 1$ singular values account for less than 90%. We refer to the smallest integer d_i as the estimated local dimension of $\mathcal{M}$ at x_i. One possible way to estimate the dimension of the manifold would be to use the mean of the estimated local dimensions $d_1, \ldots, d_n$, that is, $\hat{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$ (and then round it to the closest integer). The mean estimator minimizes the sum of squared errors $\sum_{i=1}^{n} (d_i - \hat{d})^2$. We estimate the intrinsic dimension of the manifold by the median value of all the d_i's, that is, we define the estimator $\hat{d}$ for the intrinsic dimension d as

$$\hat{d} = \text{median}\{d_1, d_2, \ldots, d_n\}.$$

The median has the property that it minimizes the sum of absolute errors $\sum_{i=1}^{n} |d_i - \hat{d}|$. As such, estimating the intrinsic dimension by the median is more robust to outliers compared to the mean estimator. In all proceeding steps of the algorithm we use the median estimator $\hat{d}$, but in order to facilitate the notation we write d instead of $\hat{d}$.

Suppose that the singular value decomposition (SVD) of B_i is given by

$$B_i = U_i \Sigma_i V_i^T.$$

The columns of the $p \times N_i$ matrix U_i are orthonormal and are known as the left singular vectors

$$U_i = \begin{bmatrix} u_{i_1} & u_{i_2} & \cdots & u_{i_{N_i}} \end{bmatrix}.$$

We define the $p \times d$ matrix O_i by the first d left singular vectors (corresponding to the largest singular values):

$$O_i = \begin{bmatrix} u_{i_1} & u_{i_2} & \cdots & u_{i_d} \end{bmatrix}. \tag{2.1}$$

The d columns of O_i are orthonormal, i.e., $O_i^T O_i = I_{d \times d}$. The columns of O_i represent an orthonormal basis to a d-dimensional subspace of $\mathbb{R}^p$. This basis is a numerical approximation to an orthonormal basis of the tangent plane $T_{x_i}\mathcal{M}$. The order of the approximation (as a function of ϵ_{PCA} and n) is established later, using the fact that the columns of O_i are also the eigenvectors (corresponding to the d largest eigenvalues) of the $p \times p$ covariance matrix Ξ_i given by

$$\Xi_i = \sum_{j=1}^{N_i} K\left(\frac{\|x_i - x_{i_j}\|_{\mathbb{R}^p}}{\sqrt{\epsilon_{\text{PCA}}}}\right)(x_{i_j} - x_i)(x_{i_j} - x_i)^T. \tag{2.2}$$

Since K is supported on the interval $[0, 1]$ the covariance matrix Ξ_i can also be represented as

$$\Xi_i = \sum_{j=1}^{n} K\left(\frac{\|x_i - x_j\|_{\mathbb{R}^p}}{\sqrt{\epsilon_{\text{PCA}}}}\right)(x_j - x_i)(x_j - x_i)^T. \tag{2.3}$$

We emphasize that the covariance matrix is never actually formed due to its excessive storage requirements, and all computations are performed with the matrix B_i. We remark that it is also possible to estimate the intrinsic dimension d and the basis O_i using the multiscaled PCA algorithm [28] that uses several different values of ϵ_{PCA} for a given x_i, but here we try to make our approach as simple as possible while being able to later prove convergence theorems.

Alignment. Suppose x_i and x_j are two nearby points whose Euclidean distance satisfies $\|x_i - x_j\|_{\mathbb{R}^p} < \sqrt{\epsilon}$, where $\epsilon > 0$ is a scale parameter different from the scale parameter ϵ_{PCA}. In fact, ϵ is much larger than ϵ_{PCA} as we later choose $\epsilon = O(n^{-\frac{2}{d+4}})$, while, as mentioned earlier, $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+1}})$ (manifolds with boundary) or $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+2}})$ (manifolds with no boundary). In any case, ϵ is small enough so that the tangent spaces $T_{x_i}\mathcal{M}$ and $T_{x_j}\mathcal{M}$ are also close.[4] Therefore, the column spaces of O_i and O_j are almost the same. If the subspaces were to be exactly the same, then the matrices O_i and O_j would have differ by a $d \times d$ orthogonal transformation O_{ij} satisfying $O_i O_{ij} = O_j$, or equivalently $O_{ij} = O_i^T O_j$. In that case, $O_i^T O_j$ is the matrix representation of the operator that transport vectors from $T_{x_j}\mathcal{M}$ to $T_{x_i}\mathcal{M}$, viewed as copies of $\mathbb{R}^d$. The subspaces, however, are usually not exactly the same, due to curvature. As a result, the matrix $O_i^T O_j$ is not necessarily orthogonal, and we define O_{ij} as its closest orthogonal matrix, i.e.,

$$O_{ij} = \operatorname*{argmin}_{O \in O(d)} \|O - O_i^T O_j\|_{HS}, \tag{2.4}$$

where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm (given by $\|A\|_{HS}^2 = \mathrm{Tr}(AA^T)$ for any real matrix A) and $O(d)$ is the set of orthogonal $d \times d$ matrices. This minimization problem has a simple solution [13, 25, 21, 1] via the SVD of $O_i^T O_j$. Specifically, if

$$O_i^T O_j = U\Sigma V^T$$

is the SVD of $O_i^T O_j$, then O_{ij} is given by

$$O_{ij} = UV^T.$$

We refer to the process of finding the optimal orthogonal transformation between bases as alignment. Later we show that the matrix O_{ij} is an approximation to the parallel transport operator (see Appendix A) from $T_{x_j}\mathcal{M}$ to $T_{x_i}\mathcal{M}$ whenever x_i and x_j are nearby.

Note that not all bases are aligned; only the bases of nearby points are aligned. We set E to be the edge set of the undirected graph over n vertices that correspond to the data points, where an edge between i and j exists iff their corresponding bases are aligned by the algorithm[5] (or equivalently, iff $0 < \|x_i - x_j\|_{\mathbb{R}^p} < \sqrt{\epsilon}$). The weights w_{ij} are defined using a kernel function K as [6]

$$w_{ij} = K\left(\frac{\|x_i - x_j\|_{\mathbb{R}^p}}{\sqrt{\epsilon}}\right), \tag{2.5}$$

where we assume that K is supported on the interval $[0, 1]$. For example, the Gaussian kernel $K(u) = \exp\{-u^2\}\chi_{[0,1]}$ leads to weights of the form $w_{ij} = \exp\{-\frac{\|x_i - x_j\|^2}{\epsilon}\}$ for $0 < \|x_i - x_j\| < \sqrt{\epsilon}$ and 0 otherwise. We emphasize that the kernel K used for the definition of the weights w_{ij} could be different than the kernel used for the previous step of local PCA.

---

[4]In the sense that their Grassmannian distance given approximately by the operator norm $\|O_i O_i^T - O_j O_j^T\|$ is small.

[5]We do not align a basis with itself, so the edge set E does not contain self loops of the form (i, i).

[6]Notice that the weights are only a function of the Euclidean distance between data points; another possibility, which we do not consider in this paper, is to include the Grassmannian distance $\|O_i O_i^T - O_j O_j^T\|_2$ into the definition of the weight.
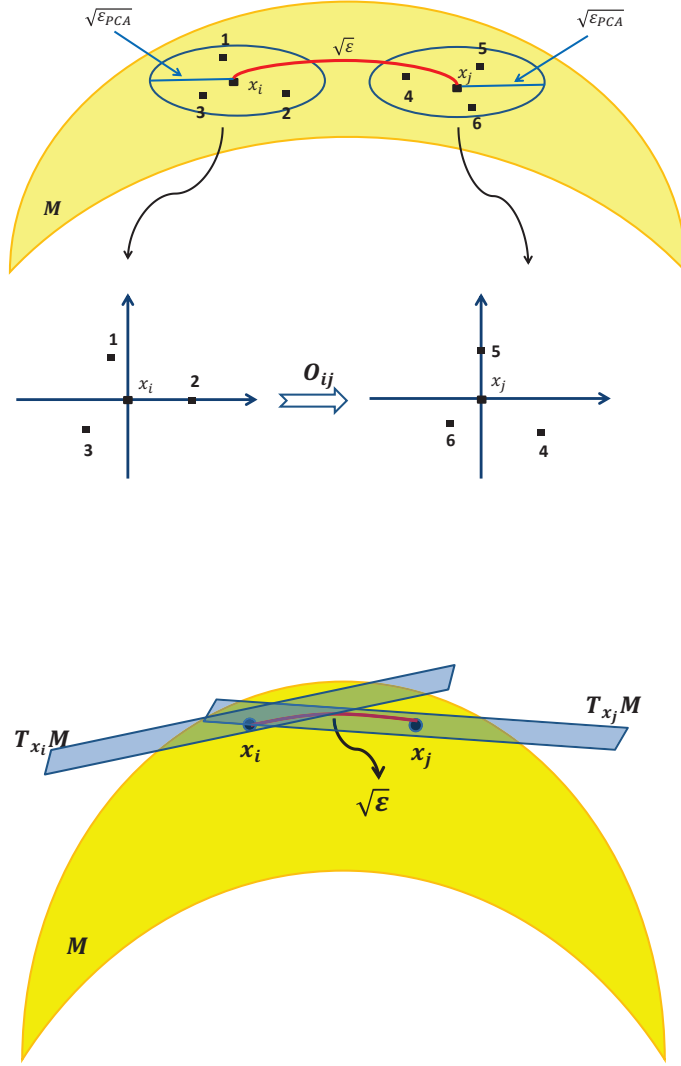
FIG. 2.1. *The orthonormal basis of the tangent plane $T_{x_i}\mathcal{M}$ is determined by local PCA using data points inside a Euclidean ball of radius $\sqrt{\varepsilon_{PCA}}$ centered at x_i. The bases for $T_{x_i}\mathcal{M}$ and $T_{x_j}\mathcal{M}$ are optimally aligned by an orthogonal transformation O_{ij} that can be viewed as a mapping from $T_{x_i}\mathcal{M}$ to $T_{x_j}\mathcal{M}$.*

3. Vector diffusion mapping. We construct the following matrix S:

$$S(i,j) = \begin{cases} w_{ij}O_{ij} & (i,j) \in E, \\ 0_{d\times d} & (i,j) \notin E. \end{cases} \tag{3.1}$$

That is, S is a block matrix, with $n \times n$ blocks, each of which is of size $d \times d$. Each block is either a $d \times d$ orthogonal transformation O_{ij} multiplied by the scalar weight

7

w_{ij}, or a zero $d \times d$ matrix.[7] The matrix S is symmetric since $O_{ij}^T = O_{ji}$ and $w_{ij} = w_{ji}$, and its overall size is $nd \times nd$. We define a diagonal matrix D of the same size, where the diagonal blocks are scalar matrices given by

$$D(i,i) = \deg(i)I_{d \times d}, \tag{3.2}$$

and

$$\deg(i) = \sum_{j:(i,j) \in E} w_{ij} \tag{3.3}$$

is the weighted degree of node i. The matrix $D^{-1}S$ can be applied to vectors v of length nd, which we regard as n vectors of length d, such that $v(i)$ is a vector in $\mathbb{R}^d$ viewed as a vector in $T_{x_i}\mathcal{M}$. The matrix $D^{-1}S$ is an averaging operator for vector fields, since

$$(D^{-1}Sv)(i) = \frac{1}{\deg(i)} \sum_{j:(i,j) \in E} w_{ij}O_{ij}v(j). \tag{3.4}$$

This implies that the operator $D^{-1}S$ transport vectors from the tangent spaces $T_{x_j}\mathcal{M}$ (that are nearby to $T_{x_i}\mathcal{M}$) to $T_{x_i}\mathcal{M}$ and then averages the transported vectors in $T_{x_i}\mathcal{M}$.

Notice that diffusion maps and other non-linear dimensionality reduction methods make use of the weight matrix $W = (w_{ij})_{i,j=1}^n$, but not of the transformations O_{ij}. In diffusion maps, the weights are used to define a discrete random walk over the graph, where the transition probability a_{ij} in a single time step from node i to node j is given by

$$a_{ij} = \frac{w_{ij}}{\deg(i)}. \tag{3.5}$$

The Markov transition matrix $A = (a_{ij})_{i,j=1}^n$ can be written as

$$A = \mathcal{D}^{-1}W, \tag{3.6}$$

where $\mathcal{D}$ is $n \times n$ diagonal matrix with

$$\mathcal{D}(i,i) = \deg(i). \tag{3.7}$$

While A is the Markov transition probability matrix in a single time step, A^t is the transition matrix for t steps. In particular, $A^t(i,j)$ sums the probabilities of all paths of length t that start at i and end at j. Coifman and Lafon [9, 26] showed that A^t can be used to define an inner product in a Hilbert space. Specifically, the matrix A is similar to the symmetric matrix $\mathcal{D}^{-1/2}W\mathcal{D}^{-1/2}$ through $A = \mathcal{D}^{-1/2}(\mathcal{D}^{-1/2}W\mathcal{D}^{-1/2})\mathcal{D}^{1/2}$. It follows that A has a complete set of real eigenvalues and eigenvectors $\{\mu_l\}_{l=1}^n$ and $\{\phi_l\}_{l=1}^n$, respectively, satisfying $A\phi_l = \mu_l\phi_l$. Their diffusion mapping Φ_t is given by

$$\Phi_t(i) = (\mu_1^t\phi_1(i), \mu_2^t\phi_2(i), \ldots, \mu_n^t\phi_n(i)), \tag{3.8}$$

---

[7] As mentioned in the previous footnote, the edge set does not contain self-loops, so $w_{ii} = 0$ and $S(i,i) = 0_{d \times d}$.

where $\phi_l(i)$ is the i'th entry of the eigenvector ϕ_l. The mapping Φ_t satisfies

$$\sum_{k=1}^{n} \frac{A^t(i,k)}{\sqrt{\deg(k)}} \frac{A^t(j,k)}{\sqrt{\deg(k)}} = \langle \Phi_t(i), \Phi_t(j) \rangle, \tag{3.9}$$

where $\langle \cdot, \cdot \rangle$ is the usual dot product over Euclidean space. The metric associated to this inner product is known as the *diffusion distance*. The diffusion distance $d_{\mathrm{DM},t}(i,j)$ between i and j is given by

$$d_{\mathrm{DM},t}^2(i,j) = \sum_{k=1}^{n} \frac{(A^t(i,k) - A^t(j,k))^2}{\deg(k)} = \langle \Phi_t(i), \Phi_t(i) \rangle + \langle \Phi_t(j), \Phi_t(j) \rangle - 2\langle \Phi_t(i), \Phi_t(j) \rangle. \tag{3.10}$$

Thus, the diffusion distance between i and j is the weighted-ℓ_2 proximity between the probability clouds of random walkers starting at i and j after t steps.

In the VDM framework, we define the affinity between i and j by considering all paths of length t connecting them, but instead of just summing the weights of all paths, we *sum the transformations*. A path of length t from j to i is some sequence of vertices $j_0, j_1, \ldots, j_t$ with $j_0 = j$ and $j_t = i$ and its corresponding orthogonal transformation is obtained by multiplying the orthogonal transformations along the path in the following order:

$$O_{j_t,j_{t-1}} \cdots O_{j_2,j_1} O_{j_1,j_0}. \tag{3.11}$$

Every path from j to i may therefore result in a different transformation. This is analogous to the parallel transport operator from differential geometry that depends on the path connecting two points whenever the manifold has curvature (e.g., the sphere). Thus, when adding transformations of different paths, cancelations may happen. We would like to define the affinity between i and j as the consistency between these transformations, with higher affinity expressing more agreement among the transformations that are being averaged. To quantify this affinity, we consider again the matrix $D^{-1}S$ which is similar to the symmetric matrix

$$\tilde{S} = D^{-1/2} S D^{-1/2} \tag{3.12}$$

through $D^{-1}S = D^{-1/2}\tilde{S}D^{1/2}$ and define the affinity between i and j as $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$, that is, as the squared HS norm of the $d \times d$ matrix $\tilde{S}^{2t}(i,j)$, which takes into account all paths of length $2t$, where t is a positive integer. In a sense, $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ measures not only the number of paths of length $2t$ connecting i and j but also the amount of agreement between their transformations. That is, for a fixed number of paths, $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ is larger when the path transformations are in agreement, and is smaller when they differ.

Since $\tilde{S}$ is symmetric, it has a complete set of eigenvectors $v_1, v_2, \ldots, v_{nd}$ and eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{nd}$. We order the eigenvalues in decreasing order of magnitude $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{nd}|$. The spectral decompositions of $\tilde{S}$ and $\tilde{S}^{2t}$ are given by

$$\tilde{S}(i,j) = \sum_{l=1}^{nd} \lambda_l v_l(i) v_l(j)^T, \quad \text{and} \quad \tilde{S}^{2t}(i,j) = \sum_{l=1}^{nd} \lambda_l^{2t} v_l(i) v_l(j)^T, \tag{3.13}$$

where $v_l(i) \in \mathbb{R}^d$ for $i = 1, \ldots, n$ and $l = 1, \ldots, nd$. The HS norm of $\tilde{S}^{2t}(i,j)$ is

calculated using the trace:

$$\|\tilde{S}^{2t}(i,j)\|_{HS}^2 = \text{Tr}\left[\tilde{S}^{2t}(i,j)\tilde{S}^{2t}(i,j)^T\right] = \sum_{l,r=1}^{nd} (\lambda_l\lambda_r)^{2t} \langle v_l(i), v_r(i)\rangle \langle v_l(j), v_r(j)\rangle.$$

(3.14)

It follows that the affinity $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ is an inner product for the finite dimensional Hilbert space $\mathbb{R}^{(nd)^2}$ via the mapping V_t:

$$V_t : i \mapsto \left((\lambda_l\lambda_r)^t \langle v_l(i), v_r(i)\rangle\right)_{l,r=1}^{nd}.$$

(3.15)

That is,

$$\|\tilde{S}^{2t}(i,j)\|_{HS}^2 = \langle V_t(i), V_t(j)\rangle.$$

(3.16)

Note that in the manifold learning setup, the embedding $i \mapsto V_t(i)$ is invariant to the choice of basis for $T_{x_i}\mathcal{M}$ because the dot products $\langle v_l(i), v_r(i)\rangle$ are invariant to orthogonal transformations. We refer to V_t as the vector diffusion mapping.

From the symmetry of the dot products $\langle v_l(i), v_r(i)\rangle = \langle v_r(i), v_l(i)\rangle$, it is clear that $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ is also an inner product for the finite dimensional Hilbert space $\mathbb{R}^{nd(nd+1)/2}$ corresponding to the mapping

$$i \mapsto \left(c_{lr}(\lambda_l\lambda_r)^t \langle v_l(i), v_r(i)\rangle\right)_{1 \le l \le r \le nd},$$

where

$$c_{lr} = \begin{cases} \sqrt{2} & l < r, \\ 1 & l = r. \end{cases}$$

We define the symmetric vector diffusion distance $d_{\text{VDM},t}(i,j)$ between nodes i and j as

$$d_{\text{VDM},t}^2(i,j) = \langle V_t(i), V_t(i)\rangle + \langle V_t(j), V_t(j)\rangle - 2\langle V_t(i), V_t(j)\rangle.$$

(3.17)

The matrices $I - \tilde{S}$ and $I + \tilde{S}$ are positive semidefinite due to the following identity:

$$v^T(I \pm D^{-1/2}SD^{-1/2})v = \sum_{(i,j)\in E} \left\|\frac{v(i)}{\sqrt{\deg(i)}} \pm \frac{w_{ij}O_{ij}v(j)}{\sqrt{\deg(j)}}\right\|^2 \ge 0,$$

(3.18)

for any $v \in \mathbb{R}^{nd}$. As a consequence, all eigenvalues λ_l of $\tilde{S}$ reside in the interval $[-1, 1]$. In particular, for large enough t, most terms of the form $(\lambda_l\lambda_r)^{2t}$ in (3.14) are close to 0, and $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ can be well approximated by using only the few largest eigenvalues and their corresponding eigenvectors. This lends itself into an efficient approximation of the vector diffusion distances $d_{\text{VDM},t}(i,j)$ of (3.17), and it is not necessary to raise the matrix $\tilde{S}$ to its $2t$ power (which usually results in dense matrices). Thus, for any $\delta > 0$, we define the truncated vector diffusion mapping V_t^δ that embeds the data set in $\mathbb{R}^{m^2}$ (or equivalently, but more efficiently in $\mathbb{R}^{m(m+1)/2}$) using the eigenvectors $v_1, \dots, v_m$ as

$$V_t^\delta : i \mapsto \left((\lambda_l\lambda_r)^t \langle v_l(i), v_r(i)\rangle\right)_{l,r=1}^{m}$$

(3.19)

10

where $m = m(t, \delta)$ is the largest integer for which $\left(\frac{\lambda_m}{\lambda_1}\right)^{2t} > \delta$ and $\left(\frac{\lambda_{m+1}}{\lambda_1}\right)^{2t} \leq \delta$.

We remark that we define V_t through $\|\tilde{S}^{2t}(i,j)\|_{HS}^2$ rather than through $\|\tilde{S}^t(i,j)\|_{HS}^2$, because we cannot guarantee that in general all eigenvalues of $\tilde{S}$ are non-negative. In Section 8, we show that in the continuous setup of the manifold learning problem all eigenvalues are non-negative. We anticipate that for most practical applications that correspond to the manifold assumption, all negative eigenvalues (if any) would be small in magnitude (say, smaller than δ). In such cases, one can use any real $t > 0$ for the truncated vector diffusion map V_t^δ.

4. Normalized Vector Diffusion Mappings. It is also possible to obtain slightly different vector diffusion mappings using different normalizations of the matrix S. These normalizations are similar to the ones used in the diffusion map framework [9]. For example, notice that

$$w_l = D^{-1/2} v_l \tag{4.1}$$

are the right eigenvectors of $D^{-1}S$, that is, $D^{-1}S w_l = \lambda_l w_l$. We can thus define another vector diffusion mapping, denoted V_t', as

$$V_t' : i \mapsto \left((\lambda_l \lambda_r)^t \langle w_l(i), w_r(i)\rangle\right)_{l,r=1}^{nd}. \tag{4.2}$$

From (4.1) it follows that V_t' and V_t satisfy the relations

$$V_t'(i) = \frac{1}{\deg(i)} V_t(i), \tag{4.3}$$

and

$$\langle V_t'(i), V_t'(j)\rangle = \frac{\langle V_t(i), V_t(j)\rangle}{\deg(i)\deg(j)}. \tag{4.4}$$

As a result,

$$\langle V_t'(i), V_t'(j)\rangle = \frac{\|\tilde{S}^{2t}(i,j)\|_{HS}^2}{\deg(i)\deg(j)} = \frac{\|(D^{-1}S)^{2t}(i,j)\|_{HS}^2}{\deg(j)^2}. \tag{4.5}$$

In other words, the Hilbert-Schmidt norm of the matrix $D^{-1}S$ leads to an embedding of the data set in a Hilbert space only upon proper normalization by the vertex degrees (similar to the normalization by the vertex degrees in (3.9) and (3.10) for the diffusion map). We define the associated vector diffusion distances as

$$d_{\text{VDM}',t}^2(i,j) = \langle V_t'(i), V_t'(i)\rangle + \langle V_t'(j), V_t'(j)\rangle - 2\langle V_t'(i), V_t'(j)\rangle. \tag{4.6}$$

The distances are related by $d_{\text{VDM}',t}^2(i,j) = \frac{d_{\text{VDM},t}^2(i,j)}{\deg(i)\deg(j)}$.

We comment that the normalized mappings $i \mapsto \frac{V_t(i)}{\|V_t(i)\|}$ and $i \mapsto \frac{V_t'(i)}{\|V_t'(i)\|}$ that map the data points to the unit sphere are equivalent, that is,

$$\frac{V_t'(i)}{\|V_t'(i)\|} = \frac{V_t(i)}{\|V_t(i)\|}. \tag{4.7}$$

This means that the angles between pairs of embedded points are the same for both mappings. For diffusion map, it has been observed that in some cases the distances

11

$\|\frac{\Phi_t(i)}{\|\Phi_t(i)\|} - \frac{\Phi_t(i)}{\|\Phi_t(i)\|}\|$ are more meaningful than $\|\Phi_t(i) - \Phi_t(j)\|$ (see, for example, [17]). This may also suggest the usage of the distances $\|\frac{V_t(i)}{\|V_t(i)\|} - \frac{V_t(i)}{\|V_t(i)\|}\|$ in the VDM framework.

Another important family of normalized diffusion mappings is obtained by the following procedure. Suppose $0 \le \alpha \le 1$, and define the symmetric matrices W_α and S_α as

$$W_\alpha = \mathcal{D}^{-\alpha} W \mathcal{D}^{-\alpha}, \tag{4.8}$$

and

$$S_\alpha = D^{-\alpha} S D^{-\alpha}. \tag{4.9}$$

We define the weighted degrees $\deg_\alpha(1), \ldots, \deg_\alpha(n)$ corresponding to W_α by

$$\deg_\alpha(i) = \sum_{j=1}^{n} W_\alpha(i, j),$$

the $n \times n$ diagonal matrix $\mathcal{D}_\alpha$ as

$$\mathcal{D}_\alpha(i, i) = \deg_\alpha(i), \tag{4.10}$$

and the $n \times n$ block diagonal matrix D_α (with blocks of size $d \times d$) as

$$D_\alpha(i, i) = \deg_\alpha(i) I_{d \times d}. \tag{4.11}$$

We can then use the matrices S_α and D_α (instead of S and D) to define the vector diffusion mappings $V_{\alpha,t}$ and $V'_{\alpha,t}$. Notice that for $\alpha = 0$ we have $S_0 = S$ and $D_0 = D$, so that $V_{0,t} = V_t$ and $V'_{0,t} = V'_t$. The case $\alpha = 1$ turns out to be especially important as discussed in the next Section.

5. Convergence to the connection-Laplacian. For diffusion maps, the discrete random walk over the data points converges to a continuous diffusion process over that manifold in the limit $n \to \infty$ and $\epsilon \to 0$. This convergence can be stated in terms of the normalized graph Laplacian L given by

$$L = \mathcal{D}^{-1} W - I.$$

In the case where the data points $\{x_i\}_{i=1}^n$ are sampled independently from the uniform distribution over $\mathcal{M}^d$, the graph Laplacian converges pointwise to the Laplace-Beltrami operator, as we have the following proposition [26, 3, 35, 20]: If $f : \mathcal{M}^d \to \mathbb{R}$ is a smooth function (e.g., $f \in C^3(\mathcal{M})$), then with high probability

$$\frac{1}{\epsilon} \sum_{j=1}^{N} L_{ij} f(x_j) = \frac{1}{2} \Delta_\mathcal{M} f(x_i) + O\left(\epsilon + \frac{1}{n^{1/2} \epsilon^{1/2 + d/4}}\right), \tag{5.1}$$

where $\Delta_\mathcal{M}$ is the Laplace-Beltrami operator on $\mathcal{M}^d$. The error consists of two terms: a bias term $O(\epsilon)$ and a variance term that decreases as $1/\sqrt{n}$, but also depends on ϵ. Balancing the two terms may lead to an optimal choice of the parameter ϵ as a function of the number of points n. In the case of uniform sampling, Belkin and Niyogi [4] have shown that the eigenvectors of the graph Laplacian converge to the

12

eigenfunctions of the Laplace-Beltrami operator on the manifold, which is stronger than the pointwise convergence given in (5.1).

In the case where the data points $\{x_i\}_{i=1}^{n}$ are independently sampled from a probability density function $p(x)$ whose support is a d-dimensional manifold $\mathcal{M}^d$ and satisfies some mild conditions, the graph Laplacian converges pointwise to the Fokker-Planck operator as stated in following proposition [26, 3, 35, 20]: If $f \in C^3(\mathcal{M})$, then with high probability

$$\frac{1}{\epsilon} \sum_{j=1}^{N} L_{ij} f(x_j) = \frac{1}{2} \Delta_{\mathcal{M}} f(x_i) + \nabla U(x_i) \cdot \nabla f(x_i) + O\left(\epsilon + \frac{1}{n^{1/2}\epsilon^{1/2+d/4}}\right), \quad (5.2)$$

where the potential term U is given by $U(x) = -2\log p(x)$. The error is interpreted in the same way as in the uniform sampling case. In [9] it is shown that it is possible to recover the Laplace-Beltrami operator also for non-uniform sampling processes using W_1 and $\mathcal{D}_1$ (that correspond to $\alpha = 1$ in (4.8) and (4.11)). The matrix $\mathcal{D}_1^{-1}W_1 - I$ converges to the Laplace-Beltrami operator independently of the sampling density function $p(x)$.

For VDM, we prove in Appendix B the following theorem, Theorem 5.1, that states that the matrix $D_\alpha^{-1} S_\alpha - I$, where $0 \le \alpha \le 1$, converges to the connection-Laplacian operator (defined via the covariant derivative, see Appendix A and [32]) plus some potential terms depending on $p(x)$. In particular, $D_1^{-1} S_1 - I$ converges to the connection-Laplacian operator, without any additional potential terms. Using the terminology of spectral graph theory, it may thus be appropriate to call $D_1^{-1} S_1 - I$ the *connection-Laplacian of the graph*.

The main content of Theorem 5.1 specifies the way in which VDM generalizes diffusion maps: while diffusion mapping is based on the heat kernel and the Laplace-Beltrami operator over scalar functions, VDM is based on the heat kernel and the connection-Laplacian over vector fields. While for diffusion maps, the computed eigenvectors are discrete approximations of the Laplacian eigenfunctions, for VDM, the l-th eigenvector v_l of $D_1^{-1} S_1 - I$ is a discrete approximation of the l-th eigen-vector field X_l of the connection-Laplacian ∇^2 over $\mathcal{M}$, which satisfies $\nabla^2 X_l = -\lambda_l X_l$ for some $\lambda_l \ge 0$.

In the formulation of the Theorem 5.1, as well as in the remainder of the paper, we slightly change the notation used so far in the paper, as we denote the observed data points in $\mathbb{R}^p$ by $\iota(x_1), \iota(x_2), \ldots, \iota(x_n)$, where $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^p$ is the embedding of the Riemannian manifold $\mathcal{M}$ in $\mathbb{R}^p$. Furthermore, we denote by $\iota_* T_{x_i} \mathcal{M}$ the d-dimensional subspace of $\mathbb{R}^p$ which is the embedding of $T_{x_i} \mathcal{M}$ in $\mathbb{R}^p$. It is important to note that in the manifold learning setup, the manifold $\mathcal{M}$, the embedding ι and the points $x_1, x_2, \ldots, x_n \in \mathcal{M}$ are assumed to exist but cannot be directly observed.

THEOREM 5.1. *Let $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^p$ be a smooth d-dim closed Riemannian manifold embedded in $\mathbb{R}^p$, with metric g induced from the canonical metric on $\mathbb{R}^p$. Let $K \in C^2([0,1))$ be a positive function. For $\epsilon > 0$, let $K_\epsilon(x_i, x_j) = K\left(\frac{\|\iota(x_i) - \iota(x_j)\|_{\mathbb{R}^p}}{\sqrt{\epsilon}}\right)$ for $0 < \|\iota(x_i) - \iota(x_j)\| < \sqrt{\epsilon}$, and $K_\epsilon(x_i, x_j) = 0$ otherwise. Let the data set $\{x_i\}_{i=1,\ldots,n}$ be independently distributed according to the probability density function $p(x)$ supported on $\mathcal{M}$, where p is uniformly bounded from below and above, that is, $0 < p_m \le p(x) \le p_M < \infty$. Define the estimated probability density distribution by*

$$p_\epsilon(x_i) = \sum_{j=1}^{n} K_\epsilon(x_i, x_j)$$

13

and for $0 \le \alpha \le 1$ define the α-normalized kernel $K_{\epsilon,\alpha}$ by

$$K_{\epsilon,\alpha}(x_i, x_j) = \frac{K_\epsilon(x_i, x_j)}{p_\epsilon^\alpha(x_i) p_\epsilon^\alpha(x_j)}.$$

Then, using $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$ for $X \in C^3(T\mathcal{M})$ and for all x_i with high probability (w.h.p.)

$$\frac{1}{\epsilon} \left[\frac{\sum_{j=1}^n K_{\epsilon,\alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1}^n K_{\epsilon,\alpha}(x_i, x_j)} - \bar{X}_i \right]$$

$$= \frac{m_2}{2dm_0} \left(\left\langle \iota_* \left\{ \nabla^2 X(x_i) + d\frac{\int_{S^{d-1}} \nabla_\theta X(x_i) \nabla_\theta (p^{1-\alpha})(x_i) d\theta}{p^{1-\alpha}(x_i)} \right\}, u_l(x_i) \right\rangle \right)_{l=1}^d \quad (5.3)$$

$$+ O\left(\epsilon^{1/2} + \epsilon^{-1} n^{-\frac{3}{d+2}} + n^{-1/2} \epsilon^{-(d/4+1/2)} \right)$$

$$= \frac{m_2}{2dm_0} \left(\left\langle \iota_* \left\{ \nabla^2 X(x_i) + d\frac{\int_{S^{d-1}} \nabla_\theta X(x_i) \nabla_\theta (p^{1-\alpha})(x_i) d\theta}{p^{1-\alpha}(x_i)} \right\}, e_l(x_i) \right\rangle \right)_{l=1}^d$$

$$+ O\left(\epsilon^{1/2} + \epsilon^{-1} n^{-\frac{3}{d+2}} + n^{-1/2} \epsilon^{-(d/4+1/2)} \right)$$

where ∇^2 is the connection-Laplacian, $\bar{X}_i \equiv (\langle \iota_* X(x_i), u_l(x_i) \rangle)_{l=1}^d \in \mathbb{R}^d$ for all i, $\{u_l(x_i)\}_{l=1,\dots,d}$ is an orthonormal basis for a d-dimensional subspace of $\mathbb{R}^p$ determined by local PCA (i.e., the columns of O_i), $\{e_l(x_i)\}_{l=1,\dots,d}$ is an orthonormal basis for $\iota_* T_{x_i} \mathcal{M}$, $m_l = \int_{\mathbb{R}^d} \|x\|^l K(\|x\|) dx$, and O_{ij} is the optimal orthogonal transformation determined by the alignment procedure. In particular, when $\alpha = 1$ we have

$$\frac{1}{\epsilon} \left[\frac{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j)} - \bar{X}_i \right] = \frac{m_2}{2dm_0} \left(\langle \iota_* \nabla^2 X(x_i), e_l(x_i) \rangle \right)_{l=1}^d \quad (5.4)$$

$$+ O\left(\epsilon^{1/2} + \epsilon^{-1} n^{-\frac{3}{d+2}} + n^{-1/2} \epsilon^{-(d/4+1/2)} \right).$$

Furthermore, for $\epsilon = O(n^{-\frac{2}{d+4}})$, almost surely,

$$\lim_{n\to\infty} \frac{1}{\epsilon} \left[\frac{\sum_{j=1}^n K_{\epsilon,\alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1}^n K_{\epsilon,\alpha}(x_i, x_j)} - \bar{X}_i \right] \quad (5.5)$$

$$= \frac{m_2}{2dm_0} \left(\left\langle \iota_* \left\{ \nabla^2 X(x_i) + d\frac{\int_{S^{d-1}} \nabla_\theta X(x_i) \nabla_\theta (p^{1-\alpha})(x_i) d\theta}{p^{1-\alpha}(x_i)} \right\}, e_l(x_i) \right\rangle \right)_{l=1}^d,$$

and

$$\lim_{n\to\infty} \frac{1}{\epsilon} \left[\frac{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j)} - \bar{X}_i \right] = \frac{m_2}{2dm_0} \left(\langle \iota_* \nabla^2 X(x_i), e_l(x_i) \rangle \right)_{l=1}^d. \quad (5.6)$$

When $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$ then the same almost surely convergence results above hold but with a slower convergence rate.

When the manifold is compact with boundary, (5.4) does not hold at the boundary. However, we have the following result for the convergence behavior near the boundary:

THEOREM 5.2. *Let* $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^p$ *be a smooth d-dim compact Riemannian manifold with smooth boundary $\partial \mathcal{M}$ embedded in $\mathbb{R}^p$, with metric g induced from the canonical metric on $\mathbb{R}^p$. Let $\{x_i\}_{i=1,...,n}$, $p(x)$, $K_{\epsilon,1}(x_i, x_j)$, $p_\epsilon(x_i)$, $\{e_l(x_i)\}_{l=1,...,d}$ and $\bar{X}_i$ be defined in the same way as in Theorem 5.1. Choose $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$. Denote $\mathcal{M}_{\sqrt{\epsilon}} = \{x \in \mathcal{M} : \min_{y \in \partial \mathcal{M}} d(x, y) \le \sqrt{\epsilon}\}$, where $d(x, y)$ is the geodesic distance between x and y. When $x_i \in \mathcal{M}_{\sqrt{\epsilon}}$, we have*

$$\frac{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1}^n K_{\epsilon,1}(x_i, x_j)} = \left(\left\langle \iota_* P_{x_i, x_0} \left(X(x_0) + \frac{m_1^\epsilon}{m_0^\epsilon} \nabla_{\partial_d} X(x_0) \right), e_l(x_i) \right\rangle \right)_{l=1}^d$$

$$+ O \left(\epsilon + n^{-\frac{3}{2(d+1)}} + n^{-1/2} \epsilon^{-(d/4 - 1/2)} \right), \qquad (5.7)$$

where $x_0 = \operatorname{argmin}_{y \in \partial \mathcal{M}} d(x_i, y)$, P_{x_i, x_0} is the parallel transport from x_0 to x_i along the geodesic linking them, m_1^ϵ and m_0^ϵ are constants defined in (B.99) and (B.100), and ∂_d is the normal direction to the boundary at x_0.

For the choice $\epsilon = O(n^{-\frac{2}{d+4}})$ (as in Theorem 5.1), the error appearing in (5.7) is $O(\epsilon^{3/4})$ which is asymptotically smaller than $O(\sqrt{\epsilon})$, which is the order of $\frac{m_1^\epsilon}{m_0^\epsilon}$. A consequence of Theorem 5.1, Theorem 5.2 and the above discussion about the error terms is that the eigenvectors of $D_1^{-1} S_1 - I$ are discrete approximations of the eigen-vector-fields of the connection-Laplacian operator with homogeneous Neumann boundary condition that satisfy

$$\begin{cases} \nabla^2 X(x) = -\lambda X(x), & \text{for } x \in \mathcal{M}, \\ \nabla_{\partial_d} X(x) = 0, & \text{for } x \in \partial \mathcal{M}. \end{cases} \qquad (5.8)$$

We remark that the Neumann boundary condition also emerges for the choice $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$. This is due to the fact that the error in the local PCA term is $O(\epsilon_{PCA}^{1/2}) = O(n^{-\frac{1}{d+2}})$, which is asymptotically smaller than $O(\epsilon^{1/2}) = O(n^{-\frac{1}{d+4}})$ error term.

Finally, Theorem 5.3 details the way in which the algorithm approximates the continuous heat kernel of the connection-Laplacian:

THEOREM 5.3. *Let $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^p$ be a smooth d-dim compact Riemannian manifold embedded in $\mathbb{R}^p$, with metric g induced from the canonical metric on $\mathbb{R}^p$ and ∇^2 be the connection Laplacian. For $0 \le \alpha \le 1$, define*

$$T_{\epsilon,\alpha} X(x) = \frac{\int_{\mathcal{M}} K_{\epsilon,\alpha}(x, y) P_{x,y} X(y) \mathrm{d}V(y)}{\int_{\mathcal{M}} K_{\epsilon,\alpha}(x, y) \mathrm{d}V(y)},$$

where $P_{x,y} : T_y \mathcal{M} \to T_x \mathcal{M}$ is the parallel transport operator from y to x along the geodesic connecting them.

Then, for any $t > 0$, the heat kernel $e^{t\nabla^2}$ can be approximated on $L^2(T\mathcal{M})$ (the space of squared-integrable vector fields) by $T_{\epsilon,1}^{\frac{t}{\epsilon}}$, that is,

$$\lim_{\epsilon \to 0} T_{\epsilon,1}^{\frac{t}{\epsilon}} = e^{t\nabla^2},$$

in the L^2 sense.

6. Numerical simulations. In all numerical experiments reported in this Section, we use the normalized vector diffusion mapping $V'_{1,t}$ corresponding to $\alpha = 1$ in (4.9) and (4.10), that is, we use the eigenvectors of $D_1^{-1} S_1$ to define the VDM. In all experiments we used the kernel function $K(u) = e^{-5u^2} \chi_{[0,1]}$ for the local PCA step

15

as well as for the definition of the weights w_{ij}. The specific choices for ϵ and ϵ_{PCA} are detailed below. We remark that the results are not very sensitive to these choices, that is, similar results are obtained for a wide regime of parameters.

The purpose of the first experiment is to numerically verify Theorem 5.1 using spheres of different dimensions. Specifically, we sampled $n = 8000$ points uniformly from S^d embedded in $\mathbb{R}^{d+1}$ for $d = 2, 3, 4, 5$. Figure 6.1 shows bar plots of the largest 30 eigenvalues of the matrix $D_1^{-1} S_1$ for $\epsilon_{\mathrm{PCA}} = 0.1$ when $d = 2, 3, 4$ and $\epsilon_{\mathrm{PCA}} = 0.2$ when $d = 5$, and $\epsilon = \epsilon_{\mathrm{PCA}}^{\frac{d+1}{d+4}}$. It is noticeable that the eigenvalues have numerical multiplicities greater than 1. Since the connection-Laplacian commutes with rotations, the dimensions of its eigenspaces can be calculated using representation theory (see Appendix C). In particular, our calculation predicted the following dimensions for the eigenspaces of the largest eigenvalues:

$$S^2 : 6, 10, 14, \ldots. \quad S^3 : 4, 6, 9, 16, 16, \ldots. \quad S^4 : 5, 10, 14, \ldots. \quad S^5 : 6, 15, 20, \ldots.$$

These dimensions are in full agreement with the bar plots shown in Figure 6.1.

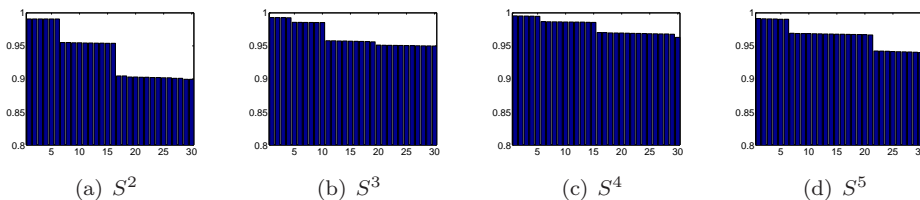| (a) S^2 | (b) S^3 | (c) S^4 | (d) S^5 |

FIG. 6.1. *Bar plots of the largest 30 eigenvalues of $D_1^{-1} S_1$ for $n = 8000$ points uniformly distributed over spheres of different dimensions.*

In the second set of experiments, we numerically compare the vector diffusion distance, the diffusion distance, and the geodesic distance for different compact manifolds with and without boundaries. The comparison is performed for the following four manifolds: 1) the sphere S^2 embedded in $\mathbb{R}^3$; 2) the torus T^2 embedded in $\mathbb{R}^3$; 3) the interval $[-\pi, \pi]$ in $\mathbb{R}$; and 4) the square $[0, 2\pi] \times [0, 2\pi]$ in $\mathbb{R}^2$. For both VDM and DM we truncate the mappings using $\delta = 0.2$, see (3.19). The geodesic distance is computed by the algorithm of Dijkstra on a weighted graph, whose vertices correspond to the data points, the edges link data points whose Euclidean distance is less than $\sqrt{\epsilon}$, and the weights $w_G(i, j)$ are the Euclidean distances, that is,

$$w_G(i, j) = \begin{cases} \|x_i - x_j\|_{\mathbb{R}^p} & \|x_i - x_j\| < \sqrt{\epsilon}, \\ +\infty & \text{otherwise.} \end{cases}$$

S^2 case: we sampled $n = 5000$ points uniformly from $S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\} \subset \mathbb{R}^3$ and set $\epsilon_{\mathrm{PCA}} = 0.1$ and $\epsilon = \sqrt{\epsilon_{\mathrm{PCA}}} \approx 0.316$. For the truncated vector diffusion distance, when $t = 10$, we find that the number of eigenvectors whose eigenvalue is larger (in magnitude) than $\lambda_1^t \delta$ is $m_{\mathrm{VDM}} = m_{\mathrm{VDM}}(t = 10, \delta = 0.2) = 16$ (recall the definition of $m(t, \delta)$ that appears after (3.19)). The corresponding embedded dimension is $m_{\mathrm{VDM}}(m_{\mathrm{VDM}}+1)/2$, which in this case is $16 \cdot 17/2 = 136$. Similarly, for $t = 100$, $m_{\mathrm{VDM}} = 6$ (embedded dimension is $6 \cdot 7/2 = 21$), and when $t = 1000$, $m_{\mathrm{VDM}} = 6$ (embedded dimension is again 21). Although the first eigenspace (corresponding the largest eigenvalue) of the connection-Laplacian over S^2 is of dimension 6, there are small discrepancies between the top 6 numerically computed eigenvalues, due to the

16

finite sampling. This numerical discrepancy is amplified upon raising the eigenvalues to the t'th power, when t is large, e.g., $t = 1000$. For demonstration purposes, we remedy this numerical effect by artificially setting $\lambda_l = \lambda_1$ for $l = 2, ..., 6$. For the truncated diffusion distance, when $t = 10$, $m_{\mathrm{DM}} = 36$ (embedded dimension is $36 - 1 = 35$), when $t = 100$, $m_{\mathrm{DM}} = 4$ (embedded dimension is 3), and when $t = 1000$, $m_{\mathrm{DM}} = 4$ (embedded dimension is 3). Similarly, we have the same numerical effect when $t = 1000$, that is, μ_2, μ_3 and μ_4 are close but not exactly the same, so we again set $\mu_l = \mu_2$ for $l = 3, 4$. The results are shown in Figure 6.2.

(a) $d_{\mathrm{VDM'},t=10}$ (b) $d_{\mathrm{VDM'},t=100}$ (c) $d_{\mathrm{VDM'},t=1000}$

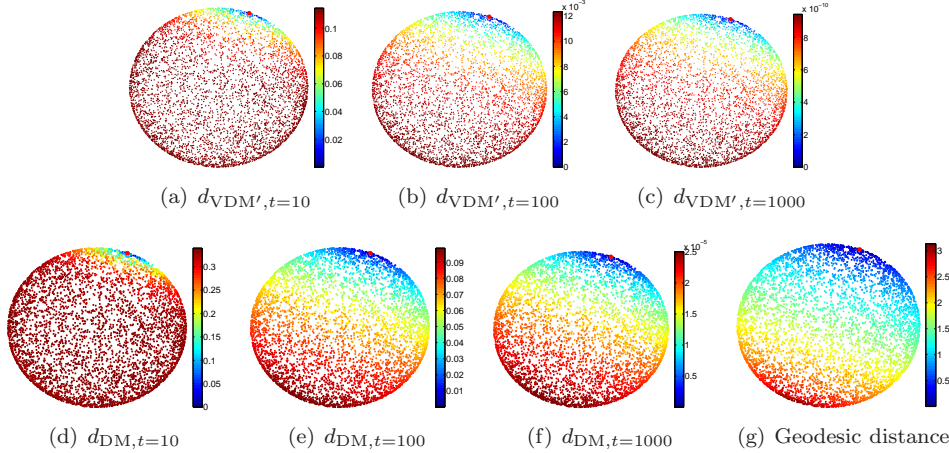(d) $d_{\mathrm{DM},t=10}$ (e) $d_{\mathrm{DM},t=100}$ (f) $d_{\mathrm{DM},t=1000}$ (g) Geodesic distance

FIG. 6.2. S^2 case. Top: truncated vector diffusion distances for $t = 10$, $t = 100$ and $t = 1000$; Bottom: truncated diffusion distances for $t = 10$, $t = 100$ and $t = 1000$, and the geodesic distance. The reference point from which distances are computed is marked in red.

T^2 case: we sampled $n = 5000$ points (u, v) uniformly over the square $[0, 2\pi) \times [0, 2\pi)$ and then mapped them to $\mathbb{R}^3$ using the following transformation that defines the surface T^2 as

$$T^2 = \{((2 + \cos(v)) \cos(u), (2 + \cos(v)) \sin(u), \sin(v)) : (u, v) \in [0, 2\pi) \times [0, 2\pi)\} \subset \mathbb{R}^3.$$

Notice that the resulting sample points are non-uniformly distributed over T^2. Therefore, the usage of S_1 and D_1 instead of S and D is important if we want the eigenvectors to approximate the eigen-vector-fields of the connection-Laplacian over T^2. We used $\epsilon_{\mathrm{PCA}} = 0.2$ and $\epsilon = \sqrt{\epsilon_{\mathrm{PCA}}} \approx 0.447$, and find that for the truncated vector diffusion distance, when $t = 10$, the embedded dimension is 2628, when $t = 100$, the embedded dimension is 36, and when $t = 1000$, the embedded dimension is 3. For the truncated diffusion distance, when $t = 10$, the embedded dimension is 130, when $t = 100$, the embedded dimension is 14, and when $t = 1000$, the embedded dimension is 2. The results are shown in Figure 6.3.

1-dim interval case: we sampled $n = 5000$ equally spaced grid points from the interval $[-\pi, \pi] \subset \mathbb{R}^1$ and set $\epsilon_{\mathrm{PCA}} = 0.01$ and $\epsilon = \epsilon_{\mathrm{PCA}}^{2/5} \approx 0.158$. For the truncated vector diffusion distance, when $t = 10$, the embedded dimension is 120, when $t = 100$, the embedded dimension is 15, and when $t = 1000$, the embedded dimension is 3. For the truncated diffusion distance, when $t = 10$, the embedded dimension is 36, when $t = 100$, the embedded dimension is 11, and when $t = 1000$, the embedded dimension is 3. The results are shown in Figure 6.4.
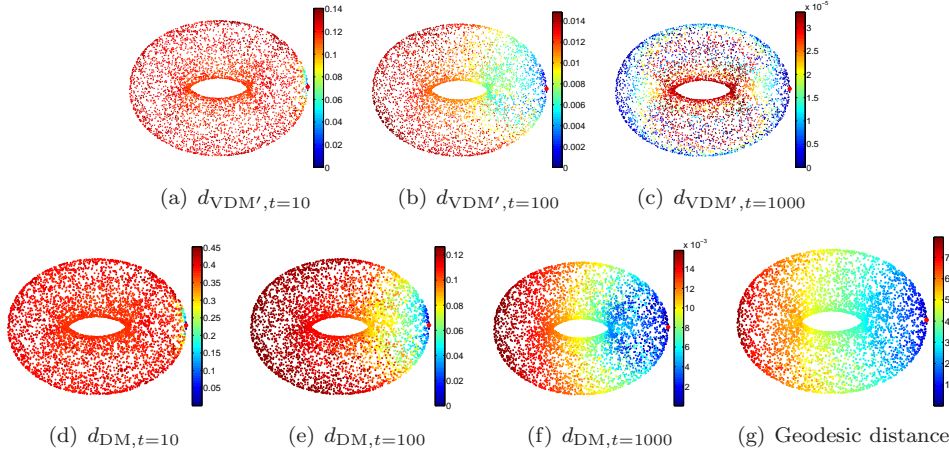
17

(a) $d_{\mathrm{VDM}',t=10}$ (b) $d_{\mathrm{VDM}',t=100}$ (c) $d_{\mathrm{VDM}',t=1000}$

(d) $d_{\mathrm{DM},t=10}$ (e) $d_{\mathrm{DM},t=100}$ (f) $d_{\mathrm{DM},t=1000}$ (g) Geodesic distance

FIG. 6.3. T^2 case. Top: truncated vector diffusion distances for $t = 10$, $t = 100$ and $t = 1000$; Bottom: truncated diffusion distances for $t = 10$, $t = 100$ and $t = 1000$, and the geodesic distance. The reference point from which distances are computed is marked in red.

(a) $d_{\mathrm{VDM}',t=10}$ (b) $d_{\mathrm{VDM}',t=100}$ (c) $d_{\mathrm{VDM}',t=1000}$

(d) $d_{\mathrm{DM},t=10}$ (e) $d_{\mathrm{DM},t=100}$ (f) $d_{\mathrm{DM},t=1000}$ (g) Geodesic distance
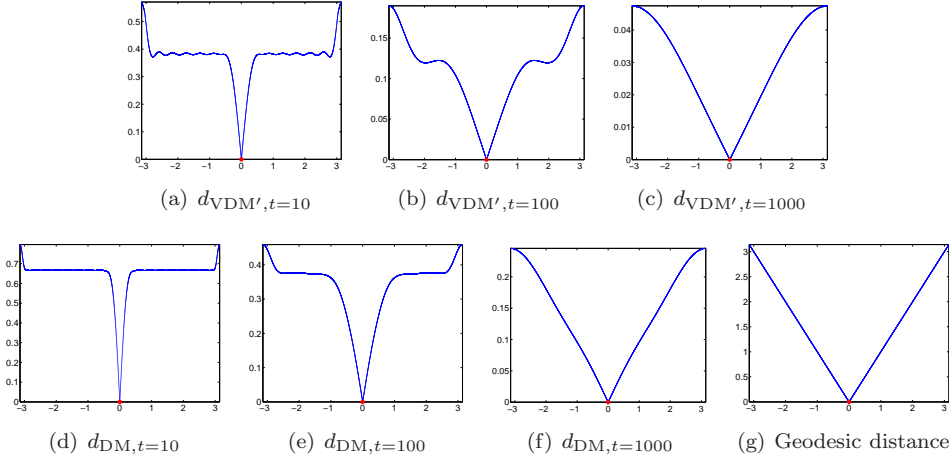
FIG. 6.4. 1-dim interval case. Top: truncated vector diffusion distances for $t = 10$, $t = 100$ and $t = 1000$; Bottom: truncated diffusion distances for $t = 10$, $t = 100$ and $t = 1000$, and the geodesic distance. The reference point from which distances are computed is marked in red.

Square case: we sampled $n = 6561 = 81^2$ equally spaced grid points from the square $[0, 2\pi] \times [0, 2\pi]$ and fix $\epsilon_{\mathrm{PCA}} = 0.01$ and $\epsilon = \sqrt{\epsilon_{\mathrm{PCA}}} = 0.1$. For the truncated vector diffusion distance, when $t = 10$, the embedded dimension is 20100 (we only calculate the first 200 eigenvalues), when $t = 100$, the embedded dimension is 1596, and when $t = 1000$, the embedded dimension is 36. For the truncated diffusion distance, when $t = 10$, the embedded dimension is 200 (we only calculate the first 200 eigenvalues), when $t = 100$, the embedded dimension is 200, and when $t = 1000$, the embedded dimension is 28. The results are shown in Figure 6.5.

7. Out-of-sample extension of vector fields. Let $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{Y} = \{y_i\}_{i=1}^m$ so that $\mathcal{X}, \mathcal{Y} \subset \mathcal{M}^d$, where $\mathcal{M}$ is embedded in $\mathbb{R}^p$ by ι. Suppose X is a smooth vector field that we observe only on $\mathcal{X}$ and want to extend to $\mathcal{Y}$. That is, we observe

(a) $d_{\mathrm{VDM'},t=10}$ (b) $d_{\mathrm{VDM'},t=100}$ (c) $d_{\mathrm{VDM'},t=1000}$

(d) $d_{\mathrm{DM},t=10}$ (e) $d_{\mathrm{DM},t=100}$ (f) $d_{\mathrm{DM},t=1000}$ (g) Geodesic distance
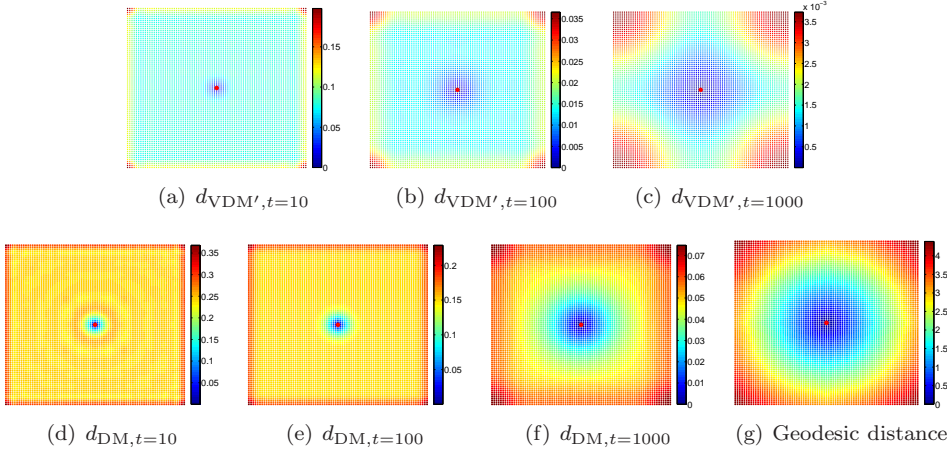
FIG. 6.5. *Square case. Top: truncated vector diffusion distances for $t = 10$, $t = 100$ and $t = 1000$; Bottom: truncated diffusion distances for $t = 10$, $t = 100$ and $t = 1000$, and the geodesic distance. The reference point from which distances are computed is marked in red.*

the vectors $\iota_* X(x_1), \ldots, \iota_* X(x_n) \in \mathbb{R}^p$ and want to estimate $\iota_* X(y_1), \ldots, \iota_* X(y_m)$. The set $\mathcal{X}$ is assumed to be fixed, while the points in $\mathcal{Y}$ may arrive on-the-fly and need to be processed in real time. We propose the following Nyström scheme for extending X from $\mathcal{X}$ to $\mathcal{Y}$.

In the preprocessing step we use the points $x_1, \ldots, x_n$ for local PCA, alignment and vector diffusion mapping as described in Sections 2 and 3. That is, using local PCA, we find the $p \times d$ matrices O_i ($i = 1, \ldots, n$), such that the columns of O_i are an orthonormal basis for a subspace that approximates the embedded tangent plane $\iota_* T_{x_i}\mathcal{M}$; using alignment we find the orthonormal $d \times d$ matrices O_{ij} that approximate the parallel transport operator from $T_{x_j}\mathcal{M}$ to $T_{x_i}\mathcal{M}$; and using w_{ij} and O_{ij} we construct the matrices S and D and compute (a subset of) the eigenvectors $v_1, v_2, \ldots, v_{nd}$ and eigenvalues $\lambda_1, \ldots, \lambda_{nd}$ of $D^{-1}S$.

We project the embedded vector field $\iota_* X(x_i) \in \mathbb{R}^p$ into the d-dimensional subspace spanned by the columns of O_i, and define $X_i \in \mathbb{R}^d$ as

$$X_i = O_i^T \iota_* X(x_i). \tag{7.1}$$

We represent the vector field X on $\mathcal{X}$ by the vector $\boldsymbol{x}$ of length nd, organized as n vectors of length d, with

$$\boldsymbol{x}(i) = X_i, \quad \text{for } i = 1, \ldots, n.$$

We use the orthonormal basis of eigen-vector-fields $v_1, \ldots, v_{nd}$ to decompose $\boldsymbol{x}$ as

$$\boldsymbol{x} = \sum_{l=1}^{nd} a_l v_l, \tag{7.2}$$

where $a_l = \boldsymbol{x}^T v_l$. This concludes the preprocessing computations.

Suppose $y \in \mathcal{Y}$ is a "new" out-of-sample point. First, we perform the local PCA step to find a $p \times d$ matrix, denoted O_y, whose columns form an orthonormal basis to a d-dimensional subspace of $\mathbb{R}^p$ that approximates the embedded tangent plane

19

$\iota_* T_y \mathcal{M}$. The local PCA step uses only the neighbors of y among the points in $\mathcal{X}$ (but not in $\mathcal{Y}$) inside a ball of radius $\sqrt{\epsilon_{PCA}}$ centered at y.

Next, we use the alignment process to compute the $d \times d$ orthonormal matrix $O_{y,i}$ between x_i and y by setting

$$O_{y,i} = \underset{O \in O(d)}{\operatorname{argmin}} \|O_y^T O_i - O\|_{HS}.$$

Notice that the eigen-vector-fields satisfy

$$v_l(i) = \frac{1}{\lambda_l} \frac{\sum_{j=1}^n K_\epsilon(\|x_i - x_j\|)O_{ij}v_l(j)}{\sum_{j=1}^n K_\epsilon(\|x_i - x_j\|)}.$$

We denote the extension of v_l to the point y by $\tilde{v}_l(y)$ and define it as

$$\tilde{v}_l(y) = \frac{1}{\lambda_l} \frac{\sum_{j=1}^n K_\epsilon(\|y - x_j\|)O_{y,j}v_l(j)}{\sum_{j=1}^n K_\epsilon(\|y - x_j\|)}. \tag{7.3}$$

To finish the extrapolation problem, we denote the extension of $\boldsymbol{x}$ to y by $\tilde{\boldsymbol{x}}(y)$ and define it as

$$\tilde{\boldsymbol{x}}(y) = \sum_{l=1}^{m(\delta)} a_l \tilde{v}_l(y), \tag{7.4}$$

where $m(\delta) = \max_l |\lambda_l| > \delta$, and $\delta > 0$ is some fixed parameter to ensure the numerical stability of the extension procedure (due to the division by λ_l in (7.3), $\frac{1}{\delta}$ can be regarded as the condition number of the extension procedure). The vector $\iota_* X(y) \in \mathbb{R}^p$ is estimated as

$$\iota_* X(y) = O_y \tilde{\boldsymbol{x}}(y). \tag{7.5}$$

8. The continuous case: heat kernels. As discussed earlier, in the limit $n \to \infty$ and $\epsilon \to 0$ considered in (5.2), the normalized graph Laplacian converges to the Laplace-Beltrami operator, which is the generator of the heat kernel for functions (0-forms). Similarly, in the limit $n \to \infty$ considered in (5.3), we get the connection Laplacian operator, which is the generator of a heat kernel for vector fields (or 1-forms). The connection Laplacian ∇^2 is a self-adjoint, second order elliptic operator defined over the tangent bundle $T\mathcal{M}$. It is well-known [16] that the spectrum of ∇^2 is discrete inside $\mathbb{R}^-$ and the only possible accumulation point is $-\infty$. We will denote the spectrum as $\{-\lambda_k\}_{k=0}^\infty$, where $0 \le \lambda_0 \le \lambda_1....$ From the classical elliptic theory, see for example [16], we know that $e^{t\nabla^2}$ has the kernel

$$k_t(x,y) = \sum_{n=0}^\infty e^{-\lambda_n t} X_n(x) \otimes \overline{X_n(y)}.$$

where $\nabla^2 X_n = -\lambda_n X_n$. Also, the eigenvector-fields X_n of ∇^2 form an orthonormal basis of $L^2(T\mathcal{M})$. In the continuous setup, we define the vector diffusion distance between $x, y \in \mathcal{M}$ using $\|k_t(x,y)\|_{HS}^2$. An explicit calculation gives

$$\|k_t(x,y)\|_{HS}^2 = \operatorname{Tr}\left[k_t(x,y)k_t(x,y)^*\right]$$
$$= \sum_{n,m=0}^\infty e^{-(\lambda_n + \lambda_m)t} \langle X_n(x), X_m(x) \rangle \overline{\langle X_n(y), X_m(y) \rangle}. \tag{8.1}$$

20

It is well known that the heat kernel $k_t(x, y)$ is smooth in x and y and analytic in t [16], so for $t > 0$ we can define a family of vector diffusion mappings V_t, that map any $x \in \mathcal{M}$ into the Hilbert space ℓ^2 by:

$$V_t : x \mapsto \left(e^{-(\lambda_n + \lambda_m)t/2} \langle X_n(x), X_m(x) \rangle \right)_{n,m=0}^{\infty}, \tag{8.2}$$

which satisfies

$$\|k_t(x, y)\|_{HS}^2 = \langle V_t(x), V_t(y) \rangle_{\ell^2}. \tag{8.3}$$

The vector diffusion distance $d_{\mathrm{VDM},t}(x, y)$ between $x \in M$ and $y \in M$ is defined as

$$d_{\mathrm{VDM},t}(x, y) := \|V_t(x) - V_t(y)\|_{\ell^2}, \tag{8.4}$$

which is clearly a distance function over $\mathcal{M}$. In practice, due to the decay of $e^{-(\lambda_n + \lambda_m)t}$, only pairs (n, m) for which $\lambda_n + \lambda_m$ is not too large are needed to get a good approximation of this vector diffusion distance. Like in the discrete case, the dot products $\langle X_n(x), X_m(x) \rangle$ are invariant to the choice of basis for the tangent space at x.

We now study some properties of the vector diffusion map V_t (8.2). First, we claim for all $t > 0$, the vector diffusion mapping V_t is an embedding of the compact Riemannian manifold $\mathcal{M}$ into ℓ^2.

THEOREM 8.1. *Given a d-dim closed Riemannian manifold $(\mathcal{M}, g)$ and an orthonormal basis $\{X_n\}_{n=0}^{\infty}$ of $L^2(T\mathcal{M})$ composed of the eigen-vector-fields of the connection-Laplacian ∇^2, then for any $t > 0$, the vector diffusion map V_t is a diffeomorphic embedding of $\mathcal{M}$ into ℓ^2.*

Proof. We show that $V_t : \mathcal{M} \to \ell^2$ is continuous in x by noting that

$$\|V_t(x) - V_t(y)\|_{\ell^2}^2 = \sum_{n,m=0}^{\infty} e^{-(\lambda_n + \lambda_m)t} (\langle X_n(x), X_m(x) \rangle - \langle X_n(y), X_m(y) \rangle)^2$$

$$= \mathrm{Tr}(k_t(x, x)k_t(x, x)^*) + \mathrm{Tr}(k_t(y, y)k_t(y, y)^*) - 2\,\mathrm{Tr}(k_t(x, y)k_t(x, y)^*) \tag{8.5}$$

From the continuity of the kernel $k_t(x, y)$, it is clear that $\|V_t(x) - V_t(y)\|_{\ell^2}^2 \to 0$ as $y \to x$. Since $\mathcal{M}$ is compact, it follows that $V_t(\mathcal{M})$ is compact in ℓ^2. Then we show that V_t is one-to-one. Fix $x \neq y$ and a smooth vector field X that satisfies $\langle X(x), X(x) \rangle \neq \langle X(y), X(y) \rangle$. Since the eigen-vector fields $\{X_n\}_{n=0}^{\infty}$ form a basis to $L^2(T\mathcal{M})$, we have

$$X(z) = \sum_{n=0}^{\infty} c_n X_n(z), \quad \text{for all } z \in \mathcal{M},$$

where $c_n = \int_{\mathcal{M}} \langle X, X_n \rangle dV$. As a result,

$$\langle X(z), X(z) \rangle = \sum_{n,m=0}^{\infty} c_n c_m \langle X_n(z), X_m(z) \rangle.$$

Since $\langle X(x), X(x) \rangle \neq \langle X(y), X(y) \rangle$, there exist $n, m \in \mathbb{N}$ such that $\langle X_n(x), X_m(x) \rangle \neq \langle X_n(y), X_m(y) \rangle$, which shows that $V_t(x) \neq V_t(y)$, i.e., V_t is one-to-one. From the fact

21

that the map V_t is continuous and one-to-one from $\mathcal{M}$, which is compact, onto $V_t(\mathcal{M})$, we conclude that V_t is an embedding.

□

Next, we demonstrate the asymptotic behavior of the vector diffusion distance $d_{\mathrm{VDM},t}(x,y)$ and the diffusion distance $d_{\mathrm{DM},t}(x,y)$ when t is small and x is close to y. The following theorem shows that in this asymptotic limit both the vector diffusion distance and the diffusion distance behave like the geodesic distance.

THEOREM 8.2. *Let* $(\mathcal{M}, g)$ *be a smooth* d*-dim closed Riemannian manifold. Suppose* $x, y \in \mathcal{M}$ *so that* $x = \exp_y v$, *where* $v \in T_y\mathcal{M}$. *For any* $t > 0$, *when* $\|v\|^2 \ll t \ll 1$ *we have the following asymptotic expansion of the vector diffusion distance:*

$$d^2_{VDM,t}(x,y) = d(4\pi)^{-d}\frac{\|v\|^2}{t^{d+1}} + O(t^{-d})$$

Similarly, when $\|v\|^2 \ll t \ll 1$, *we have the following asymptotic expansion of the diffusion distance:*

$$d^2_{DM,t}(x,y) = (4\pi)^{-d/2}\frac{\|v\|^2}{2t^{d/2+1}} + O(t^{-d/2}).$$

Proof. Fix y and a normal coordinate around y. Denote $j(x,y) = |\det(d_v \exp_y)|$, where $x = \exp_y(v)$, $v \in T_x\mathcal{M}$. Suppose $\|v\|$ is small enough so that $x = \exp_y(v)$ is away from the cut locus of y. It is well known that the heat kernel $k_t(x,y)$ for the connection Laplacian ∇^2 over the vector bundle $\mathcal{E}$ possesses the following asymptotic expansion when x and y are close: [5, p. 84] or [11]

$$\|\partial_t^k(k_t(x,y) - k_t^N(x,y))\|_l = O(t^{N-d/2-l/2-k}), \tag{8.6}$$

where $\|\cdot\|_l$ is the C^l norm,

$$k_t^N(x,y) := (4\pi t)^{-d/2}e^{-\|v\|^2/4t}j(x,y)^{-1/2}\sum_{i=0}^{N} t^i\Phi_i(x,y), \tag{8.7}$$

$N > d/2$, and Φ_i is a smooth section of the vector bundle $\mathcal{E} \otimes \mathcal{E}^*$ over $\mathcal{M} \times \mathcal{M}$. Moreover, $\Phi_0(x,y) = P_{x,y}$ is the parallel transport from $\mathcal{E}_y$ to $\mathcal{E}_x$. In the VDM setup, we take $\mathcal{E} = T\mathcal{M}$, the tangent bundle of $\mathcal{M}$. Also, by [5, Proposition 1.28], we have the following expansion:

$$j(x,y) = 1 + \mathrm{Ric}(v,v)/6 + O(\|v\|^3). \tag{8.8}$$

Equations (8.7) and (8.8) lead to the following expansion under the assumption $\|v\|^2 \ll t$:

$$\mathrm{Tr}(k_t(x,y)k_t(x,y)^*)$$
$$= (4\pi t)^{-d}e^{-\|v\|^2/2t}(1 + \mathrm{Ric}(v,v)/6 + O(\|v\|^3))^{-1}\mathrm{Tr}((P_{x,y} + O(t))((P_{x,y} + O(t))^*)$$
$$= (4\pi t)^{-d}e^{-\|v\|^2/2t}(1 - \mathrm{Ric}(v,v)/6 + O(\|v\|^3))(d + O(t))$$
$$= (d + O(t))(4\pi t)^{-d}\left(1 - \frac{\|v\|^2}{2t} + O\left(\frac{\|v\|^4}{t^2}\right)\right).$$

In particular, for $\|v\| = 0$ we have

$$\mathrm{Tr}(k_t(x,x)k_t(x,x)^*) = (d + O(t))(4\pi t)^{-d}.$$

22

Thus, for $\|v\|^2 \ll t \ll 1$, we have

$$
\begin{aligned}
d_{\mathrm{VDM},t}^2(x,y) &= \mathrm{Tr}(k_t(x,x)k_t(x,x)^*) + \mathrm{Tr}(k_t(y,y)k_t(y,y)^*) - 2\,\mathrm{Tr}(k_t(x,y)k_t(x,y)^*) \\
&= d(4\pi)^{-d}\frac{\|v\|^2}{t^{d+1}} + O(t^{-d}).
\end{aligned} \tag{8.9}
$$

By the same argument we can carry out the asymptotic expansion of the diffusion distance $d_{\mathrm{DM},t}(x,y)$. Denote the eigenfunctions and eigenvalues of the Laplace-Beltrami operator Δ by ϕ_n and μ_n. We can rewrite the diffusion distance as follows:

$$
d_{\mathrm{DM},t}^2(x,y) = \sum_{n=1}^{\infty} e^{-\mu_n t}(\phi_n(x) - \phi_n(y))^2 = \tilde{k}_t(x,x) + \tilde{k}_t(y,y) - 2\tilde{k}_t(x,y), \tag{8.10}
$$

where $\tilde{k}_t$ is the heat kernel of the Laplace-Beltrami operator. Note that the Laplace-Beltrami operator is equal to the connection-Laplacian operator defined over the trivial line bundle over $\mathcal{M}$. As a result, equation (8.7) also describes the asymptotic expansion of the heat kernel for the Laplace-Beltrami operator as

$$
\tilde{k}_t(x,y) = (4\pi t)^{-d/2}e^{-\|v\|^2/4t}(1 + \mathrm{Ric}(v,v)/6 + O(\|v\|^3))^{-1/2}(1 + O(t)).
$$

Put these facts together, we obtain

$$
d_{\mathrm{DM},t}^2(x,y) = (4\pi)^{-d/2}\frac{\|v\|^2}{2t^{d/2+1}} + O(t^{-d/2}), \tag{8.11}
$$

when $\|v\|^2 \ll t \ll 1$. $\square$

9. Application of VDM to Cryo-Electron Microscopy. Besides being a general framework for data analysis and manifold learning, VDM is useful for performing robust multi-reference rotational alignment of objects, such as one-dimensional periodic signals, two-dimensional images and three-dimensional shapes. In this Section, we briefly describe the application of VDM to a particular multi-reference rotational alignment problem of two-dimensional images that arise in the field of cryo-electron microscopy (EM). A more comprehensive study of this problem can be found in [36] and [19]. It can be regarded as a prototypical multi-reference alignment problem, and we expect many other multi-reference alignment problems that arise in areas such as computer vision and computer graphics to benefit from the proposed approach.

The goal in cryo-EM [14] is to determine 3D macromolecular structures from noisy projection images taken at unknown random orientations by an electron microscope, i.e., a random Computational Tomography (CT). Determining 3D macromolecular structures for large biological molecules remains vitally important, as witnessed, for example, by the 2003 Chemistry Nobel Prize, co-awarded to R. MacKinnon for resolving the 3D structure of the Shaker K^+ channel protein, and by the 2009 Chemistry Nobel Prize, awarded to V. Ramakrishnan, T. Steitz and A. Yonath for studies of the structure and function of the ribosome. The standard procedure for structure determination of large molecules is X-ray crystallography. The challenge in this method is often more in the crystallization itself than in the interpretation of the X-ray results, since many large proteins have so far withstood all attempts to crystallize them.

In cryo-EM, an alternative to X-ray crystallography, the sample of macromolecules is rapidly frozen in an ice layer so thin that their tomographic projections are typically disjoint; this seems the most promising alternative for large molecules that defy
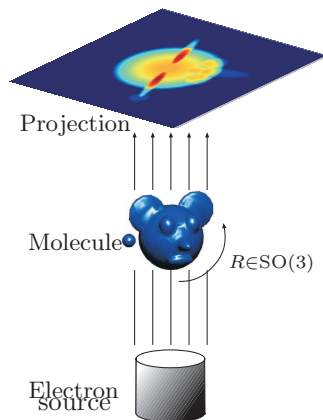
23

FIG. 9.1. *Schematic drawing of the imaging process: every projection image corresponds to some unknown 3D rotation of the unknown molecule.*

crystallization. The cryo-EM imaging process produces a large collection of tomographic projections of the same molecule, corresponding to different and unknown projection orientations. The goal is to reconstruct the three-dimensional structure of the molecule from such unlabeled projection images, where data sets typically range from 10^4 to 10^5 projection images whose size is roughly 100×100 pixels. The intensity of the pixels in a given projection image is proportional to the line integrals of the electric potential induced by the molecule along the path of the imaging electrons (see Figure 9.1). The highly intense electron beam destroys the frozen molecule and it is therefore impractical to take projection images of the same molecule at known different directions as in the case of classical CT. In other words, a single molecule can be imaged only once, rendering an extremely low signal-to-noise ratio (SNR) for the images (see Figure 9.2 for a sample of real microscope images), mostly due to shot noise induced by the maximal allowed electron dose (other sources of noise include the varying width of the ice layer and partial knowledge of the contrast function of the microscope). In the basic homogeneity setting considered hereafter, all imaged molecules are assumed to have the exact same structure; they differ only by their spatial rotation. Every image is a projection of the same molecule but at an unknown random three-dimensional rotation, and the cryo-EM problem is to find the three-dimensional structure of the molecule from a collection of noisy projection images.
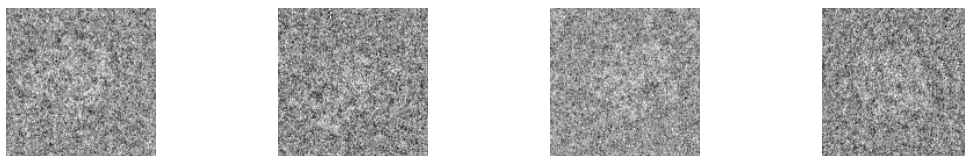

FIG. 9.2. *A collection of four real electron microscope images of the E. coli 50S ribosomal subunit; courtesy of Dr. Fred Sigworth.*

The rotation group SO(3) is the group of all orientation preserving orthogonal transformations about the origin of the three-dimensional Euclidean space $\mathbb{R}^3$ under the operation of composition. Any 3D rotation can be expressed using a 3×3 orthog-

onal matrix $R = \begin{pmatrix} | & | & | \\ R^1 & R^2 & R^3 \\ | & | & | \end{pmatrix}$ satisfying $RR^T = R^T R = I_{3\times3}$ and $\det R = 1$.

The column vectors R^1, R^2, R^3 of R form an orthonormal basis to $\mathbb{R}^3$. To each projection image P there corresponds a 3×3 unknown rotation matrix R describing its orientation (see Figure 9.1). Excluding the contribution of noise, the intensity $P(x, y)$ of the pixel located at (x, y) in the image plane corresponds to the line integral of the electric potential induced by the molecule along the path of the imaging electrons, that is,

$$P(x, y) = \int_{-\infty}^{\infty} \phi(xR^1 + yR^2 + zR^3)\,dz \tag{9.1}$$

where $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is the electric potential of the molecule in some fixed 'laboratory' coordinate system. The projection operator (9.1) is also known as the X-ray transform [29].

We therefore identify the third column R^3 of R as the imaging direction, also known as the *viewing angle* of the molecule. The first two columns R^1 and R^2 form an orthonormal basis for the plane in $\mathbb{R}^3$ perpendicular to the viewing angle R^3. All clean projection images of the molecule that share the same viewing angle look the same up to some in-plane rotation. That is, if R_i and R_j are two rotations with the same viewing angle $R_i^3 = R_j^3$ then R_i^1, R_i^2 and R_j^1, R_j^2 are two orthonormal bases for the same plane. On the other hand, two rotations with opposite viewing angles $R_i^3 = -R_j^3$ give rise to two projection images that are the same after reflection (mirroring) and some in-plane rotation.

As projection images in cryo-EM have extremely low SNR, a crucial initial step in all reconstruction methods is "class averaging" [14, 41]. Class averaging is the grouping of a large data set of n noisy raw projection images $P_1, \ldots, P_n$ into clusters, such that images within a single cluster have similar viewing angles (it is possible to artificially double the number of projection images by including all mirrored images). Averaging rotationally-aligned noisy images within each cluster results in "class averages"; these are images that enjoy a higher SNR and are used in later cryo-EM procedures such as the angular reconstitution procedure [40] that requires better quality images. Finding consistent class averages is challenging due to the high level of noise in the raw images as well as the large size of the image data set. A sketch of the class averaging procedure is shown in Figure 9.3.

Penczek, Zhu and Frank [31] introduced the rotationally invariant K-means clustering procedure to identify images that have similar viewing angles. Their Rotationally Invariant Distance $d_{\text{RID}}(i, j)$ between image P_i and image P_j is defined as the Euclidean distance between the images when they are optimally aligned with respect to in-plane rotations (assuming the images are centered)

$$d_{\text{RID}}(i, j) = \min_{\theta \in [0, 2\pi)} \|P_i - R(\theta)P_j\|, \tag{9.2}$$

where $R(\theta)$ is the rotation operator of an image by an angle θ in the counterclockwise direction. Prior to computing the invariant distances of (9.2), a common practice is to center all images by correlating them with their total average $\frac{1}{n}\sum_{i=1}^{n} P_i$, which is approximately radial (i.e., has little angular variation) due to the randomness in the rotations. The resulting centers usually miss the true centers by only a few pixels (as can be validated in simulations during the refinement procedure). Therefore, like [31],
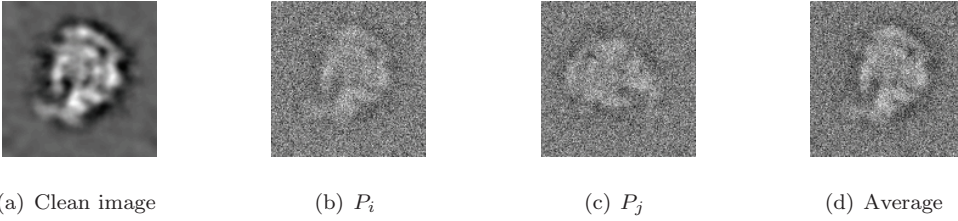
| (a) Clean image | (b) P_i | (c) P_j | (d) Average |

FIG. 9.3. *(a) A clean simulated projection image of the ribosomal subunit generated from its known volume; (b) Noisy instance of (a), denoted P_i, obtained by the addition of white Gaussian noise. For the simulated images we chose the SNR to be higher than that of experimental images in order for image features to be clearly visible; (c) Noisy projection, denoted P_j, taken at the same viewing angle but with a different in-plane rotation; (d) Averaging the noisy images (b) and (c) after in-plane rotational alignment. The class average of the two images has a higher SNR than that of the noisy images (b) and (c), and it has better similarity with the clean image (a).*

we also choose to focus on the more challenging problem of rotational alignment by assuming that the images are properly centered, while the problem of translational alignment can be solved later by solving an overdetermined linear system.

It is worth noting that the specific choice of metric to measure proximity between images can make a big difference in class averaging. The cross-correlation or Euclidean distance (9.2) are by no means optimal measures of proximity. In practice, it is common to denoise the images prior to computing their pairwise distances. Although the discussion which follows is independent of the particular choice of filter or distance metric, we emphasize that filtering can have a dramatic effect on finding meaningful class averages.

The invariant distance between noisy images that share the same viewing angle (with perhaps a different in-plane rotation) is expected to be small. Ideally, all neighboring images of some reference image P_i in a small invariant distance ball centered at P_i should have similar viewing angles, and averaging such neighboring images (after proper rotational alignment) would amplify the signal and diminish the noise.

Unfortunately, due to the low SNR, it often happens that two images of completely different viewing angles have a small invariant distance. This can happen when the realizations of the noise in the two images match well for some random in-plane rotational angle, leading to spurious neighbor identification. Therefore, averaging the nearest neighbor images can sometimes yield a poor estimate of the true signal in the reference image.

The histograms of Figure 9.5 demonstrate the ability of small rotationally invariant distances to identify images with similar viewing directions. For each image we use the rotationally invariant distances to find its 40 nearest neighbors among the entire set of $n = 40,000$ images. In our simulation we know the original viewing directions, so for each image we compute the angles (in degrees) between the viewing direction of the image and the viewing directions of its 40 neighbors. Small angles indicate successful identification of "true" neighbors that belong to a small spherical cap, while large angles correspond to outliers. We see that for SNR=1/2 there are no outliers, and all the viewing directions of the neighbors belong to a spherical cap whose opening angle is about $8°$. However, for lower values of the SNR, there are outliers, indicated by arbitrarily large angles (all the way to $180°$).

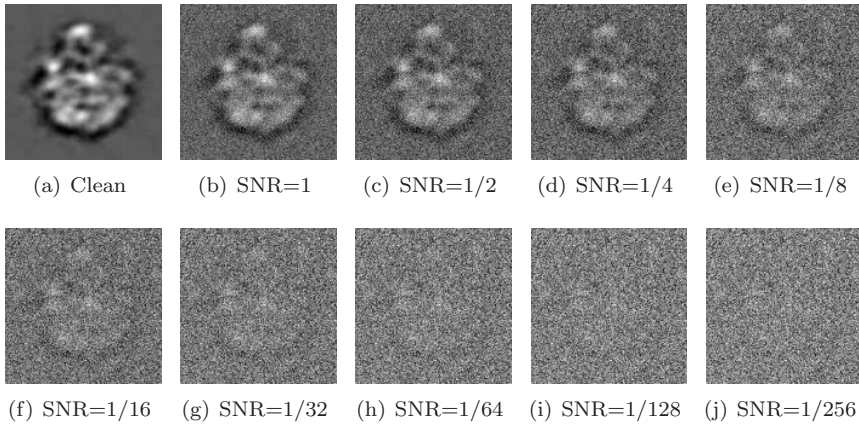Clustering algorithms, such as the K-means algorithm, perform much better than

26

(a) Clean   (b) SNR=1   (c) SNR=1/2   (d) SNR=1/4   (e) SNR=1/8

(f) SNR=1/16   (g) SNR=1/32   (h) SNR=1/64   (i) SNR=1/128   (j) SNR=1/256

FIG. 9.4. *Simulated projection with various levels of additive Gaussian white noise.*

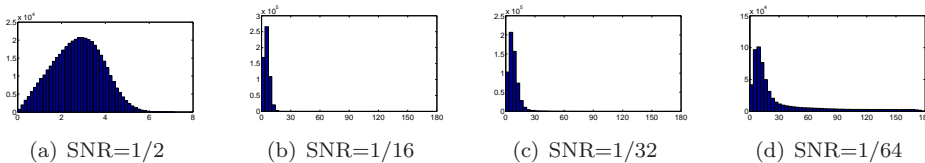(a) SNR=1/2   (b) SNR=1/16   (c) SNR=1/32   (d) SNR=1/64

FIG. 9.5. *Histograms of the angle (in degrees, x-axis) between the viewing directions of 40,000 images and the viewing directions of their 40 nearest neighboring images as found by computing the rotationally invariant distances.*

naïve nearest neighbors averaging, because they take into account all pairwise distances, not just distances to the reference image. Such clustering procedures are based on the philosophy that images that share a similar viewing angle with the reference image are expected to have a small invariant distance not only to the reference image but also to all other images with similar viewing angles. This observation was utilized in the rotationally invariant K-means clustering algorithm [31]. Still, due to noise, the rotationally invariant K-means clustering algorithm may suffer from misidentifications at the low SNR values present in experimental data.

VDM is a natural algorithmic framework for the class averaging problem, as it can further improve the detection of neighboring images even at lower SNR values. The rotationally invariant distance neglects an important piece of information, namely, the optimal angle that realizes the best rotational alignment in (9.2):

$$\theta_{ij} = \operatorname*{argmin}_{\theta \in [0, 2\pi)} \|P_i - R(\theta)P_j\|, \quad i, j = 1, \ldots, n. \tag{9.3}$$

In VDM, we use the optimal in-plane rotation angles θ_{ij} to define the orthogonal transformations O_{ij} and to construct the matrix S in (3.1). The eigenvectors and eigenvalues of $D^{-1}S$ (other normalizations of S are also possible) are then used to define the vector diffusion distances between images.

This VDM based classification method is proven to be quite powerful in practice. We applied it to a set of $n = 40,000$ noisy images with SNR=1/64. For every image we find the 40 nearest neighbors using the vector diffusion metric. In the simulation we know the viewing directions of the images, and we compute for each pair of neighbors

27

the angle (in degrees) between their viewing directions. The histogram of these angles is shown in Figure 9.6 (Left panel). About 92% of the identified images belong to a small spherical cap of opening angle 20°, whereas this percentage is only about 65% when neighbors are identified by the rotationally invariant distances (Right panel). We remark that for SNR=1/50, the percentage of correctly identified images by the VDM method goes up to about 98%.

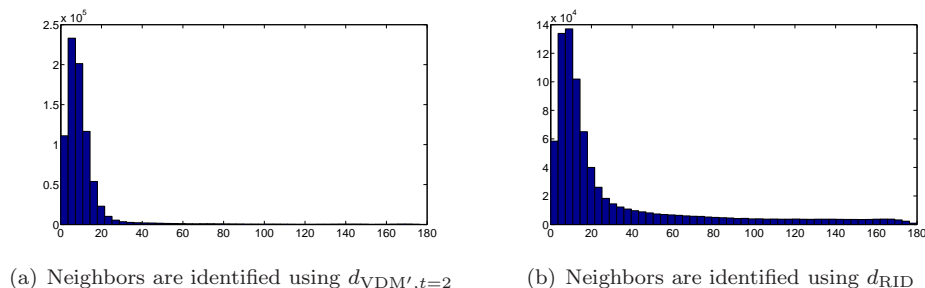(a) Neighbors are identified using $d_{\mathrm{VDM'},t=2}$

(b) Neighbors are identified using d_{RID}

FIG. 9.6. *SNR=1/64: Histogram of the angles (x-axis, in degrees) between the viewing directions of each image (out of 40000) and it 40 neighboring images. Left: neighbors are post identified using vector diffusion distances. Right: neighbors are identified using the original rotationally invariant distances d_{RID}.*

The main advantage of the algorithm presented here is that it successfully identifies images with similar viewing angles even in the presence of a large number of spurious neighbors, that is, even when many pairs of images with viewing angles that are far apart have relatively small rotationally invariant distances. In other words, the VDM-based algorithm is shown to be robust to outliers.

10. Summary and Discussion. This paper introduced vector diffusion maps, an algorithmic and mathematical framework for analyzing data sets where scalar affinities between data points are accompanied with orthogonal transformations. The consistency among the orthogonal transformations along different paths that connect any fixed pair of data points is used to define an affinity between them. We showed that this affinity is equivalent to an inner product, giving rise to the embedding of the data points in a Hilbert space and to the definition of distances between data points, to which we referred as vector diffusion distances.

For data sets of images, the orthogonal transformations and the scalar affinities are naturally obtained via the procedure of optimal registration. The registration process seeks to find the optimal alignment of two images over some class of transformations (also known as deformations), such as rotations, reflections, translations and dilations. For the purpose of vector diffusion mapping, we extract from the optimal deformation only the corresponding orthogonal transformation (rotation and reflection). We demonstrated the usefulness of the vector diffusion map framework in the organization of noisy cryo-electron microscopy images, an important step towards resolving three-dimensional structures of macromolecules. Optimal registration is often used in various mainstream problems in computer vision and computer graphics, for example, in optimal matching of three-dimensional shapes. We therefore expect the vector diffusion map framework to become a useful tool in such applications.

In the case of manifold learning, where the data set is a collection of points in a high dimensional Euclidean space, but with a low dimensional Riemannian manifold structure, we detailed the construction of the orthogonal transformations via the

28

optimal alignment of the orthonormal bases of the tangent spaces. These bases are found using the classical procedure of PCA. Under certain mild conditions about the sampling process of the manifold, we proved that the orthogonal transformation obtained by the alignment procedure approximates the parallel transport operator between the tangent spaces. The proof required careful analysis of the local PCA step which we believe is interesting of it own. Furthermore, we proved that if the manifold is sampled uniformly, then the matrix that lies at the heart of the vector diffusion map framework approximates the connection-Laplacian operator. Following spectral graph theory terminology, we call that matrix the connection-Laplacian of the graph. Using different normalizations of the matrix we proved convergence to the connection-Laplacian operator also for the case of non-uniform sampling. We showed that the vector diffusion mapping is an embedding and proved its relation with the geodesic distance using the asymptotic expansion of the heat kernel for vector fields. These results provide the mathematical foundation for the algorithmic framework that underlies the vector diffusion mapping.

We expect many possible extensions and generalizations of the vector diffusion mapping framework. We conclude by mentioning a few of them.

- *The topology of the data.* In [37] we showed how the vector diffusion mapping can determine if a manifold is orientable or non-orientable, and in the latter case to embed its double covering in a Euclidean space. To that end we used the information in the determinant of the optimal orthogonal transformation between bases of nearby tangent spaces. In other words, we used just the optimal reflection between two orthonormal bases. This simple example shows that vector diffusion mapping can be used to extract topological information from the point cloud. We expect more topological information can be extracted using appropriate modifications of the vector diffusion mapping.

- *Hodge and higher order Laplacians.* Using tensor products of the optimal orthogonal transformations it is possible to construct higher order connection-Laplacians that act on p-forms ($p \geq 1$). The index theorem [16] relates topological structure with geometrical structure. For example, the so-called Betti numbers are related to the multiplicities of the harmonic p-forms of the Hodge Laplacian. For the extraction of topological information it would therefore be useful to modify our construction in order to approximate the Hodge Laplacian instead of the connection-Laplacian.

- *Multiscale, sparse and robust PCA.* In the manifold learning case, an important step of our algorithm is local PCA for estimating the bases for tangent spaces at different data points. In the description of the algorithm, a single scale parameter ϵ_{PCA} is used for all data points. It is conceivable that a better estimation can be obtained by choosing a different, location-dependent scale parameter. A better estimation of the tangent space $T_{x_i}\mathcal{M}$ may be obtained by using a location-dependent scale parameter $\epsilon_{\mathrm{PCA},i}$ due to several reasons: non-uniform sampling of the manifold, varying curvature of the manifold, and global effects such as different pieces of the manifold that are almost touching at some points (i.e., varying "condition number" of the manifold). Choosing the correct scale $\epsilon_{\mathrm{PCA},i}$ is a problem of its own interest that was recently considered in [28], where a multiscale approach was taken to resolve the optimal scale. We recommend the incorporation of such multiscale PCA approaches into the vector diffusion mapping framework. Another difficulty that we may face when dealing with real-life data sets is that the underlying assumption

29

about the data points being located exactly on a low-dimensional manifold does not necessarily hold. In practice, the data points are expected to reside off the manifold, either due to measurement noise or due to the imperfection of the low-dimensional manifold model assumption. It is therefore necessary to estimate the tangent spaces in the presence of noise. Noise is a limiting factor for successful estimation of the tangent space, especially when the data set is embedded in a high dimensional space and noise effects all coordinates [23]. We expect recent methods for robust PCA [7] and sparse PCA [6, 24] to improve the estimation of the tangent spaces and as a result to become useful in the vector diffusion map framework.

- *Random matrix theory and noise sensitivity.* The matrix S that lies at the heart of the vector diffusion map is a block matrix whose blocks are either $d \times d$ orthogonal matrices O_{ij} or the zero blocks. We anticipate that for some applications the measurement of O_{ij} would be imprecise and noisy. In such cases, the matrix S can be viewed as a random matrix and we expect tools from random matrix theory to be useful in analyzing the noise sensitivity of its eigenvectors and eigenvalues. The noise model may also allow for outliers, for example, orthogonal matrices that are uniformly distributed over the orthogonal group $O(d)$ (according to the Haar measure). Notice that the expected value of such random orthogonal matrices is zero, which leads to robustness of the eigenvectors and eigenvalues even in the presence of large number of outliers (see, for example, the random matrix theory analysis in [34]).

- *Compact and non-compact groups and their matrix representation.* As mentioned earlier, the vector diffusion mapping is a natural framework to organize data sets for which the affinities and transformations are obtained from an optimal alignment process over some class of transformations (deformations). In this paper we focused on utilizing orthogonal transformations. At this point the reader have probably asked herself the following question: Is the method limited to orthogonal transformations, or is it possible to utilize other groups of transformations such as translations, dilations, and more? We note that the orthogonal group $O(d)$ is a compact group that has a matrix representation and remark that the vector diffusion mapping framework can be extended to such groups of transformations without much difficulty. However, the extension to non-compact groups, such as the Euclidean group of rigid transformation, the general linear group of invertible matrices and the special linear group is less obvious. Such groups arise naturally in various applications, rendering the importance of extending the vector diffusion mapping to the case of non-compact groups.

REFERENCES

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Trans. Patt. Anal. Mach. Intell.*, 9(5):698–700, 1987.

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

[3] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 486–500, 2005.

[4] M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2007.

[5] N. Berline, E. Getzler, and M. Vergne. *Heat kernels and Dirac operators*. Springer, Berlin, 2004.

[6] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[7] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Submitted for publication.

[8] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

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

[10] M. Cucuringu, Y. Lipman, and A. Singer. Sensor network localization by eigenvector synchronization over the Euclidean group. *ACM Transactions on Sensor Networks*. In press.

[11] B. DeWitt. *The global approach to Quantum field theory*. Oxford University Press, USA, 2003.

[12] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5591–5596, 2003.

[13] K. Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, 1955.

[14] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, New York, 2nd edition, 2006.

[15] W. Fulton and J. Harris. *Representation Theory: A First Course*. Springer, New York, 1991.

[16] P. Gilkey. *The Index Theorem and the Heat Equation*. Princeton, 1974.

[17] M. J. Goldberg and S. Kim. Some remarks on diffusion distances. *J. Appl. Math.*, 2010:17, 2010.

[18] R. Hadani and A. Singer. Representation theoretic patterns in three dimensional cryo-electron microscopy I - the intrinsic reconstitution algorithm. *Annals of Mathematics*. accepted for publication.

[19] R. Hadani and A. Singer. Representation theoretic patterns in three dimensional cryo-electron microscopy II - the class averaging problem. Submitted for publication.

[20] M. Hein, J. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 470–485, 2005.

[21] N. J. Higham. Computing the polar decomposition with applications. *SIAM J. Sci. Stat. Comput.*, 7:1160–1174, October 1986.

[22] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[23] I. M. Johnstone. High dimensional statistical inference and random matrices. *arXiv:math/0611589v1*, 2006.

[24] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[25] J. B. Keller. Closest unitary, orthogonal and hermitian operators to a given operator. *Mathematics Magazine*, 48(4):192–197, 1975.

[26] S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.

[27] A. B. Lee, K. S. Pedersen, and D. Mumford. The non-linear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(13):83–103, 2003.

[28] A. V. Little, J. Lee, Y.M. Jung, and M. Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 85–88, 2009.

[29] F. Natterer. *The Mathematics of Computerized Tomography*. Classics in Applied Mathematics.

SIAM: Society for Industrial and Applied Mathematics, 2001.

[30] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. In Richard Pollack, Janos Pach, and Jacob E. Goodman, editors, *Twentieth Anniversary Volume:*, pages 1–23. Springer New York.

[31] P.A. Penczek, J. Zhu, and J. Frank. A common-lines based method for determining orientations for $N \geq 3$ particle projections simultaneously. *Ultramicroscopy*, 63(3):205–218, 1996.

[32] P. Petersen. *Riemannian Geometry*. Springer, New York, 2006.

[33] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[34] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36.

[35] A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

[36] A. Singer, Z. Zhao Y. Shkolnisky, and R. Hadani. Viewing angle classification of cryo-electron microscopy images using eigenvectors. Submitted for publication.

[37] A. Singer and H.-T. Wu. Orientability and diffusion map. *Applied and Computational Harmonic Analysis*, in press.

[38] M. E. Taylor. *Noncommutative Harmonic Analysis*. AMS, 1986.

[39] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[40] M. van Heel. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*, 21(2):111–123, 1987.

[41] M. van Heel, B. Gowen, R. Matadeen, E.V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz, and A. Patwardhan. Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly Reviews of Biophysics*, 33(4):307–369, 2000.

Appendix A. Some Differential Geometry Background. The purpose of this appendix is to provide the required mathematical background for readers who are not familiar with concepts such as the parallel transport operator, connection, and the connection Laplacian. We illustrate these concepts by considering a surface $\mathcal{M}$ embedded in $\mathbb{R}^3$.

Given a function $f(x) : \mathbb{R}^3 \to \mathbb{R}$, its gradient vector field is given by

$$\nabla f := \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right).$$

Through the gradient, we can find the rate of change of f at $x \in \mathbb{R}^3$ in a given direction $v \in \mathbb{R}^3$, using the directional derivative:

$$v f(x) := \lim_{t \to 0} \frac{f(x + tv) - f(v)}{t}.$$

By chain rule we have $vf(x) = \nabla f(x)(v)$. Define $\nabla_v f(x) := \nabla f(x)(v)$.

Let X be a vector field on $\mathbb{R}^3$,

$$X(x, y, z) = (f_1(x, y, z), f_2(x, y, z), f_3(x, y, z)).$$

It is natural to extend the derivative notion to a given vector field X at $x \in \mathbb{R}^3$ by mimicking the derivative definition for functions in the following way:

$$\lim_{t \to 0} \frac{X(x + tv) - X(x)}{t} \tag{A.1}$$

where $v \in \mathbb{R}^3$. Following the same notation for the directional derivative of a function, we denote this limit by $\nabla_v X(x)$. This quantity tells us that at x, following the direction v, we compare the vector field at two points x and $x + tv$, and see how the vector field changes. While this definition looks good at first sight, we now explain

32

that it has certain shortcomings that need to be fixed in order to generalize it to the case of a surface embedded in $\mathbb{R}^3$.

Consider a two dimensional smooth surface $\mathcal{M}$ embedded in $\mathbb{R}^3$ by ι. Fix a point $x \in \mathcal{M}$ and a smooth curve $\gamma(t) : (-\epsilon, \epsilon) \to \mathcal{M} \subset \mathbb{R}^3$, where $\epsilon \ll 1$ and $\gamma(0) = x$. $\gamma'(0) \in \mathbb{R}^3$ is called a tangent vector to $\mathcal{M}$ at x. The 2 dimensional affine space spanned by the collection of all tangent vectors to $\mathcal{M}$ at x is defined to be the tangent plane at x and denoted by[8] $T_x\mathcal{M}$, which is a two dimensional affine space inside $\mathbb{R}^3$, as illustrated in Figure A.1 (left panel). Having defined the tangent plane at each point $x \in \mathcal{M}$, we define a vector field X over $\mathcal{M}$ to be a differentiable map that maps x to a tangent vector in $T_x\mathcal{M}$.[9]
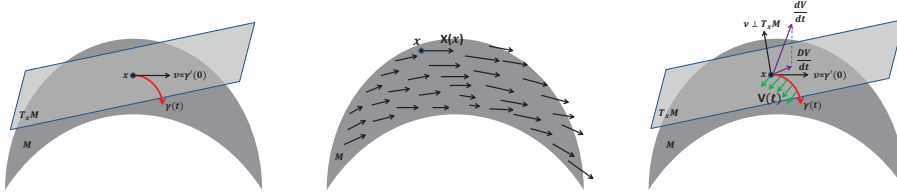

FIG. A.1. *Left: a tangent plane and a curve γ; Middle: a vector field; Right: the covariant derivative*

We now generalize the definition of the derivative of a vector field over $\mathbb{R}^3$ (A.1) to define the derivative of a vector field over $\mathcal{M}$. The first difficulty we face is how to make sense of "$X(x + tv)$", since $x + tv$ does not belong to $\mathcal{M}$. This difficulty can be tackled easily by changing the definition (A.1) a bit by considering the curve $\gamma : (-\epsilon, \epsilon) \to \mathbb{R}^3$ so that $\gamma(0) = x$ and $\gamma'(0) = v$. Thus, (A.1) becomes

$$\lim_{t \to 0} \frac{X(\gamma(t)) - X(\gamma(0))}{t} \tag{A.2}$$

where $v \in \mathbb{R}^3$. In $\mathcal{M}$, the existence of the curve $\gamma : (-\epsilon, \epsilon) \to \mathbb{R}^3$ so that $\gamma(0) = x$ and $\gamma'(0) = v$ is guaranteed by the classical ordinary differential equation theory. However, (A.2) still cannot be generalized to $\mathcal{M}$ directly even though $X(\gamma(t))$ is well defined. The difficulty we face here is how to compare $X(\gamma(t))$ and $X(x)$, that is, how to make sense of the subtraction $X(\gamma(t)) - X(\gamma(0))$. It is not obvious since a priori we do not know how $T_{\gamma(t)}\mathcal{M}$ and $T_{\gamma(0)}\mathcal{M}$ are related. The way we proceed is by defining an important notion in differential geometry called "parallel transport", which plays an essential role in our VDM framework.

Fix a point $x \in \mathcal{M}$ and a vector field X on $\mathcal{M}$, and consider a parametrized curve $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ so that $\gamma(0) = x$. Define a vector valued function $V : (-\epsilon, \epsilon) \to \mathbb{R}^3$ by restricting X to γ, that is, $V(t) = X(\gamma(t))$. The derivative of V is well defined as usual:

$$\frac{\mathrm{d}V}{\mathrm{d}t}(h) := \lim_{t \to 0} \frac{V(h + t) - V(h)}{t},$$

---

[8]Here we abuse notation slightly. Usually $T_x\mathcal{M}$ defined here is understood as the embedded tangent plane by the embedding ι of the tangent plane at x. Please see [32] for a rigorous definition of the tangent plane.

[9]See [32] for the exact notion of differentiability. Here, again, we abuse notation slightly. Usually X defined here is understood as the embedded vector field by the embedding ι of the vector field X. For the rigorous definition of a vector field, please see [32].

where $h \in (-\epsilon, \epsilon)$. The *covariant derivative* $\frac{DV}{dt}(h)$ is defined as the projection of $\frac{dV}{dt}(h)$ onto $T_{\gamma(h)}\mathcal{M}$. Then, using the definition of $\frac{DV}{dt}(h)$, we consider the following equation:

$$\begin{cases} \frac{DW}{dt}(t) = 0 \\ W(0) = w \end{cases}$$

where $w \in T_{\gamma(0)}\mathcal{M}$. The solution $W(t)$ exists by the classical ordinary differential equation theory. The solution $W(t)$ along $\gamma(t)$ is called the *parallel vector field* along the curve $\gamma(t)$, and we also call $W(t)$ the *parallel transport of w along the curve $\gamma(t)$* and denote $W(t) = P_{\gamma(t),\gamma(0)}w$.

We come back to address the initial problem: how to define the "derivative" of a given vector field over a surface $\mathcal{M}$. We define the *covariant derivative* of a given vector field X over $\mathcal{M}$ as follows:

$$\nabla_v X(x) = \lim_{t \to 0} \frac{P_{\gamma(0),\gamma(t)} X(\gamma(t)) - X(\gamma(0))}{t}, \tag{A.3}$$

where $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ with $\gamma(0) = x \in \mathcal{M}$, $\gamma'(0) = v \in T_{\gamma(0)}\mathcal{M}$. This definition says that if we want to analyze how a given vector field at $x \in \mathcal{M}$ changes along the direction v, we choose a curve γ so that $\gamma(0) = x$ and $\gamma'(0) = v$, and then "transport" the vector field value at point $\gamma(t)$ to $\gamma(0) = x$ so that the comparison of the two tangent planes makes sense. The key fact of the whole story is that without applying parallel transport to transport the vector at point $\gamma(t)$ to $T_{\gamma(0)}\mathcal{M}$, then the subtraction $X(\gamma(t)) - X(\gamma(0)) \in \mathbb{R}^3$ in general does not live on $T_x\mathcal{M}$, which distorts the notion of derivative. For comparison, let us reconsider the definition (A.1). Since at each point $x \in \mathbb{R}^3$, the tangent plane at x is $T_x\mathbb{R}^3 = \mathbb{R}^3$, the substraction $X(x+tv) - X(x)$ always makes sense. To be more precise, the true meaning of $X(x + tv)$ is $P_{\gamma(0),\gamma(t)}X(\gamma(t))$, where $P_{\gamma(0),\gamma(t)} = id$, and $\gamma(t) = x + tv$.

With the above definition, when X and Y are two vector fields on $\mathcal{M}$, we define $\nabla_X Y$ to be a new vector field on $\mathcal{M}$ so that

$$\nabla_X Y(x) := \nabla_{X(x)} Y.$$

Note that $X(x) \in T_x\mathcal{M}$. We call ∇ a connection on $\mathcal{M}$.[10]

Once we know how to differentiate a vector field over $\mathcal{M}$, it is natural to consider the second order differentiation of a vector field. The second order differentiation of a vector field is a natural notion in $\mathbb{R}^3$. For example, we can define a second order differentiation of a vector field X over $\mathbb{R}^3$ as follows:

$$\nabla^2 X := \nabla_x \nabla_x X + \nabla_y \nabla_y X + \nabla_z \nabla_z X, \tag{A.4}$$

where x, y, z are standard unit vectors corresponding to the three axes. This definition can be generalized to a vector field over $\mathcal{M}$ as follows:

$$\nabla^2 X(x) := \nabla_{E_1} \nabla_{E_1} X(x) + \nabla_{E_2} \nabla_{E_2} X(x), \tag{A.5}$$

where X is a vector field over $\mathcal{M}$, $x \in \mathcal{M}$, and E_1, E_2 are two vector fields on $\mathcal{M}$ that satisfy $\nabla_{E_i} E_j = 0$ for $i, j = 1, 2$. The condition $\nabla_{E_i} E_j = 0$ (for $i, j = 1, 2$) is needed for technical reasons. Note that in the $\mathbb{R}^3$ case (A.4), if we set $E_1 = x$, $E_2 = y$ and

---

[10]The notion of connection can be quite general. For our purposes, this definition is sufficient.

$E_3 = z$, then $\nabla_{E_i} E_j = 0$ for $i, j = 1, 2, 3$.[11] The operator ∇^2 is called the *connection Laplacian operator*, which lies in the heart of the VDM framework. The notion of *eigen-vector-field* over $\mathcal{M}$ is defined to be the solution of the following equation:

$$\nabla^2 X(x) = \lambda X(x)$$

for some $\lambda \in \mathbb{R}$. The existence and other properties of the eigen-vector-fields can be found in [16]. Finally, we comment that all the above definitions can be extended to the general manifold setup without much difficulty, where, roughly speaking, a "manifold" is the higher dimensional generalization of a surface[12].

Appendix B. Proof of Theorem 5.1, Theorem 5.2 and Theorem 5.3.

Before stating and proving the theorems, we set up the notation that is used throughout this Appendix. Let $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^p$ be a smooth d-dim compact Riemannian manifold embedded in $\mathbb{R}^p$, with metric g induced from the canonical metric on $\mathbb{R}^p$. Denote $\mathcal{M}_t = \{x \in \mathcal{M} : \min_{y \in \partial\mathcal{M}} d(x, y) \leq t\}$, where $d(x, y)$ is the geodesic distance between x and y. The data points $x_1, x_2, \ldots, x_n$ are independent samples from $\mathcal{M}$ according to the probability density function $p \in C^3(\mathcal{M})$ supported on $\mathcal{M} \subset \mathbb{R}^p$ and satisfies $0 < p(x) < \infty$. We assume that the kernels used in the local PCA step and for the construction of the matrix S are in $C^2([0,1])$. Although these kernels can be different we denote both of them by K and expect their meaning to be clear from the context. Denote τ to be the largest number having the property: the open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^p$ for every $r < \tau$ [30]. This condition holds automatically since $\mathcal{M}$ is compact. In all theorems, we assume that $\sqrt{\epsilon} < \tau$. In [30], $1/\tau$ is referred to as the "condition number" of $\mathcal{M}$. We denote $P_{y,x} : T_x\mathcal{M} \to T_y\mathcal{M}$ to be the parallel transport from x to y along the geodesic linking them. Denote by ∇ the connection over $T\mathcal{M}$ and ∇^2 the connection Laplacian over $\mathcal{M}$. Denote by $\mathcal{R}$, Ric, and s the curvature tensor, the Ricci curvature, and the scalar curvature of $\mathcal{M}$, respectively. The second fundamental form of the embedding ι is denoted by Π. To ease notation, in the sequel we use the same notation ∇ to denote different connections on different bundles whenever there is no confusion and the meaning is clear from the context.

We divide the proof of Theorem 5.1 into four theorems, each of which has its own interest. The first theorem, Theorem B.1, states that the columns of the matrix O_i that are found by local PCA (see (2.1)) form an orthonormal basis to a d-dimensional subspace of $\mathbb{R}^p$ that approximates the embedded tangent plane $\iota_* T_{x_i}\mathcal{M}$. The proven order of approximation is crucial for proving Theorem 5.1. The proof of Theorem B.1 involves geometry and probability theory.

THEOREM B.1. *If $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$ and $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$, then, with high probability (w.h.p.), the columns $\{u_l(x_i)\}_{l=1}^d$ of the $p \times d$ matrix O_i which is determined by local PCA, form an orthonormal basis to a d-dim subspace of $\mathbb{R}^p$ that deviates from $\iota_* T_{x_i}\mathcal{M}$ by $O(\epsilon_{PCA}^{3/2})$, in the following sense:*

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{PCA}^{3/2}) = O(n^{-\frac{3}{d+2}}), \qquad (\text{B.1})$$

---

[11] Please see [32] for details.
[12] We will not provide details in the manifold setting, and refer readers to standard differential geometry textbooks, such as [32].

where Θ_i is a $p \times d$ matrix whose columns form an orthonormal basis to $\iota_* T_{x_i}\mathcal{M}$. Let the minimizer in (B.1) be

$$\hat{O}_i = \underset{O \in O(d)}{\operatorname{argmin}} \|O_i^T \Theta_i - O\|_{HS}, \tag{B.2}$$

and denote by Q_i the $p \times d$ matrix

$$Q_i := \Theta_i \hat{O}_i^T, \tag{B.3}$$

and $e_l(x_i)$ the l-th column of Q_i. The columns of Q_i form an orthonormal basis to $\iota_* T_{x_i}\mathcal{M}$, and

$$\|O_i - Q_i\|_{HS} = O(\epsilon_{PCA}). \tag{B.4}$$

If $x_i \in \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$, then, w.h.p.

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{PCA}^{1/2}) = O(n^{-\frac{1}{d+2}}).$$

Better convergence near the boundary is obtained for $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$, which gives

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{PCA}^{3/4}) = O(n^{-\frac{3}{2(d+1)}}),$$

for $x_i \in \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$, and

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{PCA}^{5/4}) = O(n^{-\frac{5}{2(d+1)}}), \tag{B.5}$$

for $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$.

Theorem B.1 may seem a bit counterintuitive at first glance. When considering data points in a ball of radius $\sqrt{\epsilon_{PCA}}$, it is expected that the order of approximation would be $O(\epsilon_{PCA})$, while equation (B.1) indicates that the order of approximation is higher (3/2 instead of 1). The true order of approximation for the tangent space, as observed in (B.4) is still $O(\epsilon)$. The improvement observed in (B.1) is of relevance to Theorem B.2 and we relate it to the probabilistic nature of the PCA procedure, more specifically, to a large deviation result for the error in the law of large numbers for the covariance matrix that underlies PCA. Since the convergence of PCA is slower near the boundary, then for manifolds with boundary we need a smaller ϵ_{PCA}. Specifically, for manifolds without boundary we choose $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$ and for manifolds with boundary we choose $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$. We remark that the first choice works also for manifolds with boundary at the expense of a slower convergence rate.

The second theorem, Theorem B.2, states that the $d \times d$ orthonormal matrix O_{ij}, which is the output of the alignment procedure (2.4), approximates the parallel transport operator P_{x_i, x_j} from x_j to x_i along the geodesic connecting them. Assuming that $\|x_i - x_j\| = O(\sqrt{\epsilon})$ (here, ϵ is different than ϵ_{PCA}), the order of this approximation is $O(\epsilon_{PCA}^{3/2} + \epsilon^{3/2})$ whenever x_i, x_j are away from the boundary. This result is crucial for proving Theorem 5.1. The proof of Theorem B.2 uses Theorem B.1 and is purely geometric.

THEOREM B.2. *Consider* $x_i, x_j \notin \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$ *satisfying that the geodesic distance between* x_i *and* x_j *is* $O(\sqrt{\epsilon})$. *For* $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$, *w.h.p.,* O_{ij} *approximates* P_{x_i, x_j} *in the following sense:*

$$O_{ij}\bar{X}_j = \left(\langle \iota_* P_{x_i, x_j} X(x_j), u_l(x_i) \rangle\right)_{l=1}^d + O(\epsilon_{PCA}^{3/2} + \epsilon^{3/2}), \quad \text{for all } X \in C^3(T\mathcal{M}),$$
(B.6)

where $\bar{X}_i \equiv (\langle \iota_* X(x_i), u_l(x_i) \rangle)_{l=1}^d \in \mathbb{R}^d$, *and* $\{u_l(x_i)\}_{l=1}^d$ *is an orthonormal set determined by local PCA. For* $x_i, x_j \in \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$

$$O_{ij}\bar{X}_j = \left(\langle \iota_* P_{x_i, x_j} X(x_j), u_l(x_i) \rangle\right)_{l=1}^d + O(\epsilon_{PCA}^{1/2} + \epsilon^{3/2}), \quad \text{for all } X \in C^3(T\mathcal{M}),$$
(B.7)

For $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$, *the orders of* ϵ_{PCA} *in the error terms change according to Theorem B.1.*

The third theorem, Theorem B.3, states that the $n \times n$ block matrix $D_\alpha^{-1} S_\alpha$ is a discrete approximation of an integral operator over smooth sections of the tangent bundle. The integral operator involves the parallel transport operator. The proof of Theorem B.3 mainly uses probability theory.

THEOREM B.3. *Suppose* $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$, *and for* $0 \le \alpha \le 1$, *define the estimated probability density distribution by*

$$p_\epsilon(x_i) = \sum_{j=1}^n K_\epsilon(x_i, x_j)$$

and the normalized kernel $K_{\epsilon, \alpha}$ *by*

$$K_{\epsilon, \alpha}(x_i, x_j) = \frac{K_\epsilon(x_i, x_j)}{p_\epsilon^\alpha(x_i) p_\epsilon^\alpha(x_j)},$$

where $K_\epsilon(x_i, x_j) = K\left(\frac{\|\iota(x_i) - \iota(x_j)\|_{\mathbb{R}^p}}{\sqrt{\epsilon}}\right)$.
For $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$ *we have w.h.p.*

$$\frac{\sum_{j=1, j \ne i}^n K_{\epsilon, \alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1, j \ne i}^n K_{\epsilon, \alpha}(x_i, x_j)} = \left(\langle \iota_* T_{\epsilon, \alpha} X(x_i), u_l(x_i) \rangle\right)_{l=1}^d$$
(B.8)

$$+ O\left(\frac{1}{n^{1/2} \epsilon^{d/4 - 1/2}} + \epsilon_{PCA}^{3/2} + \epsilon^{3/2}\right),$$

where

$$T_{\epsilon, \alpha} X(x_i) = \frac{\int_\mathcal{M} K_{\epsilon, \alpha}(x_i, y) P_{x_i, y} X(y) dV(y)}{\int_\mathcal{M} K_{\epsilon, \alpha}(x_i, y) dV(y)},$$
(B.9)

$\bar{X}_i \equiv (\langle \iota_* X(x_i), u_l(x_i) \rangle)_{l=1}^d \in \mathbb{R}^d$, $\{u_l(x_i)\}_{l=1}^d$ *is the orthonormal set determined by local PCA,* $X \in C^3(T\mathcal{M})$, *and* O_{ij} *is the optimal orthogonal transformation determined by the alignment procedure.*

For $x_i \in \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$ *we have w.h.p.*

$$\frac{\sum_{j=1, j \ne i}^n K_{\epsilon, \alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1, j \ne i}^n K_{\epsilon, \alpha}(x_i, x_j)} = \left(\langle \iota_* T_{\epsilon, \alpha} X(x_i), u_l(x_i) \rangle\right)_{l=1}^d$$
(B.10)

$$+ O\left(\frac{1}{n^{1/2} \epsilon^{d/4 - 1/2}} + \epsilon_{PCA}^{1/2} + \epsilon^{3/2}\right).$$

37

For $\epsilon_{PCA} = O(n^{-\frac{2}{d+1}})$ the orders of ϵ_{PCA} in the error terms change according to Theorem B.1.

The fourth theorem, Theorem B.4, states that the operator $T_{\epsilon,\alpha}$ can be expanded in powers of $\sqrt{\epsilon}$, where the leading order term is the identity operator, the second order term is the connection-Laplacian operator plus some possible potential terms, and the first and third order terms vanish for vector fields that are sufficiently smooth. For $\alpha = 1$, the potential terms vanish, and as a result, the second order term is the connection-Laplacian. The proof is based on geometry.

THEOREM B.4. *For $X \in C^3(T\mathcal{M})$ and $x \notin \mathcal{M}_{\sqrt{\epsilon}}$ we have:*

$$T_{\epsilon,\alpha}X(x) = X(x) + \epsilon \frac{m_2}{2dm_0}\left\{\nabla^2 X(x) + d\frac{\int_{S^{d-1}}\nabla_\theta X(x)\nabla_\theta(p^{1-\alpha})(x)\mathrm{d}\theta}{p^{1-\alpha}(x)}\right\} + O(\epsilon^2),$$

(B.11)

where $T_{\epsilon,\alpha}$ is defined in (B.9), ∇^2 is the connection-Laplacian over vector fields, and $m_l = \int_{\mathbb{R}^d}\|x\|^l K(\|x\|)\mathrm{d}x$.

COROLLARY B.5. *Under the same conditions and notations as in Theorem B.4, if $X \in C^3(T\mathcal{M})$, then for all $x \notin \mathcal{M}_{\sqrt{\epsilon}}$ we have:*

$$T_{\epsilon,1}X(x) = X(x) + \epsilon\frac{m_2}{2dm_0}\nabla^2 X(x) + O(\epsilon^2).$$
(B.12)

Putting Theorems B.1 B.3 and B.4 together, we now prove Theorem 5.1:

Proof. [Proof of Theorem 5.1] Suppose $x_i \notin \mathcal{M}_{\sqrt{\epsilon}}$. By Theorem B.3, w.h.p.

$$\frac{\sum_{j=1,j\neq i}^n K_{\epsilon,\alpha}(x_i,x_j)O_{ij}\bar{X}_j}{\sum_{j=1,j\neq i}^n K_{\epsilon,\alpha}(x_i,x_j)} = (\langle \iota_* T_{\epsilon,\alpha}X(x_i), u_l(x_i)\rangle)_{l=1}^d$$
(B.13)

$$+ O\left(\frac{1}{n^{1/2}\epsilon^{d/4-1/2}} + \epsilon_{PCA}^{3/2} + \epsilon^{3/2}\right),$$

$$= (\langle \iota_* T_{\epsilon,\alpha}X(x_i), e_l(x_i)\rangle)_{l=1}^d$$

$$+ O\left(\frac{1}{n^{1/2}\epsilon^{d/4-1/2}} + \epsilon_{PCA}^{3/2} + \epsilon^{3/2}\right),$$

where $\epsilon_{PCA} = O(n^{-\frac{2}{d+2}})$, and we used Theorem B.1 to replace $u_l(x_i)$ by $e_l(x_i)$. Using Theorem B.4 for the right hand side of (B.13), we get

$$\frac{\sum_{j=1,j\neq i}^n K_{\epsilon,\alpha}(x_i,x_j)O_{ij}\bar{X}_j}{\sum_{j=1,j\neq i}^n K_{\epsilon,\alpha}(x_i,x_j)}$$

$$= \left(\left\langle \iota_* X(x_i) + \epsilon\frac{m_2}{2dm_0}\iota_*\left\{\nabla^2 X(x_i) + d\frac{\int_{S^{d-1}}\nabla_\theta X(x_i)\nabla_\theta(p^{1-\alpha})(x_i)\mathrm{d}\theta}{p^{1-\alpha}(x_i)}\right\}, e_l(x_i)\right\rangle\right)_{l=1}^d$$

$$+ O\left(\frac{1}{n^{1/2}\epsilon^{d/4-1/2}} + \epsilon_{PCA}^{3/2} + \epsilon^{3/2}\right).$$

For $\epsilon = O(n^{-\frac{2}{d+4}})$, upon dividing by ϵ, the three error terms are

$$\frac{1}{n^{1/2}\epsilon^{d/4+1/2}} = O(n^{-\frac{1}{d+4}}),$$

$$\frac{1}{\epsilon}\epsilon_{PCA}^{3/2} = O(n^{-\frac{d+8}{(d+1)(d+2)}}),$$

$$\epsilon^{1/2} = O(n^{-\frac{1}{d+4}}).$$

38

Clearly the three error terms vanish as $n \to \infty$. Specifically, the dominant error is $O(n^{-\frac{1}{d+4}})$ which is the same as $O(\sqrt{\epsilon})$. As a result, in the limit $n \to \infty$, almost surely,

$$\lim_{n \to \infty} \frac{1}{\epsilon} \left[\frac{\sum_{j=1, j\neq i}^{n} K_{\epsilon,\alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1, j\neq i}^{n} K_{\epsilon,\alpha}(x_i, x_j)} - \bar{X}_i \right]$$

$$= \frac{m_2}{2dm_0} \left(\left\langle \iota_* \left\{ \nabla^2 X(x_i) + d \frac{\int_{S^{d-1}} \nabla_\theta X(x_i) \nabla_\theta (p^{1-\alpha})(x_i) \mathrm{d}\theta}{p^{1-\alpha}(x_i)} \right\}, e_l(x_i) \right\rangle \right)_{l=1}^{d},$$

as required. □

B.1. Preliminary Lemmas. For the proofs of Theorem B.1-B.4, we need the following Lemmas.

LEMMA B.6. *In polar coordinates around $x \in \mathcal{M}$, the Riemannaian measure is given by*

$$\mathrm{d}V(\exp_x t\theta) = J(t, \theta)\mathrm{d}t\mathrm{d}\theta,$$

where $\theta \in T_x\mathcal{M}$, $\|\theta\| = 1$, $t > 0$, and

$$J(t, \theta) = t^{d-1} + t^{d+1} Ric(\theta, \theta) + O(t^{d+2}).$$

Proof. Please see [32]. □

The following Lemma is needed in Theorem B.1 and B.2.

LEMMA B.7. *Fix $x \in \mathcal{M}$ and denote $\exp_x$ the exponential map at x and $\exp^{\mathbb{R}^p}_{\iota(x)}$ the exponential map at $\iota(x)$. With the identification of $T_{\iota(x)}\mathbb{R}^p$ with $\mathbb{R}^p$, for $v \in T_x\mathcal{M}$ with $\|v\| \ll 1$ we have*

$$\iota \circ \exp_x(v) = \iota(x) + d\iota(v) + \frac{1}{2}\Pi(v, v) + \frac{1}{6}\nabla_v \Pi(v, v) + O(\|v\|^4). \tag{B.14}$$

Furthermore, for $w \in T_x\mathcal{M} \cong \mathbb{R}^d$, we have

$$d\left[\iota \circ \exp_x\right]_v(w) = d\left[\iota \circ \exp_x\right]_{v=0}(w) + \Pi(v, w) + \frac{1}{6}\nabla_v \Pi(v, w) + \frac{1}{3}\nabla_w \Pi(v, v) + O(\|v\|^3) \tag{B.15}$$

Proof. Denote $\phi = (\exp^{\mathbb{R}^p}_{\iota(x)})^{-1} \circ \iota \circ \exp_x$, that is,

$$\iota \circ \exp_x = \exp^{\mathbb{R}^p}_{\iota(x)} \circ \phi. \tag{B.16}$$

Note that $\phi(0) = 0$. Since ϕ can be viewed as a function from $T_x\mathcal{M} \cong \mathbb{R}^d$ to $T_{\iota(x)}\mathbb{R}^p \cong \mathbb{R}^p$, we can Taylor expand it to get

$$\iota \circ \exp_x(v) = \exp^{\mathbb{R}^p}_{\iota(x)} \left(d\phi|_0(v) + \frac{1}{2}\nabla d\phi|_0(v, v) + \frac{1}{6}\nabla^2 d\phi|_0(v, v, v) + O(\|v\|^4) \right).$$

We claim that

$$\nabla^k d\exp_x|_0(v, \ldots, v) = 0 \text{ for all } k \geq 1. \tag{B.17}$$

Indeed, from the definition of the exponential map we have that $d\exp_x \in \Gamma(T^*T_x\mathcal{M} \otimes T\mathcal{M})$ and

$$\nabla d\exp_x(v'(t), v'(t)) = \nabla_{d\exp_x(v'(t))} d\exp_x(v'(t)) - d\exp_x(\nabla_{v'(t)} v'(t)),$$

where $v \in T_x\mathcal{M}$, $v(t) = tv \in T_x\mathcal{M}$, and $v'(t) = v \in T_{tv}T_x\mathcal{M}$. When evaluated at $t = 0$, we get the claim (B.17) for $k = 1$. The result for $k \geq 2$ follows from a similar argument.

We view $d\iota$ as a smooth section of $Hom(T\mathcal{M}, T\mathbb{R}^p)$. Thus, by combining (B.17) with the chain rule, from (B.16) we have

$$\nabla^k d\phi|_0(v, \ldots, v) = \nabla^k d\iota|_x(v, \ldots, v) \text{ for all } k \geq 0,$$

and hence we obtain

$$\iota \circ \exp_x(v) = \exp_{\iota(x)}^{\mathbb{R}^p}\left(d\iota(v) + \frac{1}{2}\nabla d\iota(v, v) + \frac{1}{6}\nabla^2 d\iota(v, v, v) + O(\|v\|^4) \right).$$

To conclude (B.14), note that for all $v \in T_x\mathcal{M}$, we have $\exp_{\iota(x)}^{\mathbb{R}^p}(v) = \iota(x) + v$ for all $v \in T_{\iota(x)}\mathbb{R}^p$ if we identify $T_{\iota(x)}\mathbb{R}^p$ with $\mathbb{R}^p$. Next consider vector fields U, V and W around x so that $U(x) = u$, $V(x) = v$ and $W(x) = w$, where $u, v, w \in T_x\mathcal{M}$. A direct calculation gives

$$\nabla d\iota(V, W) = (\nabla_V d\iota)W = \nabla_V(d\iota(W)) - d\iota(\nabla_V W)$$

which is by definition the second fundamental form $\Pi(V, W)$ of the embedding ι. Similarly, we have

$$\begin{aligned}
\nabla^2 d\iota(U, V, W) &= (\nabla^2_{U,V} d\iota)W = (\nabla_U(\nabla_V d\iota))W - (\nabla_{\nabla_U V} d\iota)W \\
&= \nabla_U((\nabla_V d\iota)W) - \nabla_V d\iota(\nabla_U W) - (\nabla_{\nabla_U V} d\iota)W \\
&= \nabla_U(\Pi(V, W)) - \Pi(V, \nabla_U W) - \Pi(\nabla_U V, W) =: (\nabla_U\Pi)(V, W),
\end{aligned}$$

Evaluating $\nabla d\iota(V, V)$ and $\nabla^2 d\iota(V, V, V)$ at x gives us (B.14).

Next, when $w \in T_v T_x\mathcal{M}$ and $v \in T_x\mathcal{M}$, since $d\left[\iota \circ \exp_x\right]_v(w) \in T_{\iota \circ \exp_x v}\mathbb{R}^p \cong \mathbb{R}^p$, we can view $d\left[\iota \circ \exp_x\right].(w)$ as a function from $T_x\mathcal{M} \cong \mathbb{R}^d$ to $\mathbb{R}^p$. Thus, when $\|v\|$ is small enough, Taylor expansion gives us

$$d\left[\iota \circ \exp_x\right]_v(w) = d\left[\iota \circ \exp_x\right]_0(w) + \nabla(d\left[\iota \circ \exp_x\right].(w))|_0(v) + \frac{1}{2}\nabla^2(d\left[\iota \circ \exp_x\right].(w))|_0(v, v) + O(\|v\|^3),$$

here d and ∇ are understood as the ordinary differentiation over $\mathbb{R}^d$. To simplify the following calculation, for $u, v, w \in \mathbb{R}^d$, we denote

$$H_w(v) = \frac{1}{6}\nabla_w\Pi(v, v) + \frac{1}{6}\nabla_v\Pi(w, v) + \frac{1}{6}\nabla_v\Pi(v, w),$$

and

$$G_w(u, v) = \frac{1}{3}(\nabla_w\Pi(u, v) + \nabla_u\Pi(w, v) + \nabla_v\Pi(u, w)).$$

Note that we again identify $\mathbb{R}^d$ with $T_x\mathcal{M}$ in the following calculation. By (B.14), when $\|v\|$ is small enough, we have

$$\begin{aligned}
d\left[\iota \circ \exp_x\right]_v(w) &= \lim_{\delta \to 0} \frac{\iota \circ \exp_x(v + \delta w) - \iota \circ \exp_x(v)}{\delta} \\
&= \lim_{\delta \to 0} \frac{d\iota(\delta w) + \Pi(v, \delta w) + \delta H(v) + R(v + \delta w) - R(v)}{\delta},
\end{aligned}$$

where $R(v)$ is the remainder term in the Taylor expansion:

$$R(v) = \sum_{|\alpha|=4} \frac{1}{\alpha!} \left(\int_0^1 (1-t)^3 \nabla^3 (\iota \circ \exp_x)(tv) \mathrm{d}t \right) v^\alpha,$$

from which it follows that

$$\frac{R(v+\delta w) - R(v)}{\delta} = O(\|v\|\|w\|),$$

and as a result

$$d\left[\iota \circ \exp_x\right]_v (w) = d\iota(w) + \Pi(v, w) + H(v) + O(\|v\|\|w\|). \tag{B.18}$$

Similarly, from (B.18), when $\|u\|$ is small enough we have

$$\nabla(d\left[\iota \circ \exp_x\right].(w))|_u(v) = \lim_{\delta \to 0} \frac{d\left[\iota \circ \exp_x\right]_{u+\delta v}(w) - d\left[\iota \circ \exp_x\right]_u(w)}{\delta} \tag{B.19}$$
$$= \Pi(v, w) + G(u, v) + O(\|u\|\|v\|\|w\|).$$

Finally, from (B.19) we have

$$\nabla^2 d(\left[\iota \circ \exp_x\right].(w))|_0(v, v) = \lim_{\delta \to 0} \frac{\nabla(d\left[\iota \circ \exp_x\right].(w))|_{\delta v}(v) - \nabla(d\left[\iota \circ \exp_x\right].(w))|_0(v)}{\delta}$$
$$= G(v, v). \tag{B.20}$$

Thus, from (B.18) we have that

$$d\left[\iota \circ \exp_x\right]_0 (w) = d\iota(w),$$

from (B.19) we have that

$$\nabla(d\left[\iota \circ \exp_x\right].(w))|_0(v) = \Pi(v, w),$$

and from (B.20) we have that

$$G(v, v) = \frac{1}{3}\nabla_v \Pi(v, w) + \frac{2}{3}\nabla_w \Pi(v, v).$$

Putting it all together we get (B.15) as required. □

LEMMA B.8. *Suppose* $x, y \in \mathcal{M}$ *such that* $y = \exp_x(t\theta)$, *where* $\theta \in T_x\mathcal{M}$ *and* $\|\theta\| = 1$. *If* $t \ll 1$, *then* $h = \|\iota(x) - \iota(y)\| \ll 1$ *satisfies*

$$t = h + \frac{1}{24}\|\Pi(\theta, \theta)\|h^3 + O(h^4). \tag{B.21}$$

Proof. Please see [9] or apply (B.14) directly. □

LEMMA B.9. *Fix* $x \in \mathcal{M}$ *and* $y = \exp_x(t\theta)$, *where* $\theta \in T_x\mathcal{M}$ *and* $\|\theta\| = 1$. *Let* $\{\partial_l(x)\}_{l=1}^d$ *be the normal coordinate on a neighborhood* U *of* x, *then for a sufficiently small* t, *we have:*

$$\iota_* P_{y,x} \partial_l(x) = \iota_* \partial_l(x) + t\Pi(\theta, \partial_l(x)) + \frac{t^2}{6}\nabla_\theta \Pi(\theta, \partial_l(x))$$
$$+ \frac{t^2}{3}\nabla_{\partial_l(x)}\Pi(\theta, \theta) - \frac{t^2}{6}\iota_* P_{y,x}(\mathcal{R}(\theta, \partial_l(x))\theta) + O(t^3). \tag{B.22}$$

41

for all $l = 1, \ldots, d$.

Proof. Choose an open subset $U \subset \mathcal{M}$ small enough and find an open neighborhood B of $0 \in T_x\mathcal{M}$ so that $\exp_x : B \to U$ is diffeomorphic. It is well known that

$$\partial_l(\exp_x(t\theta)) = \frac{J_l(t)}{t},$$

where $J_l(t)$ is the Jacobi field with $J_l(0) = 0$ and $\nabla_t J_l(0) = \partial_l(x)$. By applying Taylor's expansion in a neighborhood of $t = 0$, we have

$$J_l(t) = P_{y,x}\left(J_l(0) + t\nabla_t J_l(0) + \frac{t^2}{2}\nabla_t^2 J_l(0) + \frac{t^3}{6}\nabla_t^3 J_l(0)\right) + O(t^4),$$

Since $J_l(0) = \nabla_t^2 J_l(0) = 0$, the following relationship holds:

$$\begin{aligned}
\partial_l(\exp_x(t\theta)) &= P_{y,x}\left(\nabla_t J_l(0) + \frac{t^2}{6}\nabla_t^3 J_l(0)\right) + O(t^3) \\
&= P_{y,x}\partial_l(x) + \frac{t^2}{6}P_{y,x}(\mathcal{R}(\theta, \partial_l(x))\theta) + O(t^3).
\end{aligned} \tag{B.23}$$

Thus we obtain

$$P_{y,x}\partial_l(x) = \partial_l(\exp_x(t\theta)) - \frac{t^2}{6}P_{y,x}(\mathcal{R}(\theta, \partial_l(x))\theta) + O(t^3), \tag{B.24}$$

On the other hand, from (B.15) in Lemma B.7 we have

$$\iota_*\partial_l(\exp_x(t\theta)) = \iota_*\partial_l(x) + t\Pi(\theta, \partial_l(x)) + \frac{t^2}{6}\nabla_\theta\Pi(\theta, \partial_l(x)) + \frac{t^2}{3}\nabla_{\partial_l(x)}\Pi(\theta, \theta) + O(t^3). \tag{B.25}$$

Putting (B.24) and (B.25) together, it follows that for $l = 1, \ldots, d$:

$$\begin{aligned}
\iota_*P_{y,x}\partial_l(x) &= \iota_*\partial_l(\exp_x(t\theta)) - \frac{t^2}{6}\iota_*P_{y,x}(\mathcal{R}(\theta, \partial_l(x))\theta) + O(t^3) \\
&= \iota_*\partial_l(x) + t\Pi(\theta, \partial_l(x)) + \frac{t^2}{6}\nabla_\theta\Pi(\theta, \partial_l(x)) \\
&\quad + \frac{t^2}{3}\nabla_{\partial_l(x)}\Pi(\theta, \theta) - \frac{t^2}{6}\iota_*P_{y,x}(\mathcal{R}(\theta, \partial_l(x))\theta) + O(t^3).
\end{aligned} \tag{B.26}$$

□

B.2. [Proof of Theorem B.1]. *Proof.* Fix $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{PCA}}}$. Denote $\{v_k\}_{k=1}^p$ the standard orthonormal basis of $\mathbb{R}^p$, that is, v_k has 1 in the k-th entry and 0 elsewhere. We can properly translate and rotate the embedding ι so that $\iota(x_i) = 0$, the first d components $\{v_1, \ldots, v_d\} \subset \mathbb{R}^p$ form the orthonormal basis of $\iota_*T_{x_i}\mathcal{M}$, and find a normal coordinate $\{\partial_k\}_{k=1}^d$ around x_i so that $\iota_*\partial_k(x_i) = v_k$. Instead of directly analyzing the matrix B_i that appears in the local PCA procedure given in (2.3), we analyze the covariance matrix $\Xi_i := B_i B_i^T$, whose eigenvectors coincide with the left singular vectors of B_i. We rewrite Ξ_i as

$$\Xi_i = \sum_{\substack{j \neq i}}^n F_j, \tag{B.27}$$

where

$$F_j = K\left(\frac{\|\iota(x_i) - \iota(x_j)\|_{\mathbb{R}^p}}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right)(\iota(x_j) - \iota(x_i))(\iota(x_j) - \iota(x_i))^T, \qquad (\mathrm{B.28})$$

and

$$F_j(k,l) = K\left(\frac{\|\iota(x_i) - \iota(x_j)\|_{\mathbb{R}^p}}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right)\langle\iota(x_j) - \iota(x_i), v_k\rangle\langle\iota(x_j) - \iota(x_i), v_l\rangle. \qquad (\mathrm{B.29})$$

Denote $B_{\sqrt{\epsilon_{\mathrm{PCA}}}}(x_i)$ to be the geodesic ball of radius $\sqrt{\epsilon_{\mathrm{PCA}}}$ around x_i. We apply the same variance error analysis as in [35, Section 3] to approximate Ξ_i. Since the points x_i are independent identically distributed (i.i.d.), F_j, $j \neq i$, are also i.i.d., by the law of large numbers one expects

$$\frac{1}{n-1}\sum_{j\neq i}^{n} F_j \approx \mathbb{E}F, \qquad (\mathrm{B.30})$$

where $F = F_1$,

$$\mathbb{E}F = \int_{B_{\sqrt{\epsilon_{\mathrm{PCA}}}}(x_i)} K_{\epsilon_{\mathrm{PCA}}}(x_i, y)(\iota(y) - \iota(x_i))(\iota(y) - \iota(x_i))^T p(y)\mathrm{d}V(y), \qquad (\mathrm{B.31})$$

and

$$\mathbb{E}F(k,l) = \int_{B_{\sqrt{\epsilon_{\mathrm{PCA}}}}(x_i)} K_{\epsilon_{\mathrm{PCA}}}(x_i, y)\langle\iota(y) - \iota(x_i), v_k\rangle\langle\iota(y) - \iota(x_i), v_l\rangle p(y)\mathrm{d}V(y). \qquad (\mathrm{B.32})$$

In order to evaluate the first moment $\mathbb{E}F(k,l)$ of (B.32), we note that for $y = \exp_{x_i} v$, where $v \in T_{x_i}\mathcal{M}$, by (B.14) in Lemma B.7 we have

$$\langle\iota(\exp_{x_i} v) - \iota(x_i), v_k\rangle = \langle\iota_* v, v_k\rangle + \frac{1}{2}\langle\Pi(v,v), v_k\rangle + \frac{1}{6}\langle\nabla_v\Pi(v,v), v_k\rangle + O(\|v\|^4). \qquad (\mathrm{B.33})$$

Substituting (B.33) into (B.32), applying Taylor's expansion, and combining Lemma B.8 and Lemma B.6, we have

$$\int_{B_{\sqrt{\epsilon_{\mathrm{PCA}}}}(x_i)} K_{\epsilon_{\mathrm{PCA}}}(x_i, y)\langle\iota(y) - \iota(x_i), v_k\rangle\langle\iota(y) - \iota(x_i), v_l\rangle p(y)\mathrm{d}V(y)$$

$$= \int_{S^{d-1}}\int_0^{\sqrt{\epsilon_{\mathrm{PCA}}}}\left[K\left(\frac{t}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right) + O\left(\frac{t^3}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right)\right] \times \qquad (\mathrm{B.34})$$

$$\left\{t^2\langle\iota_*\theta, v_k\rangle\langle\iota_*\theta, v_l\rangle + \frac{t^3}{2}\left(\langle\Pi(\theta,\theta), v_k\rangle\langle\iota_*\theta, v_l\rangle + \langle\Pi(\theta,\theta), v_l\rangle\langle\iota_*\theta, v_k\rangle\right) + O(t^4)\right\} \times$$

$$\left(p(x_i) + t\nabla_\theta p(x_i) + O(t^2)\right)\left(t^{d-1} + O(t^{d+1})\right)\mathrm{d}t\mathrm{d}\theta$$

$$= \int_{S^{d-1}}\int_0^{\sqrt{\epsilon_{\mathrm{PCA}}}}\left[K\left(\frac{t}{\sqrt{\epsilon_{\mathrm{PCA}}}}\right)\langle\iota_*\theta, v_k\rangle\langle\iota_*\theta, v_l\rangle p(x_i)t^{d+1} + O(t^{d+3})\right]\mathrm{d}t\mathrm{d}\theta, \qquad (\mathrm{B.35})$$

where (B.35) holds since integrals involving odd powers of θ must vanish due to the symmetry of the sphere S^{d-1}. Note that $\langle\iota_*\theta, v_k\rangle = 0$ when $k = d+1, \ldots, p$. Therefore,

$$\mathbb{E}F(k,l) = \begin{cases} D\epsilon_{\mathrm{PCA}}^{d/2+1} + O(\epsilon_{\mathrm{PCA}}^{d/2+2}) & \text{for } 1 \leq k = l \leq d, \\ O(\epsilon_{\mathrm{PCA}}^{d/2+2}) & \text{otherwise.} \end{cases} \qquad (\mathrm{B.36})$$

43

where $D = \int_{S^{d-1}} |\langle \iota_* \theta, v_1 \rangle|^2 \mathrm{d}\theta \int_0^1 K(u) u^{d+1} \mathrm{d}u$ is a positive constant.

Similar considerations give the second moment of $F(k,l)$ as

$$\mathbb{E}[F(k,l)^2] = \begin{cases} O(\epsilon_{\mathrm{PCA}}{}^{d/2+2}) & \text{for } k,l = 1, \ldots, d, \\ O(\epsilon_{\mathrm{PCA}}{}^{d/2+4}) & \text{for } k,l = d+1, \ldots, p, \\ O(\epsilon_{\mathrm{PCA}}{}^{d/2+3}) & \text{otherwise}. \end{cases} \tag{B.37}$$

Hence, the variance of $F(k,l)$ becomes

$$\mathrm{Var}\, F(k,l) = \begin{cases} O(\epsilon_{\mathrm{PCA}}{}^{d/2+2}) & \text{for } k,l = 1, \ldots, d, \\ O(\epsilon_{\mathrm{PCA}}{}^{d/2+4}) & \text{for } k,l = d+1, \ldots, p, \\ O(\epsilon_{\mathrm{PCA}}{}^{d/2+3}) & \text{otherwise}. \end{cases} \tag{B.38}$$

We now move on to establish a large deviation bound on the estimation of $\frac{1}{n-1}\sum_{j \neq i} F_j(k,l)$ by its mean $\mathbb{E}F_j(k,l)$. For that purpose, we measure the deviation from the mean value by α and define its probability by

$$p_{k,l}(n,\alpha) := \Pr \left\{ \left| \frac{1}{n-1}\sum_{j \neq i}^{n} F_j(k,l) - \mathbb{E}F(k,l) \right| > \alpha \right\}. \tag{B.39}$$

To establish an upper bound for the probability $p_{k,l}(n,\alpha)$, we use Bernstein's inequality, see, e.g., [22]. Define

$$Y_j(k,l) := F_j(k,l) - \mathbb{E}F(k,l).$$

Clearly $Y_j(k,l)$ are zero mean i.i.d. random variables. From the definition of $F_j(k,l)$ (see B.28 and B.29) and from the calculation of its first moment (B.36), it follows that $Y_j(k,l)$ are bounded random variables. More specifically,

$$Y_j(k,l) = \begin{cases} O(\epsilon_{\mathrm{PCA}}) & \text{for } k,l = 1, \ldots, d, \\ O(\epsilon_{\mathrm{PCA}}{}^2) & \text{for } k,l = d+1, \ldots, p, \\ O(\epsilon_{\mathrm{PCA}}{}^{3/2}) & \text{otherwise}. \end{cases} \tag{B.40}$$

Consider first the case $k,l = 1, \ldots, d$, for which Bernstein's inequality gives

$$p_{k,l}(n,\alpha) \leq \exp\left\{ -\frac{(n-1)\alpha^2}{2\mathbb{E}(Y_1(k,l)^2) + O(\epsilon_{\mathrm{PCA}})\alpha} \right\} \leq \exp\left\{ -\frac{(n-1)\alpha^2}{O(\epsilon_{\mathrm{PCA}}{}^{d/2+2}) + O(\epsilon_{\mathrm{PCA}})\alpha} \right\}. \tag{B.41}$$

From (B.41) it follows that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\mathrm{PCA}}{}^{d/4+1}}{n^{1/2}} \right),$$

provided that

$$\frac{1}{n^{1/2}\epsilon_{\mathrm{PCA}}{}^{d/4}} \ll 1. \tag{B.42}$$

Similarly, for $k,l = d+1, \ldots, p$, we have

$$p_{k,l}(n,\alpha) \leq \exp\left\{ -\frac{(n-1)\alpha^2}{O(\epsilon_{\mathrm{PCA}}{}^{d/2+4}) + O(\epsilon_{\mathrm{PCA}}{}^2)\alpha} \right\},$$

which means that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\mathrm{PCA}}^{d/4+2}}{n^{1/2}}\right)$$

provided (B.42). Finally, for $k = d+1,\ldots,p$, $l = 1,\ldots,d$ or $l = d+1,\ldots,,p$, $k = 1,\ldots,d$, we have

$$p_{k,l}(n,\alpha) \leq \exp\left\{-\frac{(n-1)\alpha^2}{O(\epsilon_{\mathrm{PCA}}^{d/2+3}) + O(\epsilon_{\mathrm{PCA}}^{3/2})\alpha}\right\},$$

which means that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\mathrm{PCA}}^{d/4+3/2}}{n^{1/2}}\right)$$

provided (B.42). The condition (B.42) is quite intuitive as it is equivalent to $n\epsilon_{\mathrm{PCA}}^{d/2} \gg 1$, which says that the expected number of points inside $B_{\sqrt{\epsilon_{\mathrm{PCA}}}}(x_i)$ is large.

As a result, when (B.42) holds, w.h.p., the covariance matrix Ξ_i is given by

$$\Xi_i = \epsilon_{\mathrm{PCA}}^{d/2+1} D \begin{bmatrix} I_{d\times d} & 0_{d\times p-d} \\ 0_{p-d\times d} & 0_{p-d\times p-d} \end{bmatrix} \tag{B.43}$$

$$+\epsilon_{\mathrm{PCA}}^{d/2+2} \begin{bmatrix} O(1) & O(1) \\ O(1) & O(1) \end{bmatrix} \tag{B.44}$$

$$+\frac{\epsilon_{\mathrm{PCA}}^{d/4+1}}{\sqrt{n}} \begin{bmatrix} O(1) & O(\epsilon_{\mathrm{PCA}}^{1/2}) \\ O(\epsilon_{\mathrm{PCA}}^{1/2}) & O(\epsilon_{\mathrm{PCA}}) \end{bmatrix}, \tag{B.45}$$

where $I_{d\times d}$ is the identity matrix of size $d \times d$, and $0_{m\times m'}$ is the zero matrix of size $m \times m'$. The error term in (B.44) is the bias term due to the curvature of the manifold, while the error term in (B.45) is the variance term due to finite sampling (i.e., finite n). In particular, under the condition in the statement of the theorem for the sampling rate, namely, $\epsilon_{\mathrm{PCA}} = O(n^{-\frac{2}{d+2}})$, we have w.h.p.

$$\Xi_i = \epsilon_{\mathrm{PCA}}^{d/2+1} D \begin{bmatrix} I_{d\times d} & 0_{d\times p-d} \\ 0_{p-d\times d} & 0_{p-d\times p-d} \end{bmatrix} \tag{B.46}$$

$$+\epsilon_{\mathrm{PCA}}^{d/2+2} \begin{bmatrix} O(1) & O(1) \\ O(1) & O(1) \end{bmatrix} + \epsilon_{\mathrm{PCA}}^{d/2+3/2} \begin{bmatrix} O(1) & O(\epsilon_{\mathrm{PCA}}^{1/2}) \\ O(\epsilon_{\mathrm{PCA}}^{1/2}) & O(\epsilon_{\mathrm{PCA}}) \end{bmatrix}$$

$$= \epsilon_{\mathrm{PCA}}^{d/2+1} \left\{ D \begin{bmatrix} I_{d\times d} & 0_{d\times p-d} \\ 0_{p-d\times d} & 0_{p-d\times p-d} \end{bmatrix} + \begin{bmatrix} O(\epsilon_{\mathrm{PCA}}^{1/2}) & O(\epsilon_{\mathrm{PCA}}) \\ O(\epsilon_{\mathrm{PCA}}) & O(\epsilon_{\mathrm{PCA}}) \end{bmatrix} \right\}.$$

Note that by definition Ξ_i is symmetric, so we rewrite (B.46) as

$$\Xi_i = \epsilon_{\mathrm{PCA}}^{d/2+1} D \begin{bmatrix} I + \epsilon_{\mathrm{PCA}}^{1/2} A & \epsilon_{\mathrm{PCA}} C \\ \epsilon_{\mathrm{PCA}} C^T & \epsilon_{\mathrm{PCA}} B \end{bmatrix}, \tag{B.47}$$

where I is the $d\times d$ identity matrix, A is a $d\times d$ symmetric matrix, C is a $d\times(p-d)$ matrix, and B is a $(p-d)\times(p-d)$ symmetric matrix. All entries of A, B, and C are $O(1)$. Denote by u_k and λ_k, $k = 1,\ldots,p$, the eigenvectors and eigenvalues of Ξ_i, where the eigenvectors are orthonormal, and the eigenvalues are ordered in a decreasing order. Using regular perturbation theory, we find that $\lambda_k = D\epsilon_{\mathrm{PCA}}^{d/2+1}\left(1 + O(\epsilon_{\mathrm{PCA}}^{1/2})\right)$

(for $k = 1, \ldots, d$), and that the expansion of the first d eigenvectors $\{u_k\}_{k=1}^d$ is given by

$$u_k = \begin{bmatrix} [w_k + O(\epsilon_{\mathrm{PCA}}^{3/2})]_{d \times 1} \\ [O(\epsilon_{\mathrm{PCA}})]_{p-d \times 1} \end{bmatrix} \in \mathbb{R}^p, \qquad (\mathrm{B}.48)$$

where $\{w_k\}_{k=1}^d$ are orthonormal eigenvectors of A satisfying $Aw_k = \lambda_k^A w_k$. Indeed, a direct calculation gives us

$$\begin{bmatrix} I + \epsilon_{\mathrm{PCA}}^{1/2} A & \epsilon_{\mathrm{PCA}} C \\ \epsilon_{\mathrm{PCA}} C^T & \epsilon_{\mathrm{PCA}} B \end{bmatrix} \begin{bmatrix} w_k + \epsilon_{\mathrm{PCA}}^{3/2} v_{3/2} + \epsilon_{\mathrm{PCA}}^2 v_2 + O(\epsilon_{\mathrm{PCA}}^{5/2}) \\ \epsilon_{\mathrm{PCA}} z_1 + \epsilon_{\mathrm{PCA}}^{3/2} z_{3/2} + O(\epsilon_{\mathrm{PCA}}^2) \end{bmatrix}$$

$$\hspace{11cm} (\mathrm{B}.49)$$

$$= \begin{bmatrix} w_k + \epsilon_{\mathrm{PCA}}^{1/2} A w_k + \epsilon_{\mathrm{PCA}}^{3/2} v_{3/2} + \epsilon_{\mathrm{PCA}}^2 (A v_{3/2} + v_2 + C z_1) + O(\epsilon_{\mathrm{PCA}}^{5/2}) \\ \epsilon_{\mathrm{PCA}} C^T w_k + \epsilon_{\mathrm{PCA}}^2 B z_1 + O(\epsilon_{\mathrm{PCA}}^{5/2}) \end{bmatrix},$$

where $v_{3/2}, v_2 \in \mathbb{R}^d$ and $z_1, z_{3/2} \in \mathbb{R}^{p-d}$. On the other hand,

$$(1 + \epsilon_{\mathrm{PCA}}^{1/2} \lambda_k^A + \epsilon_{\mathrm{PCA}}^2 \lambda_2 + O(\epsilon_{\mathrm{PCA}}^{5/2})) \begin{bmatrix} w_k + \epsilon_{\mathrm{PCA}}^{3/2} v_{3/2} + \epsilon_{\mathrm{PCA}}^2 v_2 + O(\epsilon_{\mathrm{PCA}}^{5/2}) \\ \epsilon_{\mathrm{PCA}} z_1 + \epsilon_{\mathrm{PCA}}^{3/2} z_{3/2} + O(\epsilon_{\mathrm{PCA}}^2) \end{bmatrix}$$

$$\hspace{11cm} (\mathrm{B}.50)$$

$$= \begin{bmatrix} w_k + \epsilon_{\mathrm{PCA}}^{1/2} \lambda_k^A w_k + \epsilon_{\mathrm{PCA}}^{3/2} v_{3/2} + \epsilon_{\mathrm{PCA}}^2 (\lambda_k^A v_2 + v_{3/2} + \lambda_2 w_k) + O(\epsilon_{\mathrm{PCA}}^{5/2}) \\ \epsilon_{\mathrm{PCA}} z_1 + \epsilon_{\mathrm{PCA}}^{3/2} (\lambda_k^A z_1 + z_{3/2}) + O(\epsilon_{\mathrm{PCA}}^2) \end{bmatrix},$$

where $\lambda_2 \in \mathbb{R}$. Matching orders of ϵ_{PCA} between (B.49) and (B.50), we conclude that

$$\begin{aligned} O(\epsilon_{\mathrm{PCA}}): && z_1 &= C^T w_k, \\ O(\epsilon_{\mathrm{PCA}}^{3/2}): && z_{3/2} &= -\lambda_k^A z_1, \\ O(\epsilon_{\mathrm{PCA}}^2): && (A - \lambda_k^A I) v_{3/2} &= \lambda_2 w_k - C C^T w_k. \end{aligned} \qquad (\mathrm{B}.51)$$

Note that the matrix $(A - \lambda_k^A I)$ appearing in (B.51) is singular and its null space is spanned by the vector w_k, so the solvability condition is $\lambda_2 = \|C^T w_k\|^2 / \|w_k\|^2$. We mention that A is a generic symmetric matrix generated due to random finite sampling, so almost surely the eigenvalue λ_k^A is simple.

Denote O_i the $p \times d$ matrix whose k-th column is the vector u_k. We measure the deviation of the d-dim subspace of $\mathbb{R}^p$ spanned by u_k, $k = 1, \ldots, d$, from $\iota_* T_{x_i} \mathcal{M}$ by

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS}, \qquad (\mathrm{B}.52)$$

where Θ_i is a $p \times d$ matrix whose k-th column is v_k (recall that v_k is the k-th standard unit vector in $\mathbb{R}^p$). Let $\hat{O}$ be the $d \times d$ orthonormal matrix

$$\hat{O} = \begin{bmatrix} w_1^T \\ \vdots \\ w_d^T \end{bmatrix}_{d \times d}.$$

Then,

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} \le \|O_i^T \Theta_i - \hat{O}\|_{HS} = O(\epsilon_{\mathrm{PCA}}^{3/2}), \qquad (\mathrm{B}.53)$$

46

which completes the proof for points away from the boundary.

Next, we consider $x_i \in \mathcal{M}_{\sqrt{\epsilon_{\text{PCA}}}}$. The proof is almost the same as the above, so we just point out the main differences without giving the full details. The notations Ξ_i, $F_j(k,l)$, $p_{k,l}(n,\alpha)$, $Y_j(k,l)$ refer to the same quantities. Here the expectation of $F_j(k,l)$ is:

$$\mathbb{E}F(k,l) = \int_{B_{\sqrt{\epsilon_{\text{PCA}}}}(x_i) \cap \mathcal{M}} K_{\epsilon_{\text{PCA}}}(x_i, y) \langle \iota(y) - \iota(x_i), v_k \rangle \langle \iota(y) - \iota(x_i), v_l \rangle p(y) \mathrm{d}V(y).$$

(B.54)

Due to the asymmetry of the integration domain $\exp_{x_i}^{-1}(B_{\sqrt{\epsilon_{\text{PCA}}}}(x_i) \cap \mathcal{M})$ when x_i is near the boundary, we do not expect $\mathbb{E}F_j(k,l)$ to be the same as (B.36) and (B.37), since integrals involving odd powers of θ do not vanish. In particular, when $l = d+1, \ldots, p$, $k = 1, \ldots, d$ or $k = d+1, \ldots, p$, $l = 1, \ldots, d$, (B.54) becomes

$$\mathbb{E}F(k,l) = \int_{\exp_{x_i}^{-1}(B_{\sqrt{\epsilon_{\text{PCA}}}}(x_i) \cap \mathcal{M})} K\left(\frac{t}{\sqrt{\epsilon_{\text{PCA}}}}\right) \langle \iota_* \theta, v_k \rangle \langle \Pi(\theta,\theta), v_l \rangle p(x_i) t^{d+2} \mathrm{d}t \mathrm{d}\theta + O(\epsilon_{\text{PCA}}^{d/2+2})$$

$$= O(\epsilon_{\text{PCA}}^{d/2+3/2}).$$

(B.55)

Note that for $x_i \in \mathcal{M}_{\sqrt{\epsilon_{\text{PCA}}}}$ the bias term in the expansion of the covariance matrix differs from (B.44) when $l = d+1, \ldots, p$, $k = 1, \ldots, d$ or $k = d+1, \ldots, p$, $l = 1, \ldots, d$. Similar calculations show that

$$\mathbb{E}F(k,l) = \begin{cases} O(\epsilon_{\text{PCA}}^{d/2+1}) & \text{when } k, l = 1, \ldots, d, \\ O(\epsilon_{\text{PCA}}^{d/2+2}) & \text{when } k, l = d+1, \ldots, p, \\ O(\epsilon_{\text{PCA}}^{d/2+3/2}) & \text{otherwise}, \end{cases}$$

(B.56)

$$\mathbb{E}[F(k,l)^2] = \begin{cases} O(\epsilon_{\text{PCA}}^{d/2+2}) & \text{when } k, l = 1, \ldots, d, \\ O(\epsilon_{\text{PCA}}^{d/2+4}) & \text{when } k, l = d+1, \ldots, p, \\ O(\epsilon_{\text{PCA}}^{d/2+3}) & \text{otherwise}, \end{cases}$$

(B.57)

and

$$\text{Var}\, F(k,l) = \begin{cases} O(\epsilon_{\text{PCA}}^{d/2+2}) & \text{when } k, l = 1, \ldots, d, \\ O(\epsilon_{\text{PCA}}^{d/2+4}) & \text{when } k, l = d+1, \ldots, p, \\ O(\epsilon_{\text{PCA}}^{d/2+3}) & \text{otherwise}. \end{cases}$$

(B.58)

Similarly, $Y_j(k,l)$ are also bounded random variables satisfying

$$Y_j(k,l) = \begin{cases} O(\epsilon_{\text{PCA}}) & \text{for } k, l = 1, \ldots, d, \\ O(\epsilon_{\text{PCA}}^2) & \text{for } k, l = d+1, \ldots, p, \\ O(\epsilon_{\text{PCA}}^{3/2}) & \text{otherwise}. \end{cases}$$

(B.59)

Consider first the case $k, l = 1, \ldots, d$, for which Bernstein's inequality gives

$$p_{k,l}(n,\alpha) \le \exp\left\{ -\frac{(n-1)\alpha^2}{O(\epsilon_{\text{PCA}}^{d/2+2}) + O(\epsilon_{\text{PCA}})\alpha} \right\},$$

(B.60)

From (B.60) it follows that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\text{PCA}}^{d/4+1}}{n^{1/2}}\right),$$

47

provided (B.42). Similarly, for $k, l = d+1, \ldots, p$, we have

$$p_{k,l}(n, \alpha) \leq \exp\left\{-\frac{(n-1)\alpha^2}{O(\epsilon_{\mathrm{PCA}}{}^{d/2+4}) + O(\epsilon_{\mathrm{PCA}}{}^2)\alpha}\right\},$$

which means that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\mathrm{PCA}}{}^{d/4+2}}{n^{1/2}}\right)$$

provided (B.42). Finally, for $k = d+1, \ldots, p$, $l = 1, \ldots, d$ or $l = d+1, \ldots, p$, $k = 1, \ldots, d$, we have

$$p_{k,l}(n, \alpha) \leq \exp\left\{-\frac{(n-1)\alpha^2}{O(\epsilon_{\mathrm{PCA}}{}^{d/2+3}) + O(\epsilon_{\mathrm{PCA}}{}^{3/2})\alpha}\right\},$$

which means that w.h.p.

$$\alpha = O\left(\frac{\epsilon_{\mathrm{PCA}}{}^{d/4+3/2}}{n^{1/2}}\right)$$

provided (B.42). As a result, under the condition in the statement of the theorem for the sampling rate, namely, $\epsilon_{\mathrm{PCA}} = O(n^{-\frac{2}{d+2}})$, we have w.h.p.

$$\Xi_i = \epsilon_{\mathrm{PCA}}{}^{d/2+1}\begin{bmatrix} O(1) & 0 \\ 0 & 0 \end{bmatrix}$$

$$+ \epsilon_{\mathrm{PCA}}{}^{d/2+3/2}\begin{bmatrix} O(1) & O(1) \\ O(1) & O(\epsilon_{\mathrm{PCA}}{}^{1/2}) \end{bmatrix} + \frac{\epsilon_{\mathrm{PCA}}{}^{d/4+1}}{\sqrt{n}}\begin{bmatrix} O(1) & O(\epsilon_{\mathrm{PCA}}{}^{1/2}) \\ O(\epsilon_{\mathrm{PCA}}{}^{1/2}) & O(\epsilon_{\mathrm{PCA}}) \end{bmatrix}$$

$$= \epsilon_{\mathrm{PCA}}{}^{d/2+1}\left\{\begin{bmatrix} O(1) & 0_{d\times p-d} \\ 0_{p-d\times d} & 0_{p-d\times p-d} \end{bmatrix} + \begin{bmatrix} O(\epsilon_{\mathrm{PCA}}{}^{1/2}) & O(\epsilon_{\mathrm{PCA}}{}^{1/2}) \\ O(\epsilon_{\mathrm{PCA}}{}^{1/2}) & O(\epsilon_{\mathrm{PCA}}) \end{bmatrix}\right\}.$$

Then, by the same argument as in the case when $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{\mathrm{PCA}}}}$, we conclude that

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{\mathrm{PCA}}{}^{1/2}).$$

Similar calculations show that for $\epsilon_{\mathrm{PCA}} = O(n^{-\frac{2}{d+1}})$ we get

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{\mathrm{PCA}}{}^{5/4})$$

for $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{\mathrm{PCA}}}}$, and

$$\min_{O \in O(d)} \|O_i^T \Theta_i - O\|_{HS} = O(\epsilon_{\mathrm{PCA}}{}^{3/4})$$

for $x_i \in \mathcal{M}_{\sqrt{\epsilon_{\mathrm{PCA}}}}$.
□

B.3. [Proof of Theorem B.2]. *Proof.* Denote by O_i the $p \times d$ matrix whose columns $u_l(x_i)$, $l = 1, \ldots, d$ are orthonormal inside $\mathbb{R}^p$ as determined by local PCA around x_i. As in (B.3), we denote by $e_l(x_i)$ the l-th column of Q_i, where Q_i is a $p \times d$ matrix whose columns form an orthonormal basis of $\iota_* T_{x_i} \mathcal{M}$ so by Theorem B.1

48

$\|O_i^T Q_i - Id\|_{HS} = O(\epsilon_{\text{PCA}}^{3/2})$ for $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+2}})$, which is the case of focus here (if $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+1}})$ then $\|O_i^T Q_i - Id\|_{HS} = O(\epsilon_{\text{PCA}}^{5/4})$).

Fix x_i and the normal coordinate $\{\partial_l\}_{l=1}^d$ around x_i so that $\iota_* \partial_l(x_i) = e_l(x_i)$. Let $x_j = \exp_{x_i} t\theta$, where $\theta \in T_{x_i}\mathcal{M}$, $\|\theta\| = 1$ and $t = O(\sqrt{\epsilon})$. Then, by the definition of the parallel transport, we have

$$P_{x_i,x_j} X(x_j) = \sum_{l=1}^d g(X(x_j), P_{x_j,x_i} \partial_l(x_i)) \partial_l(x_i) \tag{B.61}$$

and since the parallel transport and the embedding ι are isometric we have

$$g(P_{x_i,x_j} X(x_j), \partial_l(x_i)) = g(X(x_j), P_{x_j,x_i} \partial_l(x_i)) = \langle \iota_* X(x_j), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle. \tag{B.62}$$

Local PCA provides an estimation of an orthonormal basis spanning $\iota_* T_{x_i}\mathcal{M}$, which is free up to $O(d)$. Thus, there exists $R \in O(p)$ so that $\iota_* T_{x_j}\mathcal{M}$ is invariant under R and $e_l(x_j) = R\iota_* P_{x_j,x_i} \partial_l(x_i)$ for all $l = 1, \ldots, d$. Hence we have the following relationship:

$$\langle \iota_* X(x_j), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle = \langle \sum_{k=1}^d \langle \iota_* X(x_j), e_k(x_j) \rangle e_k(x_j), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle$$

$$= \sum_{k=1}^d \langle \iota_* X(x_j), e_k(x_j) \rangle \langle e_k(x_j), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle$$

$$= \sum_{k=1}^d \langle R\iota_* P_{x_j,x_i} \partial_k(x_i), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle \langle \iota_* X(x_j), e_k(x_j) \rangle \tag{B.63}$$

$$:= \sum_{k=1}^d \bar{R}_{l,k} \langle \iota_* X(x_j), e_k(x_j) \rangle := \bar{R} X_j,$$

where $\bar{R}_{l,k} := \langle R\iota_* P_{x_j,x_i} \partial_k(x_i), \iota_* P_{x_j,x_i} \partial_l(x_i) \rangle$, $\bar{R} := [\bar{R}_{l,k}]_{l,k=1}^d$ and $X_j = (\langle \iota_* X(x_j), e_k(x_j) \rangle)_{k=1}^d$.

On the other hand, Lemma B.9 gives us

$$Q_i^T Q_j = \left[\iota_* \partial_l(x_i)^T R\iota_* P_{x_j,x_i} \partial_k(x_i) \right]_{l,k=1}^d$$

$$= \left[\iota_* P_{x_j,x_i} \partial_l(x_i)^T R\iota_* P_{x_j,x_i} \partial_k(x_i) \right]_{l,k=1}^d$$

$$- t \left[\Pi(\theta, \partial_l(x_i))^T R\iota_* P_{x_j,x_i} \partial_k(x_i) \right]_{l,k=1}^d$$

$$- \frac{t^2}{6} \left[2\nabla_{\partial_l(x_i)} \Pi(\theta, \theta)^T R\iota_* P_{x_j,x_i} \partial_k(x_i) + \nabla_\theta \Pi(\theta, \partial_l(x_i))^T R\iota_* P_{x_j,x_i} \partial_k(x_i) \right.$$

$$\left. - (\iota_* P_{x_j,x_i} \mathcal{R}(\theta, \partial_l(x_i))\theta)^T R\iota_* P_{x_j,x_i} \partial_k(x_i) \right]_{l,k=1}^d$$

$$+ O(t^3). \tag{B.64}$$

We now analyze the right hand side of (B.64) term by term. Note that since $\iota_* T_{x_j}\mathcal{M}$ is invariant under R, we have $R\iota_* P_{x_j,x_i} \partial_k(x_i) = \sum_{r=1}^d \bar{R}_{r,k} \iota_* P_{x_j,x_i} \partial_r(x_i)$. For the

49

$O(t)$ term, we have

$$\Pi(\theta, \partial_l(x_i))^T R \iota_* P_{x_j, x_i} \partial_k(x_i) = \sum_{r=1}^d \bar{R}_{r,k} \Pi(\theta, \partial_l(x_i))^T \iota_* P_{x_j, x_i} \partial_r(x_i)$$

$$= \sum_{r=1}^d \bar{R}_{r,k} \Pi(\theta, \partial_l(x_i))^T \left[\iota_* \partial_r(x_i) + t\Pi(\theta, \partial_r(x_i)) + O(t^2) \right] \tag{B.65}$$

$$= t \sum_{r=1}^d \bar{R}_{r,k} \Pi(\theta, \partial_l(x_i))^T \Pi(\theta, \partial_r(x_i)) + O(t^2)$$

where the second equality is due to Lemma B.9 and the third equality holds since $\Pi(\theta, \partial_l(x_i))$ is perpendicular to $\iota_* \partial_r(x_i)$ for all $l, r = 1, \ldots, d$. Moreover, Gauss equation gives us

$$0 = \langle \mathcal{R}(\theta, \theta) \partial_r(x_i), \partial_l(x_i) \rangle = \Pi(\theta, \partial_l(x_i))^T \Pi(\theta, \partial_r(x_i)) - \Pi(\theta, \partial_r(x_i))^T \Pi(\theta, \partial_l(x_i)),$$

which means the matrix $S_1 := \left[\Pi(\theta, \partial_l(x_i))^T \Pi(\theta, \partial_r(x_i)) \right]_{l,r=1}^d$ is symmetric.

Fix a vector field X on a neighborhood around x_i so that $X(x_i) = \theta$. By definition we have

$$\nabla_{\partial_l} \Pi(X, X) = \nabla_{\partial_l}(\Pi(X, X)) - 2\Pi(X, \nabla_{\partial_l} X) \tag{B.66}$$

Viewing $T\mathcal{M}$ as a subbundle of $T\mathbb{R}^p$, we have the equation of Weingarten:

$$\nabla_{\partial_l}(\Pi(X, X)) = -A_{\Pi(X,X)} \partial_l + \nabla_{\partial_l}^{\perp}(\Pi(X, X)), \tag{B.67}$$

where $A_{\Pi(X,X)} \partial_l$ and $\nabla_{\partial_l}^{\perp}(\Pi(X, X))$ are the tangential and normal components of $\nabla_{\partial_l}(\Pi(X, X))$ respectively. Moreover, the following equation holds:

$$\langle A_{\Pi(X,X)} \partial_l, \iota_* \partial_k \rangle = \langle \Pi(\partial_l, \partial_k), \Pi(X, X) \rangle. \tag{B.68}$$

By evaluating (B.66) and (B.67) at x_i, we have

$$\nabla_{\partial_l(x_i)} \Pi(\theta, \theta)^T R \iota_* P_{x_j, x_i} \partial_k(x_i) = \sum_{r=1}^d \bar{R}_{r,k} \nabla_{\partial_l(x_i)} \Pi(\theta, \theta)^T \iota_* P_{x_j, x_i} \partial_r(x_i)$$

$$= \sum_{r=1}^d \bar{R}_{r,k} (-A_{\Pi(\theta,\theta)} \partial_l(x_i) + \nabla_{\partial_l(x_i)}^{\perp}(\Pi(X, X)) - 2\Pi(\theta, \nabla_{\partial_l(x_i)} X))^T \left[\iota_* \partial_r(x_i) + t\Pi(\theta, \partial_r(x_i)) + O(t^2) \right]$$

$$= -\sum_{r=1}^d \bar{R}_{r,k} (A_{\Pi(\theta,\theta)} \partial_l(x_i))^T \iota_* \partial_r(x_i) + O(t) = -\sum_{r=1}^d \bar{R}_{r,k} \langle \Pi(\theta, \theta), \Pi(\partial_l(x_i), \partial_r(x_i)) \rangle + O(t).$$

$$\tag{B.69}$$

where the third equality holds since $\Pi(\theta, \nabla_{\partial_l(x_i)} X)$ and $\nabla_{\partial_l(x_i)}^{\perp}(\Pi(X, X))$ are perpendicular to $\iota_* \partial_l(x_i)$ and the last equality holds by (B.68). Due to the symmetry of the second fundamental form, we know the matrix $S_2 = \left[\langle \Pi(\theta, \theta), \Pi(\partial_l(x_i), \partial_r(x_i)) \rangle \right]_{l,r=1}^d$ is symmetric.

50

Similarly we have

$$\nabla_\theta \Pi(\partial_l(x_i), \theta)^T R \iota_* P_{x_j, x_i} \partial_k(x_i) = \sum_{r=1}^d \bar{R}_{r,k}(A_{\Pi(\partial_l(x_i), \theta)} \theta)^T \iota_* \partial_r(x_i) + O(t). \quad \text{(B.70)}$$

Since $(A_{\Pi(\partial_l(x_i), \theta)} \theta)^T \iota_* \partial_r(x_i) = \Pi(\theta, \partial_l(x_i))^T \Pi(\theta, \partial_r(x_i))$ by (B.68), which we denoted earlier by S_1 and used Gauss equation to conclude that it is symmetric.

To estimate the last term, we work out the following calculation by using the isometry of the parallel transport:

$$(\iota_* P_{x_j, x_i} \mathcal{R}(\theta, \partial_l(x_i)) \theta)^T R \iota_* P_{x_j, x_i} \partial_k(x_i) = \sum_{r=1}^d \bar{R}_{r,k} \langle \iota_* P_{x_j, x_i} (\mathcal{R}(\theta, \partial_l(x_i)) \theta), \iota_* P_{x_j, x_i} \partial_r(x_i) \rangle$$

$$= \sum_{r=1}^d \bar{R}_{r,k} g(P_{x_j, x_i} (\mathcal{R}(\theta, \partial_l(x_i)) \theta), P_{x_j, x_i} \partial_r(x_i)) = \sum_{r=1}^d \bar{R}_{r,k} g(\mathcal{R}(\theta, \partial_l(x_i)) \theta, \partial_r(x_i))$$

$$\text{(B.71)}$$

Denote $S_3 = \left[g(\mathcal{R}(\theta, \partial_l(x_i)) \theta, \partial_r(x_i)) \right]_{l,r=1}^d$, which is symmetric by the definition of $\mathcal{R}$.

Substituting (B.65), (B.69), (B.70) and (B.71) into (B.64) we have

$$Q_i^T Q_j = \bar{R} + t^2(-S_1 - S_2/3 + S_1/6 - S_3/6)\bar{R} + O(t^3) = \bar{R} + t^2 S \bar{R} + O(t^3), \quad \text{(B.72)}$$

where $S := -S_1 - S_2/3 + S_1/6 - S_3/6$ is a symmetric matrix.

Suppose that both x_i and x_j are not in $\mathcal{M}_{\sqrt{\epsilon}}$. To finish the proof, we have to understand the relationship between $O_i^T O_j$ and $Q_i^T Q_j$, which is rewritten as:

$$O_i^T O_j = Q_i^T Q_j + (O_i - Q_i)^T Q_j + O_i^T (O_j - Q_j). \quad \text{(B.73)}$$

From (B.1) in Theorem B.1, we know

$$\|(O_i - Q_i)^T Q_i\|_{HS} = \|O_i^T Q_i - Id\|_{HS} = O(\epsilon_{PCA}^{3/2}),$$

which is equivalent to

$$(O_i - Q_i)^T Q_i = O(\epsilon_{PCA}^{3/2}). \quad \text{(B.74)}$$

Due to (B.72) we have $Q_j = Q_i \bar{R} + t^2 Q_i S \bar{R} + O(t^3)$, which together with (B.74) gives

$$(O_i - Q_i)^T Q_j = (O_i - Q_i)^T (Q_i \bar{R} + t^2 Q_i S \bar{R} + O(t^3)) = O(\epsilon_{PCA}^{3/2} + \epsilon^{3/2}). \quad \text{(B.75)}$$

Together with the fact that $Q_i^T = \bar{R} Q_j^T + t^2 S \bar{R} Q_j^T + O(t^3)$ derived from (B.72), we have

$$O_i^T (O_j - Q_j) = Q_i^T (O_j - Q_j) + (O_i - Q_i)^T (O_j - Q_j) \quad \text{(B.76)}$$
$$= (\bar{R} Q_j^T + t^2 S \bar{R} Q_j^T + O(t^3))(O_j - Q_j) + (O_i - Q_i)^T (O_j - Q_j)$$
$$= O(\epsilon_{PCA}^{3/2} + \epsilon^{3/2}) + (O_i - Q_i)^T (O_j - Q_j)$$

Recall that the following relationship between O_i and Q_i holds (B.48)

$$O_i = Q_i + \begin{bmatrix} O(\epsilon_{PCA}^{3/2}) \\ O(\epsilon_{PCA}) \end{bmatrix} \quad \text{(B.77)}$$

when the embedding ι is properly translated and rotated so that it satisfies $\iota(x_i) = 0$, the first d standard unit vectors $\{v_1, \ldots, v_d\} \subset \mathbb{R}^p$ form the orthonormal basis of $\iota_* T_{x_i} \mathcal{M}$, and the normal coordinates $\{\partial_k\}_{k=1}^d$ around x_i satisfy $\iota_* \partial_k(x_i) = v_k$. Similarly, the following relationship between O_j and Q_j holds (B.48)

$$O_j = Q_j + \begin{bmatrix} O(\epsilon_{\mathrm{PCA}}^{3/2}) \\ O(\epsilon_{\mathrm{PCA}}) \end{bmatrix} \tag{B.78}$$

when the embedding ι is properly translated and rotated so that it satisfies $\iota(x_j) = 0$, the first d standard unit vectors $\{v_1, \ldots, v_d\} \subset \mathbb{R}^p$ form the orthonormal basis of $\iota_* T_{x_j} \mathcal{M}$, and the normal coordinates $\{\partial_k\}_{k=1}^d$ around x_j satisfy $\iota_* \partial_k(x_j) = \bar{R} \iota_* P_{x_j, x_i} \partial_k(x_i) = v_k$. Also recall that $\iota_* T_{x_j} \mathcal{M}$ is invariant under the rotation R and from Lemma B.9, $e_k(x_i)$ and $e_k(x_j)$ are related by $e_k(x_j) = e_k(x_i) + O(\sqrt{\epsilon})$. Therefore,

$$O_j - Q_j = (R + O(\sqrt{\epsilon})) \begin{bmatrix} O(\epsilon_{\mathrm{PCA}}^{3/2}) \\ O(\epsilon_{\mathrm{PCA}}) \end{bmatrix} = \begin{bmatrix} O(\epsilon_{\mathrm{PCA}}^{3/2}) \\ O(\epsilon_{\mathrm{PCA}}) \end{bmatrix} \tag{B.79}$$

when expressed in the standard basis of $\mathbb{R}^p$ so that the first d standard unit vectors $\{v_1, \ldots, v_d\} \subset \mathbb{R}^p$ form the orthonormal basis of $\iota_* T_{x_i} \mathcal{M}$. Hence, plugging (B.79) into (B.76) gives

$$O_i^T(O_j - Q_j) = O(\epsilon_{\mathrm{PCA}}^{3/2}) + (O_i - Q_i)^T(O_j - Q_j) = O(\epsilon_{\mathrm{PCA}}^{3/2}) \tag{B.80}$$

Inserting (B.76) and (B.80) into (B.73) concludes

$$O_i^T O_j = Q_i^T Q_j + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}). \tag{B.81}$$

Recall that O_{ij} is defined as $O_{ij} = UV^T$, where U and V comes from the singular value decomposition of $O_i^T O_j$, that is, $O_i^T O_j = U\Sigma V^T$. As a result,

$$\begin{aligned}
O_{ij} &= \underset{O \in O(d)}{\operatorname{argmin}} \left\| O_i^T O_j - O \right\|_{HS} = \underset{O \in O(d)}{\operatorname{argmin}} \left\| Q_i^T Q_j + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - O \right\|_{HS} \\
&= \underset{O \in O(d)}{\operatorname{argmin}} \| \bar{R}^T Q_i^T Q_j + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O \|_{HS} \\
&= \underset{O \in O(d)}{\operatorname{argmin}} \| Id + t^2 \bar{R}^T S \bar{R} + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O \|_{HS}.
\end{aligned}$$

Since $\bar{R}^T S \bar{R}$ is symmetric, we rewrite $\bar{R}^T S \bar{R} = U\Sigma U^T$, where U is an orthonormal matrix and Σ is a diagonal matrix with the eigenvalues of $\bar{R}^T S \bar{R}$ on its diagonal. Thus,

$$Id + t^2 \bar{R}^T S \bar{R} + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O = U(Id + t^2 \Sigma)U^T + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O.$$

Since the Hilbert-Schmidt norm is invariant to orthogonal transformations, we have

$$\begin{aligned}
\| Id + t^2 \bar{R}^T S \bar{R} + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O \|_{HS} &= \| U(Id + t^2 \Sigma)U^T + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - \bar{R}^T O \|_{HS} \\
&= \| Id + t^2 \Sigma + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2}) - U^T \bar{R}^T O U \|_{HS}.
\end{aligned}$$

Since $U^T \bar{R}^T O U$ is orthogonal, the minimizer must satisfy $U^T \bar{R}^T O U = Id + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2})$, as otherwise the sum of squares of the matrix entries would be larger. Hence we conclude $O_{ij} = \bar{R} + O(\epsilon_{\mathrm{PCA}}^{3/2} + \epsilon^{3/2})$.

Applying (B.63) and (B.48), we conclude

$$
\begin{aligned}
O_{ij}\bar{X}_j &= \bar{R}\bar{X}_j + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}) = \bar{R}(\langle \iota_* X(x_j), u_l(x_j)\rangle)_{l=1}^d + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}) \\
&= \bar{R}(\langle \iota_* X(x_j), e_l(x_j) + O(\epsilon_{\text{PCA}}^{3/2})\rangle)_{l=1}^d + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}) \\
&= (\langle \iota_* X(x_j), \iota_* P_{x_j, x_i} \partial_l(x_i)\rangle)_{l=1}^d + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}) \\
&= \left(\langle \iota_* P_{x_i, x_j} X(x_j), e_l(x_i)\rangle\right)_{l=1}^d + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}) \\
&= \left(\langle \iota_* P_{x_i, x_j} X(x_j), u_l(x_i)\rangle\right)_{l=1}^d + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2})
\end{aligned}
$$

This concludes the proof for points away from the boundary.

When x_i and x_j are in $\mathcal{M}_{\sqrt{\epsilon_{\text{PCA}}}}$, by the same reasoning as above we get

$$
O_{ij}\bar{X}_j = \left(\langle \iota_* P_{x_i, x_j} X(x_j), u_l(x_i)\rangle\right)_{l=1}^d + O(\epsilon_{\text{PCA}}^{1/2} + \epsilon^{3/2}).
$$

This concludes the proof. We remark that similar results hold for $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+1}})$ using the results in Theorem B.1. $\blacksquare$

B.4. [Proof of Theorem B.3]. *Proof.* We demonstrate the proof for the case when the data is uniformly distributed over the manifold. The proof for the non-uniform sampling case is the same but more tedious. Note that when the data is uniformly distributed, $T_{\epsilon,\alpha} = T_{\epsilon,0}$ for all $0 < \alpha \le 1$, so in the proof we focus on analyzing $T_\epsilon := T_{\epsilon,0}$. Denote $K_\epsilon := K_{\epsilon,0}$. Fix $x_i \notin \mathcal{M}_{\sqrt{\epsilon_{\text{PCA}}}}$. We rewrite the left hand side of (B.8) as

$$
\frac{\sum_{j=1, j\neq i}^n K_\epsilon(x_i, x_j) O_{ij}\bar{X}_j}{\sum_{j=1, j\neq i}^n K_\epsilon(x_i, x_j)} = \frac{\frac{1}{n-1}\sum_{j=1, j\neq i}^n F_j}{\frac{1}{n-1}\sum_{j=1, j\neq i}^n G_j}, \tag{B.82}
$$

where

$$
F_j = K_\epsilon(x_i, x_j) O_{ij}\bar{X}_j, \qquad G_j = K_\epsilon(x_i, x_j).
$$

Since $x_1, x_2, \ldots, x_n$ are i.i.d random variables, then G_j for $j \neq i$ are also i.i.d random variables. However, the random vectors F_j for $j \neq i$ are not independent, because the computation of O_{ij} involves several data points which leads to possible dependency between O_{ij_1} and O_{ij_2}. Nonetheless, Theorem B.2 implies that the random vectors F_j are well approximated by the i.i.d random vectors F_j' that are defined as

$$
F_j' := K_\epsilon(x_i, x_j)\left(\langle \iota_* P_{x_i, x_j} X(x_j), u_l(x_i)\rangle\right)_{l=1}^d, \tag{B.83}
$$

and the approximation is given by

$$
F_j = F_j' + K_\epsilon(x_i, x_j) O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}), \tag{B.84}
$$

where we use $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+2}})$ (the following analysis can be easily modified to adjust the case $\epsilon_{\text{PCA}} = O(n^{-\frac{2}{d+1}})$).

Since G_j, when $j \neq i$, are identical and independent random variables and F_j', when $j \neq i$, are identical and independent random vectors, we hereafter replace F_j'

53

and G_j by F' and G in order to ease notation. By the law of large numbers we should expect the following approximation to hold

$$\frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} F_j}{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} G_j} = \frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n}[F_j' + G_j O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2})]}{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} G_j} \approx \frac{\mathbb{E}F'}{\mathbb{E}G} + O(\epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2}),$$

(B.85)

where

$$\mathbb{E}F' = \left(\left\langle \iota_* \int_{\mathcal{M}} K_\epsilon(x_i, y) P_{x_i, y} X(y) \mathrm{d}V(y), u_l(x_i) \right\rangle \right)_{l=1}^{d},$$

(B.86)

and

$$\mathbb{E}G = \int_{\mathcal{M}} K_\epsilon(x_i, y) \mathrm{d}V(y).$$

(B.87)

In order to analyze the error of this approximation, we make use of the result in [35] (equation (3.14), p. 132) to conclude a large deviation bound on each of the d coordinates of the error. Together with a simple union bound we obtain the following large deviation bound:

$$\Pr\left\{ \left\| \frac{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} F_j'}{\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} G_j} - \frac{\mathbb{E}F'}{\mathbb{E}G} \right\| > \alpha \right\} \leq C_1 \exp\left\{ -\frac{C_2(n-1)\alpha^2 \epsilon^{d/2}\mathrm{vol}(\mathcal{M})}{2\epsilon\left[\|\nabla|_{y=x_i}\langle \iota_* P_{x_i,y}X(y), u_l(x_i)\rangle\|^2 + O(\epsilon) \right]} \right\},$$

(B.88)

where C_1 and C_2 are some constants (related to d). This large deviation bound implies that w.h.p. the variance term is $O(\frac{1}{n^{1/2}\epsilon^{d/4-1/2}})$. As a result,

$$\frac{\sum_{j=1,j\neq i}^{n} K_{\epsilon,\alpha}(x_i, x_j) O_{ij} \bar{X}_j}{\sum_{j=1,j\neq i}^{n} K_{\epsilon,\alpha}(x_i, x_j)} = (\langle \iota_* T_{\epsilon,\alpha} X(x_i), u_l(x_i)\rangle)_{l=1}^{d}$$

$$+ O\left(\frac{1}{n^{1/2}\epsilon^{d/4-1/2}} + \epsilon_{\text{PCA}}^{3/2} + \epsilon^{3/2} \right),$$

which completes the proof for points away from the boundary. The proof for points inside the boundary is similar.
□

B.5. [Proof of Theorem B.4]. *Proof.* We begin the proof by citing the following Lemma from [9, Lemma 8]:

LEMMA B.10. *Suppose $f \in \mathcal{C}^3(\mathcal{M})$ and $x \notin \mathcal{M}_{\sqrt{\epsilon}}$, then*

$$\int_{B_{\sqrt{\epsilon}}(x)} \epsilon^{-d/2} K_\epsilon(x, y) f(y) \mathrm{d}V(y) = m_0 f(x) + \epsilon \frac{m_2}{d} \left[\frac{\Delta f(x)}{2} + w(x)f(x) \right] + O(\epsilon^2)$$

where $w(x) = s(x) + \frac{m_2' z(x)}{24|S^{d-1}|}$, $s(x)$ is the scalar curvature of the manifold at x, $m_l = \int_{B_1(0)} \|x\|^l K(\|x\|)\mathrm{d}x$, $B_1(0) = \{x \in \mathbb{R}^d : \|x\|_{\mathbb{R}^d} \leq 1\}$, $m_l' = \int_{B_1(0)} \|x\|^l K'(\|x\|)\mathrm{d}x$, $z(x) = \int_{S^{d-1}} \|\Pi(\theta, \theta)\|\mathrm{d}\theta$, and Π is the second fundamental form of $\mathcal{M}$ at x.

Without loss of generality we may assume that $m_0 = 1$ for convenience of notation. By Lemma B.10, we get

$$p_\epsilon(y) = p(y) + \epsilon \frac{m_2}{d} \left(\frac{\Delta p(y)}{2} + w(y)p(y) \right) + O(\epsilon^2),$$

(B.89)

54

which leads to

$$\frac{p(y)}{p_\epsilon^\alpha(y)} = p^{1-\alpha}(y)\left[1 - \alpha\epsilon\frac{m_2}{d}\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\right] + O(\epsilon^2). \tag{B.90}$$

Plug (B.90) into the numerator of $T_{\epsilon,\alpha}X(x)$:

$$\int_{B_{\sqrt{\epsilon}}(x)} K_{\epsilon,\alpha}(x,y)P_{x,y}X(y)p(y)\mathrm{d}V(y)$$

$$= p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p_\epsilon^{-\alpha}(y)p(y)\mathrm{d}V(y)$$

$$= p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p^{1-\alpha}(y)\left[1 - \alpha\epsilon\frac{m_2}{d}\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\right]\mathrm{d}V(y) + O(\epsilon^{d/2+2})$$

$$= p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p^{1-\alpha}(y)\mathrm{d}V(y)$$

$$\quad - \frac{m_2\epsilon}{d}\alpha p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p^{1-\alpha}(y)\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\mathrm{d}V(y) + O(\epsilon^{d/2+2})$$

$$:= p_\epsilon^{-\alpha}(x)A - \frac{m_2\epsilon}{d}\alpha p_\epsilon^{-\alpha}(x)B + O(\epsilon^{d/2+2})$$

where

$$\begin{cases} A := \int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p^{1-\alpha}(y)\mathrm{d}V(y), \\ B := \int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)p^{1-\alpha}(y)\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\mathrm{d}V(y). \end{cases}$$

Note that the α-normalized integral operator (B.9) is evaluated by changing the integration variables to the local coordinates, and the odd monomials in the integral vanish because the kernel is symmetric. Thus, applying Taylor's expansion to A leads to:

$$A = \int_{S^{d-1}}\int_0^{\sqrt{\epsilon}}\left[K\left(\frac{t}{\sqrt{\epsilon}}\right) + K'\left(\frac{t}{\sqrt{\epsilon}}\right)\frac{\|\Pi(\theta,\theta)\|t^3}{24\sqrt{\epsilon}} + O\left(\frac{t^6}{\epsilon}\right)\right]\times$$

$$\left[X(x) + \nabla_\theta X(x)t + \nabla_{\theta,\theta}^2 X(x)\frac{t^2}{2} + O(t^3)\right]\times$$

$$\left[p^{1-\alpha}(x) + \nabla_\theta(p^{1-\alpha})(x)t + \nabla_{\theta,\theta}^2(p^{1-\alpha})(x)\frac{t^2}{2} + O(t^3)\right]\left[t^{d-1} + \mathrm{Ric}(\theta,\theta)t^{d+1} + O(t^{d+2})\right]\mathrm{d}t\mathrm{d}\theta$$

$$= p^{1-\alpha}(x)X(x)\int_{S^{d-1}}\int_0^{\sqrt{\epsilon}}\left\{K\left(\frac{t}{\sqrt{\epsilon}}\right)\left[1 + \mathrm{Ric}(\theta,\theta)t^2\right]t^{d-1} + K'\left(\frac{t}{\sqrt{\epsilon}}\right)\frac{\|\Pi(\theta,\theta)\|t^{d+2}}{24\sqrt{\epsilon}}\right\}\mathrm{d}t\mathrm{d}\theta$$

$$+ p^{1-\alpha}(x)\int_{S^{d-1}}\int_0^{\sqrt{\epsilon}} K\left(\frac{t}{\sqrt{\epsilon}}\right)\nabla_{\theta,\theta}^2 X(x)\frac{t^{d+1}}{2}\mathrm{d}t\mathrm{d}\theta$$

$$+ X(x)\int_{S^{d-1}}\int_0^{\sqrt{\epsilon}} K\left(\frac{t}{\sqrt{\epsilon}}\right)\nabla_{\theta,\theta}^2(p^{1-\alpha})(x)\frac{t^{d+1}}{2}\mathrm{d}t\mathrm{d}\theta$$

$$+ \int_{S^{d-1}}\int_0^{\sqrt{\epsilon}} K\left(\frac{t}{\sqrt{\epsilon}}\right)\nabla_\theta X(x)\nabla_\theta(p^{1-\alpha})(x)t^{d+1}\mathrm{d}t\mathrm{d}\theta + O(\epsilon^{d/2+2})$$

From the definition of $z(x)$ it follows that

$$\int_{B_{\sqrt{\epsilon}}(0)}\frac{1}{\epsilon^{d/2}}K'\left(\frac{t}{\sqrt{\epsilon}}\right)\frac{\|\Pi(\theta,\theta)\|t^{d+2}}{24\sqrt{\epsilon}}\mathrm{d}t\mathrm{d}\theta = \frac{\epsilon^{d/2+1}m_3'z(x)}{24|S^{d-1}|}.$$

55

Suppose $\{E_l\}_{l=1}^d$ is an orthonormal basis of $T_x\mathcal{M}$, and express $\theta = \sum_{l=1}^d \theta_l E_l$. A direct calculation shows that

$$\int_{S^{d-1}} \nabla^2_{\theta,\theta} X(x)\mathrm{d}\theta = \sum_{k,l=1}^d \int_{S^{d-1}} \theta_l\theta_k \nabla^2_{E_l,E_k} X(x)\mathrm{d}\theta = \frac{|S^{d-1}|}{d}\nabla^2 X(x),$$

and similarly

$$\int_{S^{d-1}} \mathrm{Ric}(\theta,\theta)\mathrm{d}\theta = \frac{|S^{d-1}|}{d}s(x).$$

Therefore, the first three terms of A become

$$\epsilon^{d/2}p^{1-\alpha}(x)\left\{\left(1 + \frac{\epsilon m_2}{d}\frac{\Delta(p^{1-\alpha})(x)}{2p^{1-\alpha}(x)} + \frac{\epsilon m_2}{d}w(x)\right)X(x) + \frac{\epsilon m_2}{2d}\nabla^2 X(x)\right\}. \quad \text{(B.91)}$$

The last term is simplified to

$$\int_{S^{d-1}}\int_0^{\sqrt{\epsilon}} K\left(\frac{t}{\sqrt{\epsilon}}\right)\nabla_\theta X(x)\nabla_\theta(p^{1-\alpha})(x)t^{d+1}\mathrm{d}t\mathrm{d}\theta = \epsilon^{d/2+1}\frac{m_2}{|S^{d-1}|}\int_{S^{d-1}}\nabla_\theta X(x)\nabla_\theta(p^{1-\alpha})(x)\mathrm{d}\theta.$$

Next, we consider B. Note that since there is an ϵ in front of B, we only need to consider the leading order term. Denote $Q(y) = p^{1-\alpha}(y)\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)$ to simplify notation. Thus, applying Taylor's expansion to each of the terms in the integrand of B leads to:

$$B = \int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)P_{x,y}X(y)Q(y)\mathrm{d}V(y)$$

$$= \int_{S^{d-1}}\int_0^{\sqrt{\epsilon}}\left[K\left(\frac{t}{\sqrt{\epsilon}}\right) + K'\left(\frac{t}{\sqrt{\epsilon}}\right)\frac{\|\Pi(\theta,\theta)\|t^3}{24\sqrt{\epsilon}} + O\left(\frac{t^6}{\epsilon}\right)\right]\left[X(x) + \nabla_\theta X(x)t + O(t^2)\right]$$
$$\left[Q(x) + \nabla_\theta Q(x)t + O(t^2)\right]\left[t^{d-1} + \mathrm{Ric}(\theta,\theta)t^{d+1} + O(t^{d+2})\right]\mathrm{d}t\mathrm{d}\theta$$

$$= \epsilon^{d/2}X(x)Q(x) + O(\epsilon^{d/2+1})$$

In conclusion, the numerator of $T_{\epsilon,\alpha}X(x)$ becomes

$$\epsilon^{d/2}p_\epsilon^{-\alpha}(x)p^{1-\alpha}(x)\left\{1 + \frac{\epsilon m_2}{d}\left[\frac{\Delta(p^{1-\alpha})(x)}{2p^{1-\alpha}(x)} - \alpha\frac{\Delta p(x)}{2p(x)}\right]\right\}X(x)$$

$$+ \epsilon^{d/2+1}\frac{m_2}{2d}p_\epsilon^{-\alpha}(x)p^{1-\alpha}(x)\nabla^2 X(x) + \epsilon^{d/2+1}\frac{m_2}{|S^{d-1}|}p_\epsilon^{-\alpha}(x)\int_{S^{d-1}}\nabla_\theta X(x)\nabla_\theta(p^{1-\alpha})(x)\mathrm{d}\theta + O(\epsilon^{d/2+2})$$

Similar calculation of the denominator of the $T_{\epsilon,\alpha}X(x)$ gives

$$\int_{B_{\sqrt{\epsilon}}(x)} K_{\epsilon,\alpha}(x,y)p(y)\mathrm{d}V(y)$$

$$= p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)p^{1-\alpha}(y)\left[1 - \alpha\epsilon\frac{m_2}{d}\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\right]\mathrm{d}V(y) + O(\epsilon^{d/2+2})$$

$$= p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)p^{1-\alpha}(y)\mathrm{d}V(y)$$

$$- \frac{m_2\epsilon}{d}\alpha p_\epsilon^{-\alpha}(x)\int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y)p^{1-\alpha}(y)\left(w(y) + \frac{\Delta p(y)}{2p(y)}\right)\mathrm{d}V(y) + O(\epsilon^{d/2+2})$$

$$= p_\epsilon^{-\alpha}(x)C - \frac{\epsilon m_2}{d}\alpha p_\epsilon^{-\alpha}(x)D + O(\epsilon^{d/2+2})$$

56

where
$$\begin{cases} C := \int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y) p^{1-\alpha}(y) \mathrm{d}V(y), \\ D := \int_{B_{\sqrt{\epsilon}}(x)} K_\epsilon(x,y) p^{1-\alpha}(y) \left(w(y) + \frac{\Delta p(y)}{2p(y)} \right) \mathrm{d}V(y). \end{cases}$$

We apply Lemma B.10 to C and D:

$$C = \epsilon^{d/2} p^{1-\alpha}(x) \left[1 + \frac{\epsilon m_2}{d} \left(w(x) + \frac{\Delta(p^{1-\alpha})(x)}{2p^{1-\alpha}(x)} \right) \right] + O(\epsilon^{d/2+2}),$$

and

$$D = \epsilon^{d/2} p_\epsilon^{1-\alpha}(x) \left(s(x) + \frac{\Delta p(x)}{2p(x)} \right) + O(\epsilon^{d/2+1}).$$

In conclusion, the denominator of $T_{\epsilon,\alpha} X(x)$ is

$$\epsilon^{d/2} p_\epsilon^{-\alpha}(x) p^{1-\alpha}(x) \left\{ 1 + \epsilon \frac{m_2}{d} \left(\frac{\Delta(p^{1-\alpha})(x)}{2p^{1-\alpha}(x)} - \alpha \frac{\Delta p(x)}{2p(x)} \right) \right\} + O(\epsilon^{d/2+2})$$

Putting all the above together, we have

$$T_{\epsilon,\alpha} X(x) = X(x) + \epsilon \frac{m_2}{2d} \nabla^2 X(x) + \epsilon \frac{m_2}{|S^{d-1}|} \frac{\int_{S^{d-1}} \nabla_\theta X(x) \nabla_\theta (p^{1-\alpha})(x) \mathrm{d}\theta}{p^{1-\alpha}(x)} + O(\epsilon^2)$$

In particular, when $\alpha = 1$, we have:

$$T_{\epsilon,1} X(x) = X(x) + \epsilon \frac{m_2}{2d} \nabla^2 X(x) + O(\epsilon^2)$$

☐

B.6. **[Proof of Theorem 5.2].** *Proof.* Suppose $\min_{y \in \partial \mathcal{M}} d(x,y) = \tilde{\epsilon}$. Choose a normal coordinate $\{\partial_1, \ldots, \partial_d\}$ on the geodesic ball $B_{\epsilon^{1/2}}(x)$ around x so that $x_0 = \exp_x(\tilde{\epsilon} \partial_d(x))$. Due to Gauss Lemma, we know $\operatorname{span}\{\partial_1(x_0), \ldots, \partial_{d-1}(x_0)\} = T_{x_0} \partial \mathcal{M}$ and $\partial_d(x_0)$ is outer normal at x_0.

We focus first on the integral appearing in the numerator of $T_{\epsilon,1} X(x)$:

$$\int_{B_{\sqrt{\epsilon}}(x) \cap \mathcal{M}} \frac{1}{\epsilon^{d/2}} K_{\epsilon,1}(x,y) P_{x,y} X(y) p(y) \mathrm{d}V(y).$$

We divide the integral domain $\exp_x^{-1}(B_{\sqrt{\epsilon}}(x) \cap \mathcal{M})$ into slices S_η defined by

$$S_\eta = \{ (\boldsymbol{u}, \eta) \in \mathbb{R}^d : \|(u_1, \ldots, u_{d-1}, \eta)\| < \sqrt{\epsilon} \},$$

where $\eta \in [-\epsilon^{1/2}, \epsilon^{1/2}]$ and $\boldsymbol{u} = (u_1, \ldots, u_{d-1}) \in \mathbb{R}^{d-1}$. By Taylor's expansion and (B.90), the numerator of $T_{\epsilon,1} X$ becomes

$$\int_{B_{\sqrt{\epsilon}}(x) \cap \mathcal{M}} K_{\epsilon,1}(x,y) P_{x,y} X(y) p(y) \mathrm{d}V(y)$$

$$= p_\epsilon^{-1}(x) \int_{S_\eta} \int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K \left(\frac{\sqrt{\|\boldsymbol{u}\|^2 + \eta^2}}{\sqrt{\epsilon}} \right) \left(X(x) + \sum_{i=1}^{d-1} u_i \nabla_{\partial_i} X(x) + \eta \nabla_{\partial_d} X(x) + O(\epsilon) \right)$$

$$\left[1 - \epsilon \frac{m_2}{d} \left(w(y) + \frac{\Delta p(y)}{2p(y)} \right) + O(\epsilon^2) \right] \mathrm{d}\eta \mathrm{d}\boldsymbol{u}$$

$$= p^{-1}(x) \int_{S_\eta} \int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K \left(\frac{\sqrt{\|\boldsymbol{u}\|^2 + \eta^2}}{\sqrt{\epsilon}} \right) \left(X(x) + \sum_{i=1}^{d-1} u_i \nabla_{\partial_i} X(x) + \eta \nabla_{\partial_d} X(x) + O(\epsilon) \right) \mathrm{d}\eta \mathrm{d}\boldsymbol{u}$$

$$\tag{B.92}$$

Note that in general the integral domain S_η is not symmetric with related to $(0, \ldots, 0, \eta)$, so we will try to symmetrize S_η by defining the symmetrized slices:

$$\tilde{S}_\eta = \cap_{i=1}^{d-1}(R_i S_\eta \cap S_\eta),$$

where $R_i(u_1, \ldots, u_i, \ldots, \eta) = (u_1, \ldots, -u_i, \ldots, \eta)$. Note that from (B.23) in Lemma B.9, the orthonormal basis $\{P_{x_0,x}\partial_1(x), \ldots, P_{x_0,x}\partial_{d-1}(x)\}$ of $T_{x_0}\partial\mathcal{M}$ differ from $\{\partial_1(x_0), \ldots, \partial_{d-1}(x_0)\}$ by $O(\epsilon)$. Also note that up to error of order $\epsilon^{3/2}$, we can express $\partial\mathcal{M} \cap B_{\epsilon^{1/2}}(x)$ by a homogeneous degree 2 polynomial with variables $\{P_{x_0,x}\partial_1(x), \ldots, P_{x_0,x}\partial_{d-1}(x)\}$. Thus the difference between $\tilde{S}_\eta$ and S_η is of order ϵ and (B.92) can be reduced to:

$$p^{-1}(x)\int_{\tilde{S}_\eta}\int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K\left(\frac{\sqrt{\|\boldsymbol{u}\|^2 + \eta^2}}{\sqrt{\epsilon}}\right)\left(X(x) + \sum_{i=1}^{d-1} u_i\nabla_{\partial_i}X(x) + \eta\nabla_{\partial_d}X(x) + O(\epsilon)\right)\mathrm{d}\eta\mathrm{d}\boldsymbol{u}$$
(B.93)

Next, we apply Taylor's expansion on $X(x)$:

$$P_{x,x_0}X(x_0) = X(x) + \tilde{\epsilon}\nabla_{\partial_d}X(x) + O(\epsilon).$$

Since

$$\nabla_{\partial_d}X(x) = P_{x,x_0}(\nabla_{\partial_d}X(x_0)) + O(\epsilon^{1/2}),$$

the Taylor's expansion of $X(x)$ becomes:

$$X(x) = P_{x,x_0}(X(x_0) - \tilde{\epsilon}\nabla_{\partial_d}X(x_0) + O(\epsilon)),$$
(B.94)

Similarly for all $i = 1, \ldots, d$ we have

$$P_{x,x_0}(\nabla_{\partial_i}X(x_0)) = \nabla_{\partial_i}X(x) + O(\epsilon^{1/2})$$
(B.95)

Plugging (B.94) and (B.95) into (B.93) further reduce (B.92) into:

$$p^{-1}(x)\int_{\tilde{S}_\eta}\int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K\left(\frac{\sqrt{\|\boldsymbol{u}\|^2 + \eta^2}}{\sqrt{\epsilon}}\right)P_{x,x_0}\left(X(x_0) + \sum_{i=1}^{d-1} u_i\nabla_{\partial_i}X(x_0) + (\eta - \tilde{\epsilon})\nabla_{\partial_d}X(x_0) + O(\epsilon)\right)\mathrm{d}\eta\mathrm{d}\boldsymbol{u}$$
(B.96)

The symmetry of the kernel implies that for $i = 1, \ldots, d-1$,

$$\int_{\tilde{S}_\eta} K\left(\frac{\sqrt{\|\boldsymbol{u}\|^2 + \eta^2}}{\sqrt{\epsilon}}\right)u^i\mathrm{d}\boldsymbol{u} = 0,$$
(B.97)

and hence the numerator of $T_{1,\epsilon}X(x)$ becomes

$$p^{-1}(x)P_{x,x_0}(m_0^\epsilon X(x_0) + m_1^\epsilon\nabla_{\partial_d}X(x_0)) + O(\epsilon^{d/2+1})$$
(B.98)

where

$$m_0^\epsilon = \int_{\tilde{S}_\eta}\int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K\left(\frac{\sqrt{\|u\|^2 + \eta^2}}{\sqrt{\epsilon}}\right)\mathrm{d}\eta\mathrm{d}x = O(\epsilon^{d/2})$$
(B.99)

58

and

$$m_1^\epsilon = \int_{\tilde{S}_\eta} \int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} K\left(\frac{\sqrt{\|u\|^2 + \eta^2}}{\sqrt{\epsilon}}\right)(\eta - \tilde{\epsilon})\mathrm{d}\eta\mathrm{d}x = O(\epsilon^{d/2+1/2}). \tag{B.100}$$

Similarly, the denominator of $T_{\epsilon,1}X$ can be expanded as:

$$\int_{B_{\sqrt{\epsilon}}(x)\cap\mathcal{M}} K_{\epsilon,1}(x,y)p(y)\mathrm{d}V(y) = p^{-1}(x)m_0^\epsilon + O(\epsilon^{d/2+1/2}), \tag{B.101}$$

which together with (B.98) gives us the following asymptotic expansion:

$$T_{\epsilon,1}X(x) = P_{x,x_0}\left(X(x_0) + \frac{m_1^\epsilon}{m_0^\epsilon}\nabla_{\partial_d}X(x_0)\right) + O(\epsilon). \tag{B.102}$$

Combining (B.102) with (B.10) in Theorem B.3, we conclude the theorem. $\square$

B.7. [Proof of Theorem 5.3]. *Proof.* We denote the spectrum of ∇^2 by $\{\lambda_l\}_{l=0}^\infty$, where $0 \le \lambda_0 \le \lambda_1 \le \ldots$, and the corresponding eigenspaces by $E_l := \{X \in L^2(T\mathcal{M}): \nabla^2 X = -\lambda_l X\}$, $l = 0, 1, \ldots$. The eigen-vector-fields are smooth and form a basis for $L^2(T\mathcal{M})$, that is,

$$L^2(T\mathcal{M}) = \overline{\oplus_{l\in\mathbb{N}\cup\{0\}}E_l}.$$

Thus we proceed by considering the approximation through eigen-vector-field subspaces. To simplify notation, we rescale the kenrel K so that $\frac{m_2}{2dm_0} = 1$.

Fix $X_l \in E_l$. When $x \notin \mathcal{M}_{\sqrt{\epsilon}}$, from Corollary B.5 we have uniformly

$$\frac{T_{\epsilon,1}X_l(x) - X_l(x)}{\epsilon} = \nabla^2 X_l(x) + O(\epsilon).$$

When $x \in \mathcal{M}_{\sqrt{\epsilon}}$, from Theorem 5.2 and the Neumann condition, we have uniformly

$$T_{\epsilon,1}X_l(x) = P_{x,x_0}X_l(x_0) + O(\epsilon). \tag{B.103}$$

Note that we have

$$P_{x,x_0}X_l(x_0) = X_l(x) + P_{x,x_0}\sqrt{\epsilon}\nabla_{\partial_d}X_l(x_0) + O(\epsilon),$$

thus again by the Neumann condition at x_0, (B.103) becomes

$$T_{\epsilon,1}X_l(x) = X_l(x) + O(\epsilon).$$

In conclusion, when $x \in \mathcal{M}_{\sqrt{\epsilon}}$ uniformly we have

$$\frac{T_{\epsilon,1}X_l(x) - X_l(x)}{\epsilon} = O(1).$$

Note that when the boundary of the manifold is smooth, the measure of $\mathcal{M}_{\sqrt{\epsilon}}$ is $O(\epsilon^{1/2})$. We conclude that in the L^2 sense,

$$\left\|\frac{T_{\epsilon,1}X_l - X_l}{\epsilon} - \nabla^2 X_l\right\|_{L^2} = O(\epsilon^{1/4}), \tag{B.104}$$

59

Next we show how $T_{\epsilon,1}^{t/\epsilon}$ converges to $e^{-t\nabla^2}$. We know $I + \epsilon\nabla^2$ is invertible on E_l with norm $\frac{1}{2} \leq \|I + \epsilon\nabla^2\| < 1$ when $\epsilon < \frac{1}{2\lambda_l}$. Next, note that if B is a bounded operator with norm $\|B\| < 1$, we have the following bound for any $s > 0$ by the binomial expansion:

$$
\begin{aligned}
\|(I+B)^s - I\| &= \left\| sB + \frac{s(s-1)}{2!}B^2 + \frac{s(s-1)(s-2)}{3!}B^3 + \cdots \right\| \\
&\leq s\|B\| + \frac{s(s-1)}{2!}\|B\|^2 + \frac{s(s-1)(s-2)}{3!}\|B\|^3 + \cdots \\
&= s\|B\| \left\{ 1 + \frac{s-1}{2!}\|B\| + \frac{(s-1)(s-2)}{3!}\|B\|^2 + \cdots \right\} \\
&\leq s\|B\| \left\{ 1 + \frac{s-1}{1!}\|B\| + \frac{(s-1)(s-2)}{2!}\|B\|^2 + \cdots \right\} \\
&= s\|B\|(1 + \|B\|)^{s-1}
\end{aligned}
\tag{B.105}
$$

On the other hand, note that on E_l

$$
e^{t\nabla^2} = (I + \epsilon\nabla^2)^{\frac{t}{\epsilon}} + O(\epsilon).
\tag{B.106}
$$

Indeed, for $X \in E_l$, we have $e^{t\nabla^2}X = (1 - t\lambda_l + t^2\lambda_l^2/2 + \ldots)X$ and $(I + \epsilon\nabla^2)^{\frac{t}{\epsilon}}X = (1 - t\lambda_l + t^2\lambda_l^2/2 - t\epsilon\lambda_l^2/2 + \ldots)X$ by the binomial expansion. Thus we have the claim.

Put all the above together, over E_l, for all $l \geq 0$, when $\epsilon < \frac{1}{2\lambda_l}$ we have:

$$
\begin{aligned}
\|T_{\epsilon,1}^{\frac{t}{\epsilon}} - e^{t\nabla^2}\| &= \left\| (I + \epsilon\nabla^2 + O(\epsilon^{\frac{5}{4}}))^{\frac{t}{\epsilon}} - (I + \epsilon\nabla^2)^{\frac{t}{\epsilon}} + O(\epsilon) \right\| \\
&\leq \left\| (I + \epsilon\nabla^2)^{\frac{t}{\epsilon}} \right\| \left\| \left[I + (I + \epsilon\nabla^2)^{-1}O(\epsilon^{5/4}) \right]^{\frac{t}{\epsilon}} - I + O(\epsilon) \right\| \\
&= (1 + t + O(\epsilon))(\epsilon^{1/4}t + O(\epsilon)) = O(\epsilon^{\frac{1}{4}}),
\end{aligned}
$$

where the first equality comes from (B.104) and (B.106), the third inequality comes from (B.105). Thus we have $\|T_{\epsilon,1}^{\frac{t}{\epsilon}} - e^{t\nabla^2}\| \leq O(\epsilon^{1/4})$ on $\overline{\oplus_{l:\lambda_l < \frac{1}{2\epsilon}} E_l}$. By taking $\epsilon \to 0$, the proof is completed. $\square$

Appendix C. Multiplicities of eigen-1-forms of Connection Laplacian over S^n. All results and proofs in this section can be found in [15] and [38]. Consider the following setting:

$$
G = SO(n+1), \quad K = SO(n), \quad \mathcal{M} = G/K = S^n, \quad \mathfrak{g} = so(n+1), \quad \mathfrak{k} = so(n)
$$

Denote $\Omega^1(S^n)$ the complexified smooth 1 forms, which is a G-module by $(g \cdot s)(x) = g \cdot s(g^{-1}x)$ for $g \in G$, $s \in \Omega^1(S^n)$, and $x \in S^n$. Over $\mathcal{M}$ we have Haar measure $d\mu$ and Hodge Laplacian operator $\Delta = d\delta + \delta d$. Since Δ is a self-adjoint and uniform second order elliptic operator on $\Omega^p(S^n)$, the eigenvalues λ_i are discrete and non-negative real numbers, with only accumulation point at ∞, and their related eigenspaces E_i are of finite dimension. We also know $\oplus_{i=1}^{\infty} E_i$ is dense in $\Omega^1(S^n)$ in the topology defined by the inner product $(f, g)_{S^n} \equiv \int_{S^n} \langle f, g \rangle d\mu$, where $\langle \cdot, \cdot \rangle$ is the left invariant hermitian metric defined on S^n.

Since Δ is a G-invariant differential operator, its eigenspaces E_λ are G-modules. We will count the multiplicity of E_λ by first counting how many copies of E_λ are inside $\Omega^1(S^n)$ through the Frobenius reciprocity law, followed by the branching theorem

60

and calculating $\dim E_\lambda$. On the other hand, since S^n is a symmetric space, we know $\Delta = -C$ over $\Omega^1(S^n)$, where C is the Casimir operator on G, and we can determine the eigenvalue of C over any finite dimensional irreducible submodule of $\Omega^1(S^n)$ by Freudenthal's Formula. Finally we consider the relationship between real forms of g and complex forms of g.

Note that $\mathfrak{g}/\mathfrak{k} \otimes_{\mathbb{R}} \mathbb{C} \cong \mathbb{C}^n$ when $G = SO(n+1)$ and $K = SO(n)$. Denote $V = \mathbb{C}^n = \Lambda^p(\mathfrak{g}/\mathfrak{k}) \otimes_{\mathbb{R}} \mathbb{C}$ as the standard representation of $SO(n)$.

There are two steps toward calculating the multiplicity of eigenforms over S^n.

Step 1 Clearly $\Omega^1(S^n)$ is a reducible G-module. For $\lambda \in \mathrm{Irr}(G, \mathbb{C})$, construct a G-homomorphism

$$\mathrm{Hom}_G(\Gamma_\lambda, \Omega^1(S^n)) \otimes_{\mathbb{C}} \Gamma_\lambda \to \Omega^1(S^n)$$

by $\phi \otimes v \mapsto \phi(v)$. We call the image the V_λ-isotypical summand in $\Omega^1(S^n)$ with multiplicity $\dim_{\mathbb{C}} \mathrm{Hom}_G(\Gamma_\lambda, \Omega^1(S^n))$. Then we apply Frobenius reciprocity law:

$$\mathrm{Hom}_G(\Gamma_\lambda, \Omega^1(S^n)) \cong \mathrm{Hom}_K(\mathrm{res}_K^G \Gamma_\lambda, \Lambda^1 V)$$

Thus if we can calculate $\dim \mathrm{Hom}_K(\mathrm{res}_K^G \Gamma_\lambda, \Lambda^1 V)$, we know how many copies of the irreducible representation Γ_λ inside $\Omega^1(S^n)$. To calculate it, we apply the following facts. When V_i and W_j are irreducible representations of G, we have by Schur's lemma:

$$\mathrm{Hom}_G(\oplus_{i=1}^N V_i, \oplus_{j=1}^M W_j) \cong \oplus_{i=1, j=1}^{N,M} \mathrm{Hom}_G(V_i, W_j),$$

Denote $L_1 ... L_n$ the basis for the dual space of Cartan subalgebra of $so(2n)$ or $so(2n+1)$. Then $L_1 ... L_n$, together with $\frac{1}{2} \sum_{i=1}^n L^i$ generate the weight lattice. The Weyl chamber of $SO(2n+1)$ is

$$\mathcal{W} = \left\{ \sum a_i L_i : a_1 \geq a_2 \geq ... \geq a_n \geq 0 \right\},$$

and the edges of the $\mathcal{W}$ are thus the rays generated by the vectors $L_1, L_1 + L_2, ..., L_1 + L_2 ... + L_n$; for $SO(2n)$, the Weyl chamber is

$$\mathcal{W} = \left\{ \sum a_i L_i : a_1 \geq a_2 \geq ... \geq |a_n| \right\},$$

and the edges are thus the rays generated by the vectors $L_1, L_1 + L_2, ..., L_1 + L_2 ... + L_{n-2}, L_1 + L_2 ... + L_n$ and $L_1 + L_2 ... - L_n$.

To keep notations unified, we denote the fundamental weights ω_i separately. When $G = SO(2n)$, denote

$$\begin{array}{lll} \omega_0 = 0 & \text{when } p = 0 & \\ \omega_p = \sum_{i=1}^p \lambda_i & \text{when } 1 \leq p \leq n-1 & \text{(C.1)} \\ \omega_n = \frac{1}{2} \sum_{i=1}^n \lambda_i & \text{when } p = n; & \end{array}$$

when $G = SO(2n+1)$, denote

$$\begin{array}{lll} \omega_0 = 0 & \text{when } p = 0 & \\ \omega_p = \sum_{i=1}^p \lambda_i & \text{when } 1 \leq p \leq n-2 & \\ \omega_{n-1} = \frac{1}{2} \left(\sum_{i=1}^{n-1} \lambda_i - \lambda_n \right) & \text{when } p = n & \text{(C.2)} \\ \omega_n = \frac{1}{2} \left(\sum_{i=1}^{n-1} \lambda_i + \lambda_n \right) & \text{when } p = n & \end{array}$$

61

THEOREM C.1.

(1) When $m = 2n + 1$, the exterior powers $\Lambda^p V$ of the standard representation V of $so(2n + 1)$ is the irreducible representation with the highest weight ω_p, when $p < n$ and $2\omega_n$ when $p = n$.

(2) When $m = 2n$, the exterior powers $\Lambda^p V$ of the standard representation V of $so(2n)$ is the irreducible representation with the highest weight ω_p, when $p \leq n - 1$; when $p = n$, $\Lambda^n V$ splits into two irreducible representations with the highest weight $2\omega_{m-1}$ and $2\omega_m$.

Proof. Please see [15] for details. ☐

THEOREM C.2. *(Branching theorem)*

When $G = SO(2n)$ or $G = SO(2n + 1)$, the restriction of the irreducible representations of G will be decomposed as direct sum of the irreducible representations of $K = SO(2n - 1)$ or $K = SO(2n)$ in the following way. Let Γ_λ be an irreducible G-module over $\mathbb{C}$ with the highest weight $\lambda = \sum_{i=1}^{n} \lambda_i L_i \in \mathcal{W}$

then as a K-module, Γ_λ decomposes into K-irreducible modules as follows:

(1) if $m = 2n$ ($G = SO(2n)$ and $K = SO(2n - 1)$),

$$\Gamma_\lambda = \oplus \Gamma_{\sum_{i=1}^{n} \lambda_i' L_i}$$

where $\oplus$ runs over all λ_i' such that

$$\lambda_1 \geq \lambda_1' \geq \lambda_2 \geq \lambda_2' \geq ... \geq \lambda_{n-1}' \geq |\lambda_n|$$

with λ_i' and λ_i simultaneously all integers or all half integers. Here $\Gamma_{\sum_{i=1}^{n} \lambda_i' L_i}$ is the irreducible K-module with the highest weight $\sum_{i=1}^{n} \lambda_i' L_i$

(2) if $m = 2n + 1$ ($G = SO(2n + 1)$ and $K = SO(2n)$),

$$\Gamma_\lambda = \oplus \Gamma_{\sum_{i=1}^{n} \lambda_i' L_i}$$

where $\oplus$ runs over all λ_i' such that

$$\lambda_1 \geq \lambda_1' \geq \lambda_2 \geq \lambda_2' \geq ... \geq \lambda_{n-1}' \geq \lambda_n \geq |\lambda_n'|$$

with λ_i' and λ_i simultaneously all integers or all half integers. Here $\Gamma_{\sum_{i=1}^{n} \lambda_i' L_i}$ is the irreducible K-module with the highest weight $\sum_{i=1}^{n} \lambda_i' L_i$

Proof. Please see [15] for details. ☐

Based on the above theorems, we know how to calculate $\dim_{\mathbb{C}} \text{Hom}_G(\Gamma_\lambda, \Omega^1(M))$. To be more precise, since the right hand side of $\text{Hom}_K(\text{res}_K^G \Gamma_\lambda, \Lambda^1 V)$ is the irreducible representation of K with the highest weight ω_1 (or splits in the low dimension case), we know the $\dim \text{Hom}_K(\text{res}_K^G \Gamma_\lambda, \Lambda^1 V)$ can be 1 only if $\text{res}_K^G \Gamma_\lambda$ has the same highest weight by Schur's lemma and classification theorem. Please see [15] for details.

Step 2 In this step we relate the irreducible representation $\Gamma_\lambda \subset \Omega^1(S^n)$ to the eigenvalue of eigenforms. Consider the Laplace operator on $SO(n + 1)$, which is endowed with a bi-invariant Riemannian metric. Since $SO(n + 1)$ is semi-simple, we can take the metric on $\mathfrak{g}$ to be given by the negative of the Killing form: $(X, Y) = -\text{tr}(adX adY)$. The Laplace operator Δ related to this metric is the Hodge-Laplace operator which enjoys the following relationship between its eigenvalue and its related highest weight.

THEOREM C.3. *Suppose $\Gamma_\mu \subset \Omega^1(S^n)$ is an irreducible G-module with the highest weight μ, then we have*

$$\Delta = \langle \mu + 2\rho, \mu \rangle Id_{\Gamma_\mu}$$

62

where $f \in \Gamma_\mu$, ρ is the half sum of all positive roots, and $\langle \cdot, \cdot \rangle$ is induced inner product on the dual Cartan subalgebra of $\mathfrak{g}$ from the Killing form B.

Proof. Please see [38] for details. $\square$

Note that for $SO(2n+1)$, $\rho = \sum_{i=1}^{n} \left(n + \frac{1}{2} - i\right) L_i$ since $R^+ = \{L_i - L_j, L_i + L_j, L_i : i < j\}$; for $SO(2n)$, $\rho = \sum_{i=1}^{n} (n-i) L_i$ since $R^+ = \{L_i - L_j, L_i + L_j : i < j\}$.

Combining these theorems, we know if Γ_μ is an irreducible representation of G, then it is an eigenspace of Δ with eigenvalue $\lambda = \langle \mu + 2\rho, \mu \rangle$. In particular, if we can decompose the eigenform space $E_\lambda \subset \Omega^1(S^n)$ into irreducible G-module, we can not only determine the eigenvalue λ but also its multiplicity. Indeed, we know $E_\lambda = \Gamma_\mu^{\oplus N}$, where $\lambda = \langle \mu + 2\rho, \mu \rangle$, is the isotypical summand of Γ_μ inside $\Omega^1(S^n)$.

Step 3 Now we apply Weyl Character Formula to calculate the dimension of Γ_λ for G.

THEOREM C.4.

(1) When $m = 2n+1$, consider $\lambda = \sum_{i=1}^{n} \lambda_i L_i$, where $\lambda_1 \geq ... \geq \lambda_n \geq 0$, the highest weight of an irreducible representation Γ_λ. Then $\dim \Gamma_\lambda = \Pi_{i<j} \frac{l_i - l_j}{j - i} \Pi_{i \leq j} \frac{l_i + l_j}{2n+1-i-j}$, where $l_i = \lambda_i + n - i + 1/2$. In particular, when $\lambda = \omega_p$, $p \leq n$, $\dim \Gamma_\lambda = C_p^{2n+1}$.

(2) When $m = 2n$, consider $\lambda = \sum_{i=1}^{n} \lambda_i L_i$, where $\lambda_1 \geq ... \geq |\lambda_n|$, the highest weight of an irreducible representation Γ_λ. Then $\dim \Gamma_\lambda = \Pi_{i<j} \frac{l_i - l_j}{j - i} \frac{l_i + l_j}{2n-i-j}$, where $l_i = \lambda_i + n - 1$. In particular, when $\lambda = \omega_p$, $\dim \Gamma_\lambda = C_p^{2n}$ when $p < n$ and $\dim V_\lambda = C_n^{2n}/2$.

Proof. Please see [15] for details. $\square$

Step 4 We need the following theorem about real representations of G to solve the original problem:

THEOREM C.5. *(1) When n is odd. Let ω_i be the highest weight of the representation $\Lambda^i V$ os $so_{2n+1}\mathbb{C}$. For any weight $\lambda = a_1 \omega_1 + ... + a_{n-1}\omega_{n-1} + a_n \omega_n/2$ or $so_{2n+1}\mathbb{C}$, the irreducible representation Γ_λ with highest weight λ is real if a_n is even, or if $n \cong 0$ or $3 \mod 4$; if a_n is odd and $n \equiv 1$ or $2 \mod 4$, then Γ_λ is quaternionic.*

(2) When n is even. The representation Γ_λ of $so_{2n}\mathbb{R}$ with highest weight $\lambda = a_1 \omega_1 + ... + a_{n-2}\omega_{n-2} + a_{n-1}\omega_{n-1} + a_n \omega_n$ will be complex if n is odd and $a_{n-1} \neq a_n$; it will be quaternionic if $n \equiv 2 \mod 4$ and $a_{n-1} + a_n$ is odd; and it will be real otherwise.

Proof. Please see [15] for details. $\square$

Combining the following Tables C.1 and C.2 and this Theorem we know all the eigen-1-form spaces are real form.

Step 5 Now we put all the above together. All the eigen-1-form spaces of S^n as an irreducible representation of $SO(n+1)$, $n \geq 4$, are listed in the table C.1 and C.2. λ is the highest weight. The S^3 and S^2 cases are listed in C.3 and C.4. S^3 is separated since it has a different highest weight $kL_1 - L_2$, which happens in general for n-forms in S^{2n+1}.

Consider $SO(3)$ for example. In this case, $n = 1$ and $\mathcal{M} = S^2$. From the analysis in cryo-EM [18], we know that the multiplicities of eigenvectors are $6, 10, ...$, which echoes the above analysis.

In conclusion, we can see the following compatible first few multiplicities:

S^2: 6, 10, 14
S^3: 4, 6, 9, 16, 16
S^4: 5, 10, 14
S^5: 6, 15, 20
S^6: 7, 21, 27

TABLE C.1

Eigenvalues and their multiplicity of S^{2n}, where $n \geq 2$. $l_i = \lambda_i + n - i$ and $m_i = n - i$.

λ	eigenvalues	multiplicity
$kL_1, \, k \geq 1$	$k(k + 2n - 1)$	$\Pi_{i<j} \frac{l_i^2 - l_j^2}{m_i^2 - m_j^2}$
$kL_1 + L_2, \, k \geq 1$	$(k+1)(k + 2n - 2)$	$\Pi_{i<j} \frac{l_i^2 - l_j^2}{m_i^2 - m_j^2}$

TABLE C.2

Eigenvalues and their multiplicity of S^{2n+1}, where $n \geq 2$. $l_i = \lambda_i + n - i + 1/2$ and $m_i = n - i + 1/2$.

λ	eigenvalues	multiplicity
$kL_1, \, k \geq 1$	$k(k + 2n)$	$\Pi_{i<j} \frac{l_i^2 - l_j^2}{m_i^2 - m_j^2} \Pi_i \frac{l_i}{m_i}$
$kL_1 + L_2, \, k \geq 1$	$(k+1)(k + 2n - 1)$	$\Pi_{i<j} \frac{l_i^2 - l_j^2}{m_i^2 - m_j^2} \Pi_i \frac{l_i}{m_i}$

TABLE C.3

Eigenvalues and their multiplicity of S^3.

λ	eigenvalues	multiplicity
$kL_1, \, k \geq 1$	$k(k+2)$	$(k+1)^2$
$kL_1 + L_2, \, k \geq 1$	$(k+1)^2$	$k(k+2)$
$kL_1 - L_2, \, k \geq 1$	$(k+1)^2$	$k(k+2)$

TABLE C.4

Eigenvalues and their multiplicity of S^2.

λ	eigenvalues	multiplicity
$kL_1 \, k \geq 1$	$k(k+1)$	$2(2k+1)$