

Hodge Laplacians and Hodge Diffusion Maps

ALVARO ALMEIDA GOMEZ¹ AND JORGE DUQUE FRANCO²

Abstract

We introduce Hodge Diffusion Maps, a novel manifold learning algorithm designed to analyze and extract topological information from high-dimensional data-sets. This method approximates the exterior derivative acting on differential forms, thereby providing an approximation of the Hodge Laplacian operator. Hodge Diffusion Maps extend existing non-linear dimensionality reduction techniques, including vector diffusion maps, as well as the theories behind diffusion maps and Laplacian Eigenmaps. Our approach captures higher-order topological features of the data-set by projecting it into lower-dimensional Euclidean spaces using the Hodge Laplacian. We develop a theoretical framework to estimate the approximation error of the exterior derivative, based on sample points distributed over a real manifold. Numerical experiments support and validate the proposed methodology.

Keywords: Machine learning, Pattern recognition, Dimensionality reduction, Diffusion maps, Hodge Theory, Hodge Laplacians, Exterior derivative.

Mathematics Subject Classification: Primary: 68P05, 68T10, 68T45, 68W25; Secondary: 20G10, 35J05, 58J35, 58A14 .

1 Introduction

Dimensionality reduction is an essential technique for analyzing complex, high-dimensional datasets. It helps uncover important patterns and structures while overcoming the challenges of the **curse of dimensionality**. One popular non-linear dimensionality reduction method is Diffusion Maps (DM) [CL06, Laf04], a graph-based kernel method. Diffusion Maps captures the intrinsic geometry of data through a nonlinear embedding by using diffusion processes on a graph. This approach measures local connectivity between data points, revealing both local and global structures. The method is based on the manifold learning assumption, which assumes that the dataset consists of sample points distributed over a smooth manifold, and uses the Laplace-Beltrami operator to capture the topological information of the data through the diffusion process.

¹Universidad de Chile, Centro de Modelamiento Matemático (CNRS IRL2807), Beaucheff 851, Santiago, Chile, alvaroalmeidagomez182@gmail.com

²Universidad de Chile, Departamento de Matemáticas, Campus Juan Gómez Millas, Las Palmeras 3425, Santiago, Chile, jorge.duque@algebraicgeometry.cl

Vector Diffusion Maps (VDM) [SW12] extend the theory of Diffusion Maps by replacing real-valued function weights with vector-valued functions. This approach captures connectivity by considering linear orthogonal transformations that encode changes of basis between tangent spaces at different data points, while simultaneously approximating parallel transport. By incorporating these geometric relationships, VDM extract richer structural information from the dataset. The theory of VDM has been applied in various fields, including cryo-electron microscopy [SS11, TSL23]. The methodology is rooted in the connection Laplacian, which operates on vector fields and is approximated using a discrete formulation of the connection Laplacian operator. This operator is related to the first-order Hodge Laplacian via the Weitzenböck identity.

The k -th Hodge Laplacian generalizes the Laplace-Beltrami operator and plays a crucial role in capturing the topology of a manifold through the k -th De Rham cohomology group. Recent studies have extended the Hodge Laplacian to graphs and combinatorics, applying it to various fields such as ranking, game theory [RGWC⁺24, JLYY11, Lim20], and biomolecular structure analysis [WWLX22]. In this paper, we aim to approximate the Hodge Laplacian, defined over a real manifold.

While Vector Diffusion Maps leverage the first-order Hodge Laplacian to extract geometric information from data, an open question remains, as posed by [SW12]: How can higher-order geometric structures of a dataset be captured using the Hodge Laplacian? One of the main goals of this paper is to address this question.

In this work, we introduce **Hodge Diffusion Maps (HDM)**, a novel extension of Vector Diffusion Maps that utilizes the k -th order Hodge Laplacian for any $k \geq 1$. This approach overcomes the limitations of traditional methods such as Diffusion Maps and Vector Diffusion Maps, which primarily focus on low-order geometric features and often miss important higher-order structures. HDM provides a powerful framework for nonlinear dimensionality reduction that preserves the intrinsic geometry of the data. By projecting the dataset onto the dot product of the leading eigenforms of the Hodge Laplacian, our method captures meaningful geometric patterns and reveals the underlying structure of the data at multiple scales.

In this paper, we first construct the Hodge Laplacian over a Riemannian manifold by inferring the exterior derivative operator on differential forms. The main technical contribution is the approximation of the exterior derivative, defined over differential forms, using sample points distributed on an unknown manifold with an unknown intrinsic geometry. We provide analytical estimates for this approximation, which in turn allow us to approximate the Hodge Laplacian operator. This approximation generalizes the gradient operator approximation presented in [GNZ23], which uses asymmetric kernels to infer the diffusion properties of the dataset [GNZ21, HHYH23, HHS⁺23]. We summarize the key contributions of this work as follows:

- We propose an approximation of the exterior derivative operator based on sample point distributions, with the construction of the approximation independent of the dataset’s distribution. Error bound estimates for this approxi-

mation are provided in [Theorem 3.1](#).

- Based on this exterior derivative approximation, we construct a sample-based approximation of the Hodge Laplacian operator acting on differential forms defined over the manifold representing the dataset.
- Using the approximation of the Hodge Laplacian, we introduce the Hodge Diffusion Maps algorithm, which projects the dataset onto the dot products of the eigenforms of the Hodge Laplacian. This methodology extends the Vector Diffusion Maps algorithm, which is defined over first-order differential forms, to higher-order differential forms for $k \geq 1$.

The paper is structured as follows: In [Section 2](#), we briefly review the theory of Vector Diffusion Maps and the Hodge theory for real manifolds. [Section 3](#) details how to approximate the exterior derivative operator using a sample collection of points, as presented in [Theorem 3.1](#). In [Section 4](#), we apply [Theorem 3.1](#) to give a matrix-based approximation of the exterior derivative operator and provide the numerical implementation in [Algorithm 2](#). [Section 5](#) builds upon the results in [Section 4](#) to compute the Hodge Laplacian operator and define the Hodge Diffusion Maps and the Hodge Diffusion distance. [Section 6](#) presents numerical experiments on synthetic data, comparing the proposed methodology against Diffusion Maps, PCA, and t-SNE algorithms. [Section 7](#) presents the conclusions of the paper, highlighting future directions and potential applications of the proposed methodology. Finally, [Appendices A](#) and [B](#) provide the technical details related to the proof of the main result in [Theorem 3.1](#).

2 Preliminaries

Throughout this paper, we denote by \mathcal{M} a closed (i.e., compact without boundary) Riemannian manifold of dimension d , embedded in the ambient space \mathbb{R}^n . For a detailed exposition of diffusion map theory, we refer the reader to [\[CL06, Laf04\]](#), and for an introduction to Hodge Laplacians, to [\[War83\]](#).

2.1 Diffusion Maps

We briefly explain the Diffusion Maps and Vector Diffusion Maps algorithms. Given a set of data points, $X = \{x_1, x_2, \dots, x_N\} \subseteq \mathcal{M}$, the algorithm follows these steps:

First, we create a graph by measuring the similarity between pairs of data points. In classical diffusion maps, the weight g_{ij} of the edge between points x_i and x_j is calculated using the Gaussian Kernel:

$$g_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon}\right)$$

where ϵ is a scaling parameter. In the case of vector diffusion maps, the weights \bar{g}_{ij} are calculated using an orthogonal transformation ORT_{ij} along with the Gaussian kernel:

$$\bar{g}_{ij} = ORT_{ij} g_{ij}$$

The next step is to normalize the similarity matrix. In classical diffusion maps, the normalization is given by:

$$P_{ij} = \frac{g_{ij}}{\sum_k g_{ik}}$$

In vector diffusion maps, the normalization is:

$$P_{ij} = ORT_{ij} \frac{g_{ij}}{\sum_k g_{ik}}$$

This matrix P represents the probabilities of moving from one point to another in the diffusion process. In the final step, the algorithm then computes the eigenvalues and eigenvectors of the matrix P . Let $\phi_1, \phi_2, \dots, \phi_k$ be the eigenvectors corresponding to the largest eigenvalues. The diffusion maps (and vector diffusion maps) project the data points into a lower-dimensional space based on the entries of these eigenvectors:

$$\psi_i = (\lambda_1 \phi_1(i), \lambda_2 \phi_2(i), \dots, \lambda_k \phi_k(i))$$

where λ_i are the eigenvalues and $\phi_i(i)$ is the i -th entry of the eigenvector ϕ_i . This process reduces the data's dimensionality while maintaining its intrinsic geometric structure. Additionally, the matrices P_{ij} approximate the Laplace-Beltrami and Connection Laplacian operators.

In comparison with classical diffusion maps and vector diffusion maps, our proposed method considers the spectral decomposition of matrices of the form:

$$P_{ij} = \frac{1}{\sum_k g_{ik}} \det [g_{ij} L_i(x_i - x_j)^T, L_{ij}]$$

where L_i and L_{ij} are linear transformations depending on the indices i and j , respectively. We use this spectral decomposition to extract topological information from the dataset. Additionally, we show that the form of these matrices can approximate the exterior derivative operator, which is explained in more detail in [Section 4](#).

2.2 Hodge Laplacians

The fundamental object in the Hodge theory for a real manifolds \mathcal{M} are the Hodge Laplacians, sometimes also called Laplace-de Rham operator. Hodge Laplacians Δ^k are linear operators defined over the set k -differential forms $\Omega^k(\mathcal{M})$. The set of Hodge Laplacians generalizes the Laplace-Beltrami operator Δ in the sense that for $k = 0$, the two notions of Laplacian coincide $\Delta^0 = \Delta$, up to a sign. The importance of these operators lies in the fact that their kernels correspond to algebraic invariants

that encode geometric and topological information about the manifold. Let us briefly recall the definition of the Hodge Laplacian. Let (\mathcal{M}, g) be an oriented Riemannian manifold of dimension d . A Riemannian metric on a smooth manifold is a smooth assignment of an inner product to each tangent space. The Riemannian metric g induces an isomorphism $T_x\mathcal{M} \simeq T_x^*\mathcal{M}$ for every $x \in \mathcal{M}$, allowing the inner product in $T_x\mathcal{M}$ to be naturally transferred to $T_x^*\mathcal{M}$. This inner product in $T_x^*\mathcal{M}$ extends to the exterior algebra $\bigwedge^k T_x^*\mathcal{M}$ via the determinant:

$$\langle w_1 \wedge \cdots \wedge w_k, v_1 \wedge \cdots \wedge v_k \rangle := \det[\langle w_i, v_j \rangle], \quad w_i, v_j \in T_x^*\mathcal{M}. \quad (1)$$

Using the metric and the orientation, one defines the Hodge star operator

$$\star : \Omega^k(\mathcal{M}) \rightarrow \Omega^{d-k}(\mathcal{M})$$

which, for a k -differential form ω , is uniquely determined by the relation

$$\eta \wedge (\star \omega) = \langle \eta, \omega \rangle dVol$$

for every k -differential form η , where $\langle \eta, \omega \rangle$ is the pointwise inner product defined by Equation (1) and $dVol$ is the volume form induced by the Riemannian metric g . The adjoint of the exterior derivative, $\mathbf{d}_k^* : \Omega^k(\mathcal{M}) \rightarrow \Omega^{k-1}(\mathcal{M})$, is given by

$$\mathbf{d}_k^* := (-1)^{d(k+1)+1} \star \mathbf{d}_{d-k} \star .$$

The Hodge Laplacian is then defined as

$$\Delta^k := \mathbf{d}_{k+1}^* \mathbf{d}_k + \mathbf{d}_{k-1} \mathbf{d}_k^*$$

which is an endomorphism of $\Omega^k(\mathcal{M})$. The Hodge Laplacian provides important information about the cohomology elements, which intuitively correspond to k -dimensional holes. This follows from the Hodge theorem, which states that the space of k -harmonic forms

$$\mathcal{H}^k(\mathcal{M}) := \ker(\Delta^k)$$

is isomorphic to the k -th de Rham cohomology group:

$$H_{dR}^k(\mathcal{M}) \simeq \mathcal{H}^k(\mathcal{M}).$$

Thus, the Hodge Laplacian, which is constructed using the exterior derivative \mathbf{d}_k , encodes topological information about the manifold. In this paper, we focus on inferring the exterior derivative from an observable set of sample points distributed over the manifold.

3 Exterior derivative approximation

In this section, we extend the Diffusion Maps method from smooth functions to k -differential forms, aiming to approximate the exterior derivative operator on a manifold \mathcal{M} using sample points from \mathcal{M} . Our approach builds upon and generalizes the gradient estimation introduced in [GNZ23]. We assume that \mathcal{M} is a compact, d -dimensional Riemannian submanifold of \mathbb{R}^n without boundary, where the Riemannian metric on \mathcal{M} is induced by the ambient space.

3.1 Differential forms and differential arrays

We recall that a k -differential form is a smooth section $\omega : \mathcal{M} \rightarrow \wedge^k T^*\mathcal{M}$ defined from the manifold \mathcal{M} to the k -th exterior power of the cotangent bundle $T^*\mathcal{M}$, such that for every $x \in \mathcal{M}$, the value of ω at x is an element $\omega_x \in \wedge^k T_x^*\mathcal{M}$. In other words, at each point $x \in \mathcal{M}$, ω_x is a linear functional

$$\omega_x : \bigwedge^k T_x\mathcal{M} \rightarrow \mathbb{R},$$

or equivalently, an k -alternating form on the tangent space $T_x\mathcal{M}$,

$$\omega_x : \underbrace{T_x\mathcal{M} \times \cdots \times T_x\mathcal{M}}_{k\text{-times}} \rightarrow \mathbb{R}.$$

See Definition 1 for further details. The set of all k -differential forms on \mathcal{M} is denoted by $\Omega^k(\mathcal{M})$. Now, consider a local coordinate system (v_1, \dots, v_d) on \mathcal{M} . In these coordinates, any k -differential form $\omega \in \Omega^k(\mathcal{M})$ can be expressed as

$$\omega = \sum_I a_I dv_{i_1} \wedge \cdots \wedge dv_{i_k}.$$

The exterior derivative $\mathbf{d}_k : \Omega^k(\mathcal{M}) \rightarrow \Omega^{k+1}(\mathcal{M})$, acts on differential forms, and in these coordinates, it is given by

$$\mathbf{d}_k \omega = \mathbf{d}_k \left(\sum_I a_I dv_{i_1} \wedge \cdots \wedge dv_{i_k} \right) = \sum_I \sum_j \frac{\partial a_I}{\partial v_j} dv_j \wedge dv_{i_1} \wedge \cdots \wedge dv_{i_k},$$

While differential forms are abstract mathematical objects, we emphasize that, for computational purposes, we will represent them using alternating arrays, see Definition 2. From Proposition A.1, we know that $\wedge^k(T_x^*\mathcal{M}) \simeq \Theta^k(T_x\mathcal{M})$, allowing us to introduce the following space:

$$\Theta^k(T\mathcal{M}) = \bigsqcup_{x \in \mathcal{M}} \Theta^k(T_x\mathcal{M})$$

which admits a vector bundle structure such that $\wedge^k(T^*\mathcal{M}) \simeq \Theta^k(T\mathcal{M})$; see [Lee12, Lemma 10.6] for technical details. A section $W : \mathcal{M} \rightarrow \Theta^k(T\mathcal{M})$, called a k -differential array, is, via this isomorphism, simply a differential form on \mathcal{M} . We

denote the space of k -differential arrays on \mathcal{M} by $\Theta^k(\mathcal{M})$, which corresponds to the space of differential forms $\Omega^k(\mathcal{M})$, where each W_x is regarded as a k -alternating array in $T_x\mathcal{M}$. See [Appendix A](#) for a detailed explanation of this identification.

From now on, we will use differential arrays, but the reader should keep in mind that they are fundamentally differential forms, represented in a way that is more suitable for numerical computations.

3.2 Approximation of the exterior derivative using sample points

The goal of this section is to approximate the exterior derivative using sample points distributed on a manifold \mathcal{M} according to a smooth density $q(x)$. To achieve this, we first consider the heat kernel on $\mathcal{M} \times \mathcal{M}$, given by

$$e^{-\frac{\|y-x\|^2}{2t^2}},$$

where $\|\cdot\|$ denotes the Euclidean norm. We then introduce the following normalization:

$$d_t(x) = \int_{\mathcal{M}} e^{-\frac{\|y-x\|^2}{2t^2}} q(y) dVol(y), \quad (2)$$

where $dVol$ is the volume form induced by the Riemannian metric. Using this, we define the asymmetric vector-valued kernel

$$K_t(x, y) = (y - x) \frac{e^{-\frac{\|y-x\|^2}{2t^2}}}{d_t(x)}. \quad (3)$$

For fixed x, y, t , each vector $K_t(x, y) \in \mathbb{R}^n$ is a 1-differential array and, under the respective identification, can also be regarded as a 1-alternating array. Consequently, for any $W \in \Theta^k(\mathcal{M})$, it makes sense to consider

$$K_t(x, y) \wedge W(x)(n_1, n_2, \dots, n_{k+1})$$

for fixed $x \in \mathcal{M}$, which defines a $(k+1)$ -alternating arrays on \mathbb{R}^n . Now, for every differential array $W \in \Theta^k(\mathcal{M})$, we define a differential array $\mathbf{P}_t W \in \Theta^{k+1}(\mathbb{R}^n)$ by

$$\mathbf{P}_t W(x)(n_1, n_2, \dots, n_{k+1}) = \int_{\mathcal{M}} (K_t(x, y) \wedge (W(y) - W(x)))(n_1, n_2, \dots, n_{k+1}) q(y) dVol(y),$$

where $1 \leq n_i \leq n$. Here, the integral is well-defined since for fixed x and (n_1, \dots, n_{k+1}) ,

$$(K_t(x, y) \wedge (W(y) - W(x)))(n_1, n_2, \dots, n_{k+1})$$

can be regarded as a real-valued function of y defined on \mathcal{M} . Note that in this integral, we interpret $W(y)$ as being defined in the same space as $W(x)$, namely on $T_x\mathcal{M}$. Thus, \mathbf{P}_t defines a linear operator on $\Theta^k(\mathcal{M})$. For any $W \in \Theta^k(\mathcal{M})$ and $x \in \mathcal{M}$, we denote

$$\mathbf{P}_t W(x) = \int_{\mathcal{M}} K_t(x, y) \wedge (W(y) - W(x)) q(y) dVol(y) \quad (4)$$

Remark 3.1. $\mathbf{P}_t W(x)$ defines a $(k+1)$ -alternating array on \mathbb{R}^n but not necessary a $(k+1)$ -alternating array on $T_x \mathcal{M}$. However, according to [Remark A.2](#), its orthogonal projection onto $\wedge^{k+1} T_x \mathcal{M}$, denoted by

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_t W(x))$$

defines a $(k+1)$ -alternating form on $T_x \mathcal{M}$.

The following theorem establishes the relation between the operator \mathbf{P}_t and the exterior derivative over k -differential arrays.

Theorem 3.1. *Let $W \in \Theta^k(\mathcal{M})$ be a k -differential array, and let $x \in \mathcal{M}$. For any δ satisfying*

$$\frac{1}{2} < \delta < 1 - \frac{d}{2(d+2)} < 1. \quad (5)$$

where d is the dimension of \mathcal{M} , the following estimate holds:

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_t W(x)) = t^2 (\mathbf{d}_k W(x)) + O(t^f)$$

where the exponent f is given by

$$\min \{4\delta - 2, 2(1 - \delta)(d + 2)\}.$$

In particular, taking the limit as $t \rightarrow 0^+$, we obtain

$$\lim_{t \rightarrow 0^+} \frac{1}{t^2} \mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_t W(x)) = \mathbf{d}_k W(x). \quad (6)$$

First, we note that the set of values for δ satisfying [Equation \(5\)](#) is nonempty. Indeed, since $0 < \frac{d}{2(d+2)} < \frac{1}{2}$, it follows that

$$\frac{1}{2} < 1 - \frac{d}{2(d+2)} < 1.$$

[Equation \(6\)](#) provides a method for estimating the exterior derivative $\mathbf{d}_k W$ based on a set of sample points observed on the manifold. Specifically, for small $t > 0$, we have

$$\frac{1}{t^2} \mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_t W(x)) \approx \mathbf{d}_k W(x).$$

By the Law of Large Numbers (LLN), the operator \mathbf{P}_t can be approximated as:

$$\mathbf{P}_t W(x) \approx \sum_{i=1}^N \bar{K}_t(x, x_i) \wedge (W(x_i) - W(x)),$$

where

$$\bar{K}_t(x, y) = (y - x) \frac{e^{-\frac{\|y-x\|^2}{2t^2}}}{\bar{d}_t(x)}, \quad \text{and} \quad \bar{d}_t(x) = \sum_{i=1}^N e^{-\frac{\|x_i-x\|^2}{2t^2}}. \quad (7)$$

Consequently, the exterior derivative at x_j can be estimated as:

$$\mathbf{d}_k W(x_j) \approx \mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}} \left(\frac{1}{t^2} \sum_{i=1}^N \bar{K}(x_j, x_i) \wedge (W(x_i) - W(x_j)) \right) \quad (8)$$

Observe that, by the Law of Large Numbers, both terms \mathbf{P}_t and $\bar{d}_t(x)$ should have an average factor of $1/N$, but the variable N cancels out in the division involved in approximating the exterior derivative $\mathbf{d}_k W(x_j)$.

Notably, the right-hand side of Equation (8) does not depend on the distribution $q(x)$. Instead, it is computed purely from the dataset x_1, x_2, \dots, x_N . This result enables the estimation of the exterior derivative independently of the underlying distribution $q(x)$, making it a robust method for data-driven differential analysis.

4 Matrix-based computations

In this section, we derive a matrix representation for the approximation of the exterior derivative given in Equation (8). To achieve this, we first express the space of k -differential arrays in matrix form. The section is structured as follows: in Section 4.1, we present the matrix representation of the set of k -differential arrays $\Theta^k(\mathcal{M})$. Then, in Section 4.2, we present a matrix formulation for computing the exterior derivative of a k -differential array. This matrix-based approach provides a practical framework for numerical implementation and analysis.

4.1 Matrix representation of differential arrays

A key question in the proposed approach is how to reconstruct the space of k -differential arrays, denoted by $\Theta^k(\mathcal{M})$, from a finite set of N sample points $X = \{x_1, x_2, \dots, x_N\}$. These points are realizations of N independent and identically distributed (i.i.d.) random variables X_1, X_2, \dots, X_N drawn from a smooth $q(\cdot)$ over an **unknown** d -dimensional manifold \mathcal{M} . To address this, we first describe the local construction of k -differential arrays. Given a k -differential array W , its evaluation at x_i can be expressed as

$$W(x_i) = \sum_J f_J(x_i) O_J(x_i), \quad (9)$$

where the sum is taken over all k -tuples $J = (j_1, j_2, \dots, j_k)$ with $1 \leq j_1 < j_2 < \dots < j_k \leq d$. Here, $f_J(x_i)$ are real-value function, and $O_J(x_i)$ is the orthonormal basis of $\Theta^k(T_{x_i} \mathcal{M})$ defined as the wedge product

$$O_J(x_i) = \frac{1}{\sqrt{k!}} O_{j_1}(x_i) \wedge \dots \wedge O_{j_k}(x_i), \quad (10)$$

where $\{O_1(x_i), \dots, O_d(x_i)\}$ is an orthonormal basis for the tangent space $T_{x_i}\mathcal{M}$. This orthonormal basis is constructed using the Local PCA methodology described in [SW12] and [SW11]. The local PCA algorithm computes an orthonormal basis for the tangent space $T_{x_i}\mathcal{M}$ and the dimension of the manifold \mathcal{M} as follows:

1. **Neighborhood Selection:** Given a positive parameter r , consider the set of points $\{x_{i_1}, x_{i_2}, \dots, x_{i_l}\}$ that lie within the local neighborhood

$$U(x_i, r) = \{y \in \mathcal{M} \mid \|x_i - y\|_{\mathbb{R}^n} < r\}.$$

2. **Matrix construction:** Define the Matrix M_{x_i} as

$$M_{x_i} = \left[(x_{i_1} - x_i) e^{\frac{\|x_{i_1} - x_i\|^2}{2t^2}}, (x_{i_2} - x_i) e^{\frac{\|x_{i_2} - x_i\|^2}{2t^2}}, \dots, (x_{i_l} - x_i) e^{\frac{\|x_{i_l} - x_i\|^2}{2t^2}} \right].$$

3. **Estimating the Intrinsic Dimension:** Let $\sigma_1, \sigma_2, \dots, \sigma_{\min(i_l, n)}$ denote the singular values of M_{x_i} . We introduce a threshold parameter $\gamma \in (0, 1)$, typically set to $\gamma \approx 0.9$, and define the intrinsic dimension d_i at x_i as the largest integer satisfying

$$\frac{\sum_{j=1}^{d_i} \|\sigma_j\|}{\sum_{j=1}^{i_l} \|\sigma_j\|} < \gamma.$$

The parameter d_i provides an estimate of the dimension of the local tangent space at x_i . The intrinsic dimension d of the manifold \mathcal{M} is then obtained as the median of all local estimates:

$$d = \text{median}(d_1, d_2, \dots, d_N).$$

4. **Extracting the Tangent Space Basis:** Compute the singular value decomposition (SVD) of M_{x_i} , and take the first d left-singular vectors $O_1(x_i), \dots, O_d(x_i)$ as an orthonormal basis for the tangent space $T_{x_i}\mathcal{M}$.

The local PCA algorithm allows to express a k -differential array W in matrix form as

$$\mathbf{O}_k * \mathbf{f}$$

where $*$ denotes the standard matrix multiplication, and the matrices \mathbf{O}_k and \mathbf{f} are defined as follows:

Definition of \mathbf{f} : The matrix \mathbf{f} consists of N blocks, each of size $\binom{d}{k} \times 1$. The i -th block is given by

$$\mathbf{f}(i) = \begin{bmatrix} f_{J_1}(x_i) \\ f_{J_2}(x_i) \\ \vdots \\ f_{J_{\binom{d}{k}}}(x_i) \end{bmatrix}$$

where the multi-indexes $J_1, J_2, \dots, J_{\binom{d}{k}}$ correspond to all possible k -tuples

$$J_l = (j_1^l, \dots, j_k^l) \text{ with } 1 \leq j_1^l < \dots < j_k^l \leq d.$$

Thus, the full matrix \mathbf{f} has size $\binom{d}{k}N \times 1$.

Definition of \mathbf{O}_k : The matrix \mathbf{O}_k consists of $N \times N$ blocks, each of size $n^k \times \binom{d}{k}$. The block at position (i, j) is defined as

$$\mathbf{O}_k(i, j) = \begin{cases} \overline{O}_k(i) & \text{if } i = j \\ 0_{n^k \times \binom{d}{k}}, & \text{if } i \neq j \end{cases}$$

for $i, j \in \{1, \dots, N\}$. Here, $\overline{O}_k(i)$ is the matrix

$$\overline{O}_k(i) = \begin{bmatrix} O_{J_1}(x_i) & O_{J_2}(x_i) & \dots & O_{J_{\binom{d}{k}}}(x_i) \end{bmatrix},$$

where each $O_{J_l}(x_i)$ is defined as in Equation (10) and is considered as a column vector embedded in \mathbb{R}^{n^k} . Overall, \mathbf{O}_k has size $n^k N \times \binom{d}{k} N$. The values of the k -differential array $W(x_i)$ correspond to the i -th block of the product $\mathbf{O}_k * \mathbf{f}$.

4.2 Matrix-based computation of the exterior derivative

In this section, we derive a matrix expression for the approximation of the exterior derivative in k -differential arrays, as given in Equation (8), using the results from the previous section. According to Equation (9), any k -differential array W at the point x_j can be written as:

$$W(x_j) = \sum_J f_J(x_j) O_J(x_j).$$

We denote by $O(x_j)_{n \times d}$ the matrix whose columns form a basis for the tangent space $T_{x_j}\mathcal{M}$ given by

$$O_1(x_j), O_2(x_j), \dots, O_d(x_j).$$

Next, the projection of this basis onto the tangent space at the point x_i , denoted $T_{x_i}\mathcal{M}$, is given by the matrix product:

$$O(x_i)O(x_i)^T O(x_j).$$

Let \mathcal{P}_V denote the orthogonal projection onto the space V . Then, $O_J(x_j)$ decomposes as

$$\begin{aligned} O_J(x_j) &= \frac{1}{\sqrt{k!}} O_{j_1}(x_j) \wedge \dots \wedge O_{j_k}(x_j) \\ &= \frac{1}{\sqrt{k!}} \mathcal{P}_{T_{x_i}\mathcal{M}} O_{j_1}(x_j) \wedge \dots \wedge \mathcal{P}_{T_{x_i}\mathcal{M}} O_{j_k}(x_j) + \xi, \end{aligned}$$

where ξ consists of wedge product terms in which one factor of each wedge product belongs to the orthogonal complement $T_{x_i}\mathcal{M}^\perp$. Furthermore, the projected term

$$\mathcal{P}_{T_{x_i}\mathcal{M}} O_{j_1}(x_j) \wedge \dots \wedge \mathcal{P}_{T_{x_i}\mathcal{M}} O_{j_k}(x_j)$$

can be written as:

$$\sum_L \det(O_L^T(x_i)O^J(x_j))O_{l_1}(x_i) \wedge O_{l_2}(x_i) \wedge \cdots \wedge O_{l_k}(x_i),$$

where $O_L^T(x_i)$ denotes the submatrix of $O(x_i)^T$ formed by the rows indexed by $L = (l_1, \dots, l_k)$, while $O^J(x_j)$ is the submatrix of $O(x_j)$ consisting of the columns indexed by $J = (j_1, \dots, j_k)$. By Combining this with Equation (9), we obtain that the orthogonal projection onto $\wedge^k T_{x_i}\mathcal{M}$ is given by

$$\mathcal{P}_{\wedge^k T_{x_i}\mathcal{M}}W(x_j) = \sum_J f_J(x_j) \sum_L \det(O_L^T(x_i)O^J(x_j))O_L(x_i) \quad (11)$$

The previous equation helps to implement Theorem 3.1 as follows. According to Equation (8), we can approximate the exterior derivative $\mathbf{d}_k(W)(x_i)$ as

$$\begin{aligned} \frac{1}{t^2} \mathcal{P}_{\wedge^{k+1} T_{x_i}\mathcal{M}} \left(\sum_{j=1}^N \bar{K}_t(x_i, x_j) \wedge (W(x_j) - W(x_i)) \right) = \\ \frac{1}{t^2} \sum_{j=1}^N \left(\mathcal{P}_{T_{x_i}\mathcal{M}}(\bar{K}_t(x_i, x_j)) \wedge (\mathcal{P}_{\wedge^k T_{x_i}\mathcal{M}}(W(x_j) - W(x_i))) \right) \end{aligned} \quad (12)$$

Observe that by definition of the kernel $\bar{K}_t(x_j, x_i)$ as in Equation (7):

$$\mathcal{P}_{T_{x_i}\mathcal{M}}(\bar{K}_t(x_i, x_j)) = \frac{1}{d_t(x_i)} e^{\frac{-\|x_i - x_j\|^2}{2t^2}} \sum_{s=1}^N \langle x_j - x_i, O_s(x_i) \rangle O_s(x_i). \quad (13)$$

Using the Laplace expansion of the determinant and the identity

$$O_s(x_i) \wedge O_L(x_i) = \frac{1}{\sqrt{k!}} O_s(x_i) \wedge O_{l_1}(x_i) \wedge \cdots \wedge O_{l_k}(x_i) = \sqrt{k+1} O_{(s, l_1, \dots, l_k)}(x_i),$$

we obtain the following expression:

$$\begin{aligned} \sum_{s=1}^N \langle x_j - x_i, O_s(x_i) \rangle O_s(x_i) \wedge \sum_L \det(O_L^T(x_i)O^J(x_j))O_L(x_i) = \\ \sqrt{k+1} \sum_M \det([A_M(i, j), O_M^T(x_i)O^J(x_j)])O_M(x_i) \end{aligned} \quad (14)$$

where the sum runs over all $k+1$ -tuples $M = (m_1, m_2, \dots, m_{k+1})$ satisfying $1 \leq m_1 < \cdots < m_{k+1} \leq d$. Here $A(i, j)$ is the column vector defined by

$$A(i, j) = e^{\frac{-\|x_i - x_j\|^2}{2t^2}} O(x_i)^T(x_j - x_i)$$

and $A_M(i, j)$ is the submatrix of $A(i, j)$ consisting of the rows indexed by M . Additionally,

$$[A_M(i, j), O_M^T(x_i)O^L(x_j)] \quad (15)$$

denotes the concatenated matrix whose first column is $A_M(i, j)$. Therefore, combining [Equations \(11\), \(13\) and \(14\)](#), we obtain the following wedge product identity:

$$\begin{aligned} \sum_{j=1}^N \left(\mathcal{P}_{T_{x_i}\mathcal{M}}(\bar{K}_t(x_i, x_j)) \wedge (\mathcal{P}_{\wedge^k T_{x_i}\mathcal{M}} W(x_j)) \right) = \\ \sqrt{k+1} \frac{1}{\bar{d}_t(x_i)} \sum_{j=1}^N \sum_J f_J(x_j) \sum_M \det([A_M(i, j), O_M^T(x_i) O^J(x_j)]) O_M(x_i). \end{aligned} \quad (16)$$

Similarly

$$\begin{aligned} \sum_{j=1}^N \left(\mathcal{P}_{T_{x_i}\mathcal{M}}(\bar{K}_t(x_j, x_i)) \wedge W(x_i) \right) = \\ \sqrt{k+1} \frac{1}{\bar{d}_t(x_i)} \sum_{j=1}^N \sum_J f_J(x_i) \sum_M \det([A_M(i, j), O_M^T(x_i) O^J(x_i)]) O_M(x_i) \end{aligned} \quad (17)$$

Recall that [Equation \(12\)](#) provides an approximation of the exterior derivative $\mathbf{d}_k(W)(x_i)$. Note that, up to factor of $\frac{1}{t^2}$, [Equation \(12\)](#) corresponds to the difference between [Equation \(16\)](#) and [Equation \(17\)](#). Furthermore, [Equation \(16\)](#) represents the i -th block of the following matrix multiplication:

$$\sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{ED}_k^1 * \mathbf{f} \quad (18)$$

where \mathbf{ED}_k^1 is the block matrix

$$\mathbf{ED}_k^1(i, j) = \begin{bmatrix} ED^1(i, j, M_1, J_1) & ED^1(i, j, M_1, J_2) & \cdots & ED^1(i, j, M_1, J_{\binom{d}{k}}) \\ ED^1(i, j, M_2, J_1) & ED^1(i, j, M_2, J_2) & \cdots & ED^1(i, j, M_2, J_{\binom{d}{k}}) \\ \vdots & \vdots & \ddots & \vdots \\ ED^1(i, j, M_{\binom{d}{k+1}}, J_1) & ED^1(i, j, M_{\binom{d}{k+1}}, J_2) & \cdots & ED^1(i, j, M_{\binom{d}{k+1}}, J_{\binom{d}{k}}) \end{bmatrix} \quad (19)$$

with entries defined as

$$ED^1(i, j, M, J) = \frac{1}{\bar{d}_t(x_i)} \det([A_M(i, j), O_M^T(x_i) O^J(x_j)])$$

Similarly, [Equation \(17\)](#) represents the i -th block of the matrix multiplication.

$$\sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{ED}_k^2 * \mathbf{f} \quad (20)$$

where \mathbf{ED}_k^2 is the diagonal block matrix

$$\mathbf{ED}_k^2(i, j) = \begin{cases} \overline{ED}_k^2(i) & \text{if } i = j \\ 0_{\binom{d}{k+1} \times \binom{d}{k}}, & \text{if } i \neq j. \end{cases}$$

The block matrix $\overline{ED}_k^2(i)$ is given by

$$\overline{ED}_k^2(i) = \begin{bmatrix} ED^2(i, M_1, J_1) & ED^2(i, M_1, J_2) & \cdots & ED^2(i, M_1, J_{\binom{d}{k}}) \\ ED^2(i, M_2, J_1) & ED^2(i, M_2, J_2) & \cdots & ED^2(i, M_2, J_{\binom{d}{k}}) \\ \vdots & \vdots & \ddots & \vdots \\ ED^2(i, M_{\binom{d}{k+1}}, J_1) & ED^2(i, M_{\binom{d}{k+1}}, J_2) & \cdots & ED^2(i, M_{\binom{d}{k+1}}, J_{\binom{d}{k}}) \end{bmatrix} \quad (21)$$

where each block is defined by

$$\begin{aligned} ED^2(i, M, J) &= \frac{1}{d_t(x_i)} \sum_{l=1}^N \det([A_M(i, l), O_M^T(x_i) O^J(x_i)]) \\ &= \frac{1}{d_t(x_i)} \det(\sum_{l=1}^N A_M(i, l), O_M^T(x_i) O^J(x_i)) \end{aligned} \quad (22)$$

Now, recall that the k -differential array W can be express as

$$W = \mathbf{O}_k * \mathbf{f},$$

which implies that

$$\mathbf{f} = \mathbf{O}_k^T * W.$$

By combining Equations (12), (18) and (20), we obtain that the approximation of exterior derivative $\mathbf{d}_k(W)$ at the point x_i , is given by the i -th block of the matrix multiplication:

$$\frac{1}{t^2} \sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{ED}_k * \mathbf{f} = \frac{1}{t^2} \sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{ED}_k * \mathbf{O}_k^T * W \quad (23)$$

where \mathbf{ED}_k is defined as

$$\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2.$$

Thus, the matrix

$$\frac{1}{t^2} \sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{ED}_k * \mathbf{O}_k^T \quad (24)$$

represents the matrix approximation of the exterior derivative operator \mathbf{d}_k acting on k -differential arrays.

Note that the matrix \mathbf{O}_{k+1} has orthonormal columns and depends on the ambient space dimension n , whereas \mathbf{ED}_k encapsulates information about the manifold of dimension d . Consequently, \mathbf{ED}_k encodes more information about the intrinsic manifold through the exterior derivative \mathbf{d}_k .

Remark 4.1. An important observation is that if we choose a different orthonormal basis $O'_{j_1}(x_i), \dots, O'_{j_k}(x_i)$, the associated matrix \mathbf{ED}'_k , as given in Equation (24), is equivalent to \mathbf{ED}_k , in the following sense:

$$\mathbf{ED}'_k = ((\mathbf{O}'_{k+1})^T * \mathbf{O}_{k+1}) * \mathbf{ED}_k * (\mathbf{O}_k^T * \mathbf{O}'_k). \quad (25)$$

Here, the matrices $(\mathbf{O}'_{k+1})^T * \mathbf{O}_{k+1}$ and $\mathbf{O}_k^T * \mathbf{O}'_k$ are orthonormal, since the (i, i) blocks of \mathbf{O}_{k+1} , \mathbf{O}_{k+1} and \mathbf{O}_k , \mathbf{O}_k form orthonormal bases for $\Theta^{k+1}T_{x_i}\mathcal{M}$ and $\Theta^kT_{x_i}\mathcal{M}$, respectively. Therefore, the matrix \mathbf{ED}_k is unique, up to the change of basis induced by $(\mathbf{O}_{k+1}')^T * \mathbf{O}_{k+1}$ and $\mathbf{O}_k^T * \mathbf{O}'_k$.

4.3 Implementation of the Algorithm

In this section, we summarize the results from the previous sections and outline a practical algorithm for analyzing data sets using the matrix representation of the exterior derivative, as defined in Equation (24) in Section 4. The primary objective here is to compute the matrix $\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2$ as specified in Equation (24).

We assume that $X = \{x_1, x_2, \dots, x_N\}$ are sampled points, representing N independent and identically distributed (i.i.d.) random variables X_1, X_2, \dots, X_N , drawn from a smooth distribution $q(\cdot)$ over an unknown d -dimensional manifold \mathcal{M} .

The first step in the algorithm is to compute the tangent vectors $O_1(x_j), \dots, O_d(x_j)$ using the Local PCA method described in [SW12] and [SW11]. For this, we take as input the number K , which represents the total number of points in the open neighborhood of a point x , defined as:

$$U(x, r) = \{y \mid \|x - y\|_{\mathbb{R}^n} < r\} \quad (26)$$

In the implemented algorithm, the number K is the same for all points and does not depend on the indices i or j . Algorithm 1 summarizes the local PCA method, which is explained in Section 4.1.

Algorithm 1 Local PCA method

input Data-set $X = \{x_1, x_2, \dots, x_N\}$, and K , the number of points in the neighborhood $U(x_i, r)$ and scaling parameter t .

1. **for** $i = 1$ to N **do**

- Find the K -closest points to x_i , denoted $x_{i_1}, x_{i_2}, \dots, x_{i_K}$.
- Compute the matrix

$$M_{x_i} = \left[(x_{i_1} - x_i) e^{-\frac{\|x_{i_1} - x_i\|^2}{2t^2}}, (x_{i_2} - x_i) e^{-\frac{\|x_{i_2} - x_i\|^2}{2t^2}}, \dots, (x_{i_K} - x_i) e^{-\frac{\|x_{i_K} - x_i\|^2}{2t^2}} \right].$$

- Compute d_{x_i} , the rank of the matrix M_{x_i} .

2. **end for**

3. Let $d = \text{median}(d_1, d_2, \dots, d_N)$

4. **for** $i = 1$ to N **do**

- Let $O_1(x_i), \dots, O_d(x_i)$ be the d left singular vectors from the singular value decomposition of M_{x_i} .
- Compute the matrix $O(x_i)$ as

$$O(x_i) = [O_1(x_i), O_2(x_i), \dots, O_K(x_i)].$$

5. **end for**

return The orthonormal vectors $O_1(x_i), O_2(x_i), \dots, O_K(x_i)$ of the tangent space $T_{x_i}\mathcal{M}$, the matrix $O(x_i)$ and the dimension d of the manifold.

The next step in the proposed method is to compute the matrix $\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2$, as explained in [Section 4.2](#). Since the exterior derivative at point x_i depends only on information from the neighborhood $U(x, r)$ (see [Equation \(26\)](#)), we can reduce the number of points required to construct the matrices \mathbf{ED}_k^1 and \mathbf{ED}_k^2 , thereby lowering the computational complexity.

The key idea is that, for each index i , we compute the block matrices $\mathbf{ED}_k(i, j)$ only if x_j is among the K -nearest points to x_i . If x_j is not one of the K -nearest points, we set $\mathbf{ED}_k(i, j) = 0$. Similarly, when computing \mathbf{ED}_k^2 , we calculate the i -th block (as shown in [Equation \(22\)](#)) by summing over the K -nearest points to x_i . Specifically, we compute $ED^2(i, M, J)$ as:

$$ED^2(i, M, J) = \frac{1}{\overline{d_t}(x_i)} \det \left(\sum_{l=1}^K A_M(i, i_l), O_M^T(x_i) O^J(x_i) \right), \quad (27)$$

where i_1, \dots, i_K are the indices of the K -nearest points $x_{i_1}, \dots, x_{i_K} \in X$ to x_i . This simplification does not significantly affect the expression for $\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2$, since the exponential term

$$e^{-\frac{\|x_j - x_i\|^2}{2t^2}}$$

vanishes when x_j is far from x_i . The computation of the matrix $\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2$ is summarized in [Algorithm 2](#).

Algorithm 2 Computation of $\mathbf{ED}_k = \mathbf{ED}_k^1 - \mathbf{ED}_k^2$

input Data-set $X = \{x_1, x_2, \dots, x_N\}$, and K , the number of points in the neighborhood $U(x_i, r)$, scaling parameter t .

1. Apply [Algorithm 1](#) and assign $[O(x_i), d] \leftarrow \text{LocalPCA}(X, K)$
2. Initialize the array \mathbf{ED}_k as a zero matrix with $N \times N$ blocks.
3. **for** $i = 1$ to N **do**
 - Find the K -closest points to x_i , denoted $x_{i_1}, x_{i_2}, \dots, x_{i_K}$.
 - For all $1 \leq l \leq K$, compute the column vector $A(i, i_l)$ as in [Eq \(15\)](#).
 - **for** $l = 1$ to K **do**
 - **If** $i \neq i_l$
 - Assign the value $\mathbf{ED}_k(i, i_l) \leftarrow \mathbf{ED}_k^1(i, i_l)$ as shown in [Equation \(19\)](#).
 - **Else**
 - Assign the value $\mathbf{ED}_k(i, i) \leftarrow \overline{ED}_k^2(i)$ based on [Equations \(21\) and \(27\)](#).
 - **End If**
 - **end for**
4. **end for**

return The matrix \mathbf{ED}_k , which contains the intrinsic information of the exterior derivative.

5 Hodge Diffusion-Maps

In this section, we use the approximation of the exterior derivative provided in [Section 3](#) to construct a matrix approximation based on observable sample points of the Hodge Laplacian operator. Additionally, we define the Hodge diffusion maps and the Hodge diffusion distance, which are generalizations of the Vector Diffusion Maps methodology [\[SW12\]](#).

5.1 Hodge Laplacians approximation

Based on the results from [Sections 3 and 4](#), and specifically from [Equation \(24\)](#), we can approximate the exterior derivative \mathbf{d}_k on k -differential arrays using the matrix:

$$\frac{1}{t^2} \sqrt{k+1} \mathbf{O}_{k+1} * \mathbf{E} \mathbf{D}_k * \mathbf{O}_k^T$$

With this approximation, we can also approximate the Hodge Laplacian Δ_k , which is defined on k -differential arrays as:

$$\Delta_k = \mathbf{d}_{k+1}^* \circ \mathbf{d}_k + \mathbf{d}_{k-1} \circ \mathbf{d}_k^*,$$

(see [Section 2.2](#) for more details). It turns out that the matrix representation of the approximation of the adjoint \mathbf{d}_k^* of the exterior derivative corresponds to the transpose of the matrix representation of \mathbf{d}_{k-1} . In other words, this approximation is represented by the matrix:

$$\frac{1}{t^2} \sqrt{k} \mathbf{O}_{k-1} * \mathbf{E} \mathbf{D}_{k-1}^T * \mathbf{O}_k^T$$

From this, we have that the matrix representation of the Hodge Laplacian $\Delta_k(W)$ is given by

$$\frac{1}{t^4} \mathbf{O}_k * ((k+1) \mathbf{E} \mathbf{D}_k^T \mathbf{E} \mathbf{D}_k + k \mathbf{E} \mathbf{D}_{k-1} \mathbf{E} \mathbf{D}_{k-1}^T) * \mathbf{O}_k^T * W$$

where W is a k -differential array. Therefore, the intrinsic information of the Hodge-Laplacian Δ_k can be captured through the matrix:

$$\mathbf{H}_{k,t} = \frac{1}{t^4} ((k+1) \mathbf{E} \mathbf{D}_k^T \mathbf{E} \mathbf{D}_k + k \mathbf{E} \mathbf{D}_{k-1} \mathbf{E} \mathbf{D}_{k-1}^T) \quad (28)$$

We define $\mathbf{H}_{k,t}$ as the **Hodge-Laplacian** matrix of order k .

Similarly to the case of the exterior derivative, where the exterior derivative matrix $\mathbf{E} \mathbf{D}_k$ is unique up to equivalence between matrices (as shown in [Equation \(25\)](#)), the Hodge-Laplacian matrix $\mathbf{H}_{k,t}$ is also unique up to a similar equivalence. Specifically, for a different choice of the orthonormal basis $O'_{j_1}(x_i), \dots, O'_{j_k}(x_i)$, the corresponding **Hodge-Laplacian** matrix $\mathbf{H}'_{k,t}$, defined analogously to [Equation \(28\)](#) satisfies:

$$\mathbf{H}'_{k,t} = ((\mathbf{O}'_k)^T * \mathbf{O}_k) * \mathbf{H}_{k,t} * ((\mathbf{O}'_k)^T * \mathbf{O}_k)^T \quad (29)$$

where $(\mathbf{O}'_k)^T * \mathbf{O}_k$ is an orthonormal matrix.

5.2 Hodge Diffusion-Maps and Hodge Diffusion-Distance

As in the definition of affinity in vector diffusion maps [\[SW12, Page 1078\]](#), we use the Hodge-Laplacian matrix, $\mathbf{H}_{k,t}$, to define an affinity between two points, x_i and x_j .

In this section, we describe the type of affinity that the Hodge-Laplacian captures within the dataset.

Consider the matrix \mathbf{ED}_k as defined in Equation (24). This matrix is constructed by incorporating terms of the form

$$e^{\frac{-\|x_i - x_j\|^2}{2t^2}} \det \left(O(x_i)^T (x_j - x_i), O_M^T(x_i) O^J(x_j) \right),$$

where the exponential factor, $e^{\frac{-\|x_i - x_j\|^2}{2t^2}}$, encodes the local proximity between points x_i and x_j , reflecting their spatial relationship. The remaining factors in the determinant represent the area of the parallelogram spanned by the vectors $O(x_i)^T (x_j - x_i)$ and the matrix product $O_M^T(x_i) O^J(x_j)$, which accounts for the change of basis of the k tangent vectors at both x_i and x_j .

Thus, the block (i, j) of the \mathbf{ED}_k matrix quantifies both the proximity between x_i and x_j and the area of the parallelogram formed by these vectors.

By construction, the (i, j) block of the Hodge-Laplacian matrix measures the local connectivity between x_i and x_j , along with the geometric structure defined by the k and $k+1$ -dimensional change of basis vectors at each point, in relation to other points in the dataset.

To be more specific, this affinity is defined as the squared Frobenius norm of the \mathbf{tm} -power of the (i, j) -block of $\mathbf{H}_{k,t}$, i.e.,

$$\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2 = \text{Tr}(\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)^T * \mathbf{H}_{k,t}^{\mathbf{tm}}(i, j))$$

This affinity quantifies how information from the Hodge-Laplacian matrix propagates from x_i to x_j along a path of length \mathbf{tm} . Additionally, it reflects how concentrated the information from the \mathbf{tm} -th power of the **Hodge-Laplacian** is when passing information from the j -th node to the i -th node. Specifically, for any k -differential array W , the Hodge-Laplacian at the i -th point can be approximated as

$$\Delta_k^{\mathbf{tm}} W(x_i) \approx \sum_{j=1}^N O_k(x_i) * \mathbf{H}_{k,t}^{\mathbf{tm}}(i, j) * O_k^T(x_j) * W(x_j)$$

Thus, the norm $\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2$ is large when the differential array W at x_j plays a significant role computing the \mathbf{tm} -th power of the **Hodge-Laplacian** at the point x_i .

Remark 5.1. An important observation is that the affinity definition using the Frobenius norm $\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2$ is independent of the choice of orthonormal basis for the tangent space $T_{x_i}\mathcal{M}$

$$O_1(x_j), O_2(x_j), \dots, O_d(x_j).$$

Indeed, if $O'_1(x_j), O'_2(x_j), \dots, O'_d(x_j)$ is another orthonormal basis, and $\mathbf{H}'_{k,t}$ is the corresponding Hodge-Laplacian matrix, then, by Equation (29), the (i, j) -th block matrix of the \mathbf{tm} power of $\mathbf{H}_{k,t}$ and $\mathbf{H}'_{k,t}$ satisfy

$$(\mathbf{H}'_{k,t})^{\mathbf{tm}}(i,j) = A * \mathbf{H}_{k,t}^{\mathbf{tm}}(i,j) * B^T,$$

for some orthonormal matrices A and B . This implies that

$$\left((\mathbf{H}'_{k,t})^{\mathbf{tm}}(i,j)\right)^T * (\mathbf{H}'_{k,t})^{\mathbf{tm}}(i,j)$$

and

$$\mathbf{H}_{k,t}^{\mathbf{tm}}(i,j)^T * \mathbf{H}_{k,t}^{\mathbf{tm}}(i,j)$$

are similar matrices and therefore the same trace. Consequently, the Frobenius norms are equal, proving the claim.

We now define Hodge Diffusion Maps. By construction, the **Hodge-Laplacian** matrix $\mathbf{H}_{k,t}$ is symmetric and non-negative definite. Thus, by the spectral theorem, it admits a complete set of eigenvectors $b_1, b_2, \dots, b_{N^{(d)_k}}$ in $\mathbb{R}^{N^{(d)_k}}$, with corresponding non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{N^{(d)_k}}$. Each vector b_j is considered as a block vector, where each block has size $N \times 1$ and consists of a column vector of dimension $\binom{d}{k} \times 1$. We denote the i -th block of b_j by $b_j(i)$. Using this orthonormal eigenbasis, the (i,j) -th block of $\mathbf{H}_{k,t}^{\mathbf{tm}}$ can be written as

$$\mathbf{H}_{k,t}^{\mathbf{tm}}(i,j) = \sum_{l=1}^{N^{(d)_k}} \lambda_l^{\mathbf{tm}} b_l(i) \otimes b_l(j).$$

Consequently, the affinity measure $\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i,j)\|_F^2$ takes the form

$$\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i,j)\|_F^2 = \sum_{l_1, l_2=1}^{N^{(d)_k}} \lambda_{l_1}^{\mathbf{tm}} \lambda_{l_2}^{\mathbf{tm}} \langle b_{l_1}(i), b_{l_2}(i) \rangle \langle b_{l_1}(j), b_{l_2}(j) \rangle.$$

This representation allows us to define an embedding for the dataset. For $1 \leq m \leq N^{(d)_k}$, we define the truncated k -th **Hodge diffusion map** at time \mathbf{tm} and truncation level m , denoted by $\eta_{k,m}^{\mathbf{tm}}$, as the embedding that maps the dataset $X = \{x_1, x_2, \dots, x_N\} \subseteq \mathbb{R}^n$ into $\mathbb{R}^{m \times m}$, via the square matrix:

$$\eta_{k,m}^{\mathbf{tm}}(x_i) = \left[\sqrt{\lambda_{l_1}^{\mathbf{tm}}} \sqrt{\lambda_{l_2}^{\mathbf{tm}}} \langle b_{l_1}(i), b_{l_2}(i) \rangle \right]_{\substack{\mathbb{R}^{(d)_k} \\ 1 \leq l_1, l_2 \leq m}}. \quad (30)$$

Here, $\langle \cdot, \cdot \rangle_{\mathbb{R}^{(d)_k}}$ denotes the standard inner product in $\mathbb{R}^{(d)_k}$. The affinity between two points x_i and x_j can then be approximated as

$$\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i,j)\|_F^2 \approx \langle \eta_{k,m}^{\mathbf{tm}}(x_i), \eta_{k,m}^{\mathbf{tm}}(x_j) \rangle_F.$$

Based on the vector diffusion distance [SW12], which measures the connectivity of points using the connected Laplacian, we use the Hodge Laplacian to define the **Hodge Diffusion-Distance** d_{Hodge} between two points x_i and x_j as:

$$d_{\mathbf{Hodge}}^2(x_i, x_j) = \|\eta_{k,m}^{\mathbf{tm}}(x_i)\|_F^2 + \|\eta_{k,m}^{\mathbf{tm}}(x_j)\|_F^2 - 2\langle \eta_{k,m}^{\mathbf{tm}}(x_i), \eta_{k,m}^{\mathbf{tm}}(x_j) \rangle_F.$$

Although the embedding $\eta_{k,m}^{\mathbf{tm}}$ is computed using only the first m eigenvalues, an important question is when it provides a good approximation for the affinity measure $\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2$. Specifically, we are interested in when the error in the approximation, given by the absolute value of the difference

$$\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2 - \langle \eta_{k,m}^{\mathbf{tm}}(x_i), \eta_{k,m}^{\mathbf{tm}}(x_j) \rangle_F,$$

is small enough. In practice, this is not guaranteed, since the eigenvalues $\lambda_l^{\mathbf{tm}}$ for $l > m$ could still be large, especially if $\lambda_l > 1$. To address this issue, we normalize the affinity measure, the Hodge diffusion maps, and the **Hodge Diffusion-Distance** by the factor $1/\lambda_1^{\mathbf{tm}}$. Instead of considering $\|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2$, $\eta_{k,m}^{\mathbf{tm}}(x_i)$, and $d_{\mathbf{Hodge}}^2(x_i, x_j)$, we use their normalized counterparts:

$$\frac{1}{\lambda_1^{2\mathbf{tm}}} \|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2, \quad (31)$$

$$\frac{1}{\lambda_1^{\mathbf{tm}}} \eta_{k,m}^{\mathbf{tm}}(x_i), \quad (32)$$

and

$$\frac{1}{\lambda_1^{2\mathbf{tm}}} d_{\mathbf{Hodge}}^2(x_i, x_j).$$

With these normalizations, the error in the normalized embedding and the normalized affinity measure is bounded by:

$$\frac{1}{\lambda_1^{2\mathbf{tm}}} \left| \|\mathbf{H}_{k,t}^{\mathbf{tm}}(i, j)\|_F^2 - \langle \eta_{k,m}^{\mathbf{tm}}(x_i), \eta_{k,m}^{\mathbf{tm}}(x_j) \rangle_F \right| \leq \left(\frac{\lambda_{m+1}}{\lambda_1} \right)^{\mathbf{tm}} \left((N \binom{d}{k})^2 - m^2 \right).$$

If m is chosen so that the $(m+1)$ -th eigenvalue satisfies

$$\frac{\lambda_{m+1}}{\lambda_1} < 1,$$

then as $\mathbf{tm} \rightarrow \infty$, the error approaches zero. This ensures that the normalized embedding, given in Equation (32), provides a good approximation of the normalized affinity measure in Equation (31).

In Section 6, we perform several numerical experiments using these normalized quantities to demonstrate that the **Hodge Diffusion-Map** accurately approximates the affinity measure with only a small number of terms, m .

6 Numerical Experiments

In this section, we provide a numerical validation of the proposed methodology using sample points from the two-dimensional torus T^2 and the two-dimensional sphere S^2 . Our focus is on the normalized versions of the affinity measure, the Hodge Diffusion Maps, and the Hodge Diffusion Distance, as defined in [Equations \(31\)](#) and [\(32\)](#), respectively.

We compare the proposed methodology against several established algorithms: Vector Diffusion Maps [\[SW12\]](#), Diffusion Maps [\[CL06\]](#), t-distributed Stochastic Neighbor Embedding (t-SNE), and Principal Component Analysis (PCA). The implementation of Hodge Diffusion Maps follows the procedure described in [Algorithm 2](#). As a preliminary step, we apply local PCA using [Algorithm 1](#) to estimate the intrinsic dimensionality of the manifold structure underlying the dataset X .

In our experiments, we use the parameter settings specified in [Table 1](#) for the Hodge Diffusion-maps. The parameter K denotes the number of sample points in the neighborhood used to run [Algorithms 1](#) and [2](#). We set $K = 30$ to ensure a reasonable number of points without significantly impacting the computational cost. The threshold parameter γ , used in the Local PCA procedure described in [Section 4.1](#), is set to $\gamma = 0.9$ to estimate the intrinsic dimension d of the manifold.

The parameter m represents the number of truncated terms used to compute the embedding $\eta_{k,m}^{\mathbf{tm}}$ of the Hodge diffusion maps, as defined in [Equation \(30\)](#). Since $\eta_{k,m}^{\mathbf{tm}}$ is a symmetric matrix, we only consider the components in the form (i, j) where $1 \leq i \leq j \leq m$. We use $m = 3$ to visualize the results based on the first three terms.

The parameter \mathbf{tm} indicates the number of paths used to measure the connectivity between two points using the Hodge Laplacian Matrix to the power \mathbf{tm} . In our experiments, we set $\mathbf{tm} = 1$, though similar results were obtained with different values of \mathbf{tm} . These results suggest some stable behavior on the parameter \mathbf{tm} and should be further investigated in future work.

For a dataset $X = \{x_i\}_{i=1}^N$, the parameter t is the diffusion scaling factor. It is set as the average of the minimum distances between each point x_i and all other points x_j in the dataset. The choice of t is based on the need to select a small enough value to capture the data's structure, but not too small, as this could cause the term $e^{-\|x_i - x_j\|^2/2t^2}$ to vanish, losing important topological information.

Additionally, by applying the Cauchy–Schwarz inequality, we observe that the (l_1, l_2) component of both normalized Hodge diffusion maps and vector diffusion maps at a point x_i is dominated by the square root of the diagonal components (l_1, l_1) and (l_2, l_2) . This suggests that the diagonal components (l_k, l_k) encode information about the intensity of the diffusion of the embedding elements $\eta_{k,m}^{\mathbf{tm}}$.

In our numerical experiments, we plot the diagonal embedding of the normalized Hodge diffusion maps. Specifically, we plot the map:

$$x_i \rightarrow \frac{1}{\lambda_1} \left(\eta_{k,3}^1(x_i)(l, l) \right)_{1 \leq l \leq m}$$

Parameter	Value	Description
K	30	Number of points in the neighborhood
γ	0.9	Threshold parameter to estimate d
m	3	Truncation level
tm	1	Number of paths used to measure the connectivity between two points
t	$\mathbf{mean}_i \min_{j \neq i} \ x_i - x_j\ $	Diffusion scaling parameter

Table 1: Parameters specification for the Hodge diffusion-Map

We refer to this representation as the diagonal of the normalized Hodge diffusion maps.

In the following experiments, we examine two-dimensional manifolds, namely the torus T^2 and the sphere S^2 , each sampled with 2500 points distributed across them as described below.

For the torus T^2 , we use the parametrization:

$$\Omega(u, v) = [(2 + \cos(2\pi v)) \cos(2\pi u), (2 + \cos(2\pi v)) \sin(2\pi u), \sin(2\pi v)]$$

where $-\frac{1}{2} \leq u, v \leq \frac{1}{2}$. To construct the dataset, we define 50 evenly spaced sample points u_1, u_2, \dots, u_{50} within the interval $[-\frac{1}{2}, \frac{1}{2})$ using:

$$u_i = \frac{i-1}{50} - \frac{1}{2} \quad \text{for } 1 \leq i \leq 50.$$

Using this grid, the dataset X is then:

$$X = \{\Omega(u_i, u_j)\}_{1 \leq i, j \leq 50},$$

resulting in 2500 points distributed over T^2 .

For the sphere S^2 , we use the following parametrization:

$$\Omega(u, v) = [\cos(2\pi u) \sin(\pi v), \sin(2\pi u) \sin(\pi v), \cos(\pi v)]$$

where $0 \leq u, v \leq 1$. To create the dataset, we define 50 evenly spaced sample points u_1, u_2, \dots, u_{50} within the interval $[0, 1)$, given by:

$$u_i = \frac{i-1}{50} \quad \text{for } 1 \leq i \leq 50.$$

The resulting dataset X is defined as:

$$X = \{\Omega(u_i, u_j)\}_{1 \leq i, j \leq 50},$$

yielding in 2500 points distributed over S^2 .

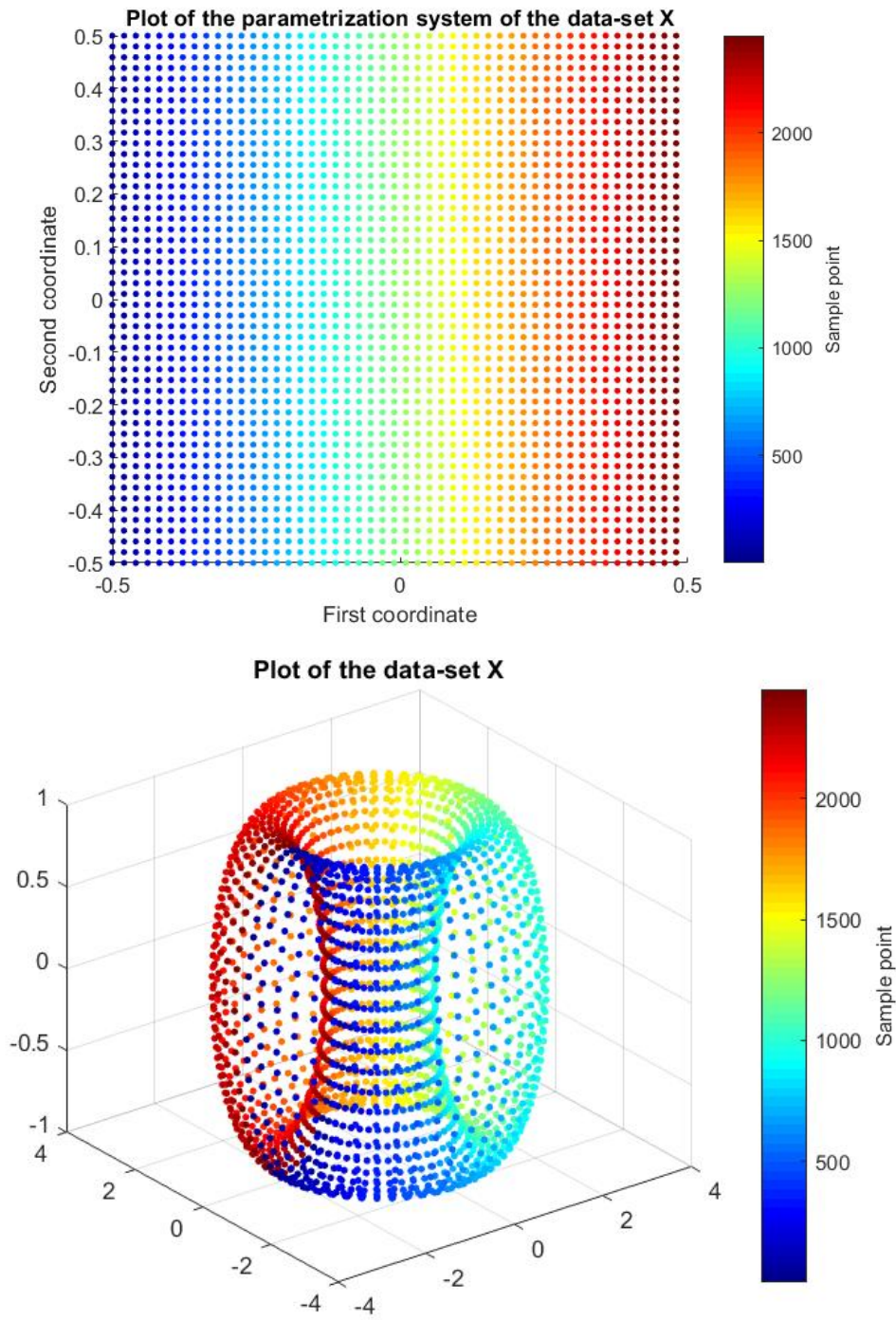


Figure 1: Top: The first and second coordinates of the parametrization system for the torus. Bottom: The dataset X plotted on the torus T^2 , with the colorbar indicating the order of the sample points.

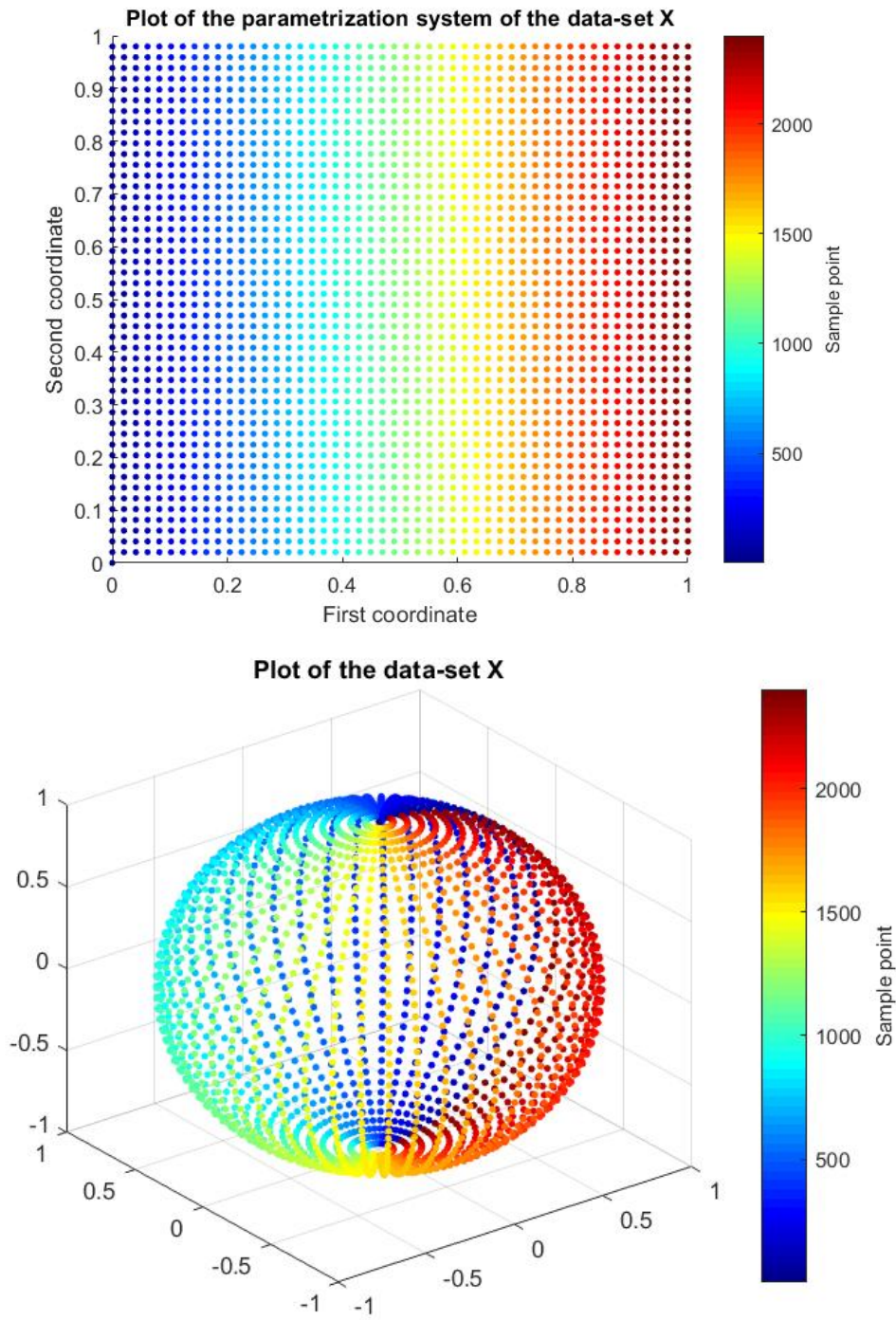


Figure 2: Top: The first and second coordinates of the parametrization system for the sphere. Bottom: The dataset X plotted on the sphere S^2 , with the colorbar indicating the order of the sample points.

In both parametrization systems, u_i and u_j correspond to the first and second coordinates, respectively. In [Figures 1 and 2](#), we visualize the datasets sampled over the torus and sphere, respectively. The colorbars indicate the ordering of the sample points, providing a reference for their distribution across the surfaces.

Although the dataset X consists of points sampled from each manifold, the number of points may not be large enough to fully capture the entire manifold. As a result, the dataset X could potentially represent a submanifold within the manifold or a totally different manifold from the one theoretically assumed.

The goal of this experiment is to explore how the Hodge Diffusion Maps method can be used to extract topological information from the sample dataset in X . Using the local PCA algorithm ([Algorithm 1](#)), we estimate the intrinsic dimension of both the torus and sphere datasets to be $d = 2$. Consequently, we can apply the Hodge diffusion-maps embedding up to the second order, that is, for $k \in \{1, 2\}$.

In the next section, we present the results and analysis for each manifold. All experiments were performed using MATLAB software on a laptop equipped with an Intel Core i5-1235U 1.30 GHz processor and 8 GB of RAM. The algorithms used in our implementation are available in the GitHub repository [[GF25](#)].

6.1 Results over two-dimensional torus

The first-order normalized Hodge diffusion maps embedding ($k = 1$) is shown in [Figure 3](#), while the second-order Hodge diffusion maps embedding ($k = 2$) is presented in [Figure 4](#). Additionally, in [Figure 5](#), we show the vector diffusion maps embedding. The computational time for running the Hodge diffusion maps was 98.67 seconds for the first order and 14.69 seconds for the second order.

The Hodge diffusion map embeddings, for both first and second orders, reveal two distinct regions with different features. One region is concentrated around points where u_2 is close to 0, while the other lies outside this area. Within each region, the values of the (i, j) component exhibit similar characteristics, as indicated by distinct color patterns unique to each region. This shows that the Hodge diffusion map successfully identifies two regions with different structural characteristics.

Similarly, the vector diffusion map identifies two regions: one near points where u_2 is approximately -0.5 or 0.5, and another outside this area. Both algorithms, thus detect a partition of the dataset into two regions. While the specific regions identified by each method are not identical, they are closely related through the Weitzenböck identity, which connects the Hodge Laplacian with the Connection Laplacian operator.

Additionally, in [Figure 6](#), we plot the diagonal of the normalized Hodge and vector diffusion maps. The first and second rows show the diagonals of the normalized Hodge diffusion maps for the first ($k = 1$) and second ($k = 2$) orders, respectively, while the third row shows the vector diffusion maps. The left column contains the first two components, and the right column contains the first three components of the respective diagonals.

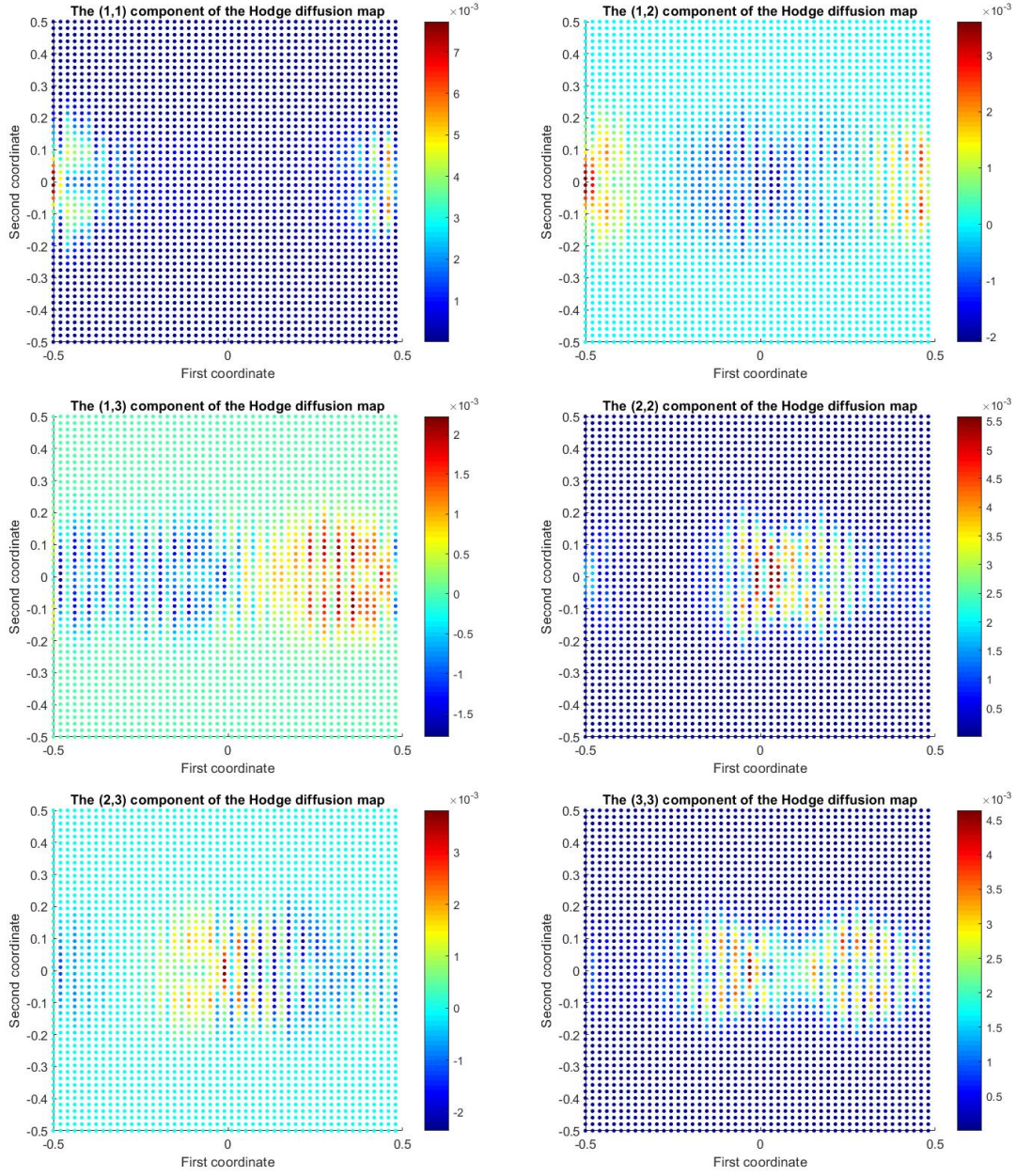


Figure 3: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of the first order normalized Hodge Diffusion Maps $\eta_{1,3}^1$ as given in Eq. (32).

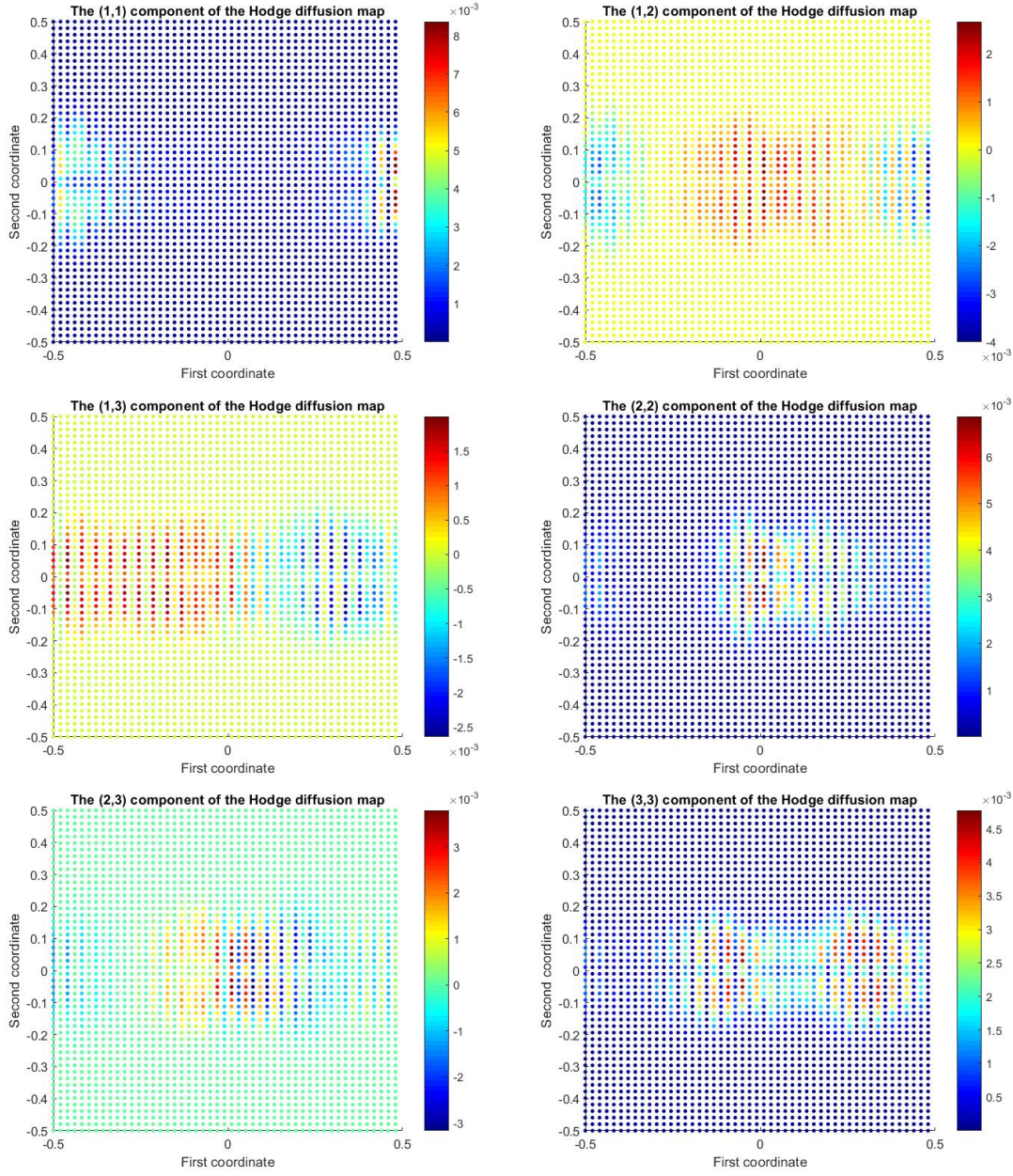


Figure 4: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of the second order normalized Hodge Diffusion Maps $\eta_{2,3}^1$ as given in Eq. (32)

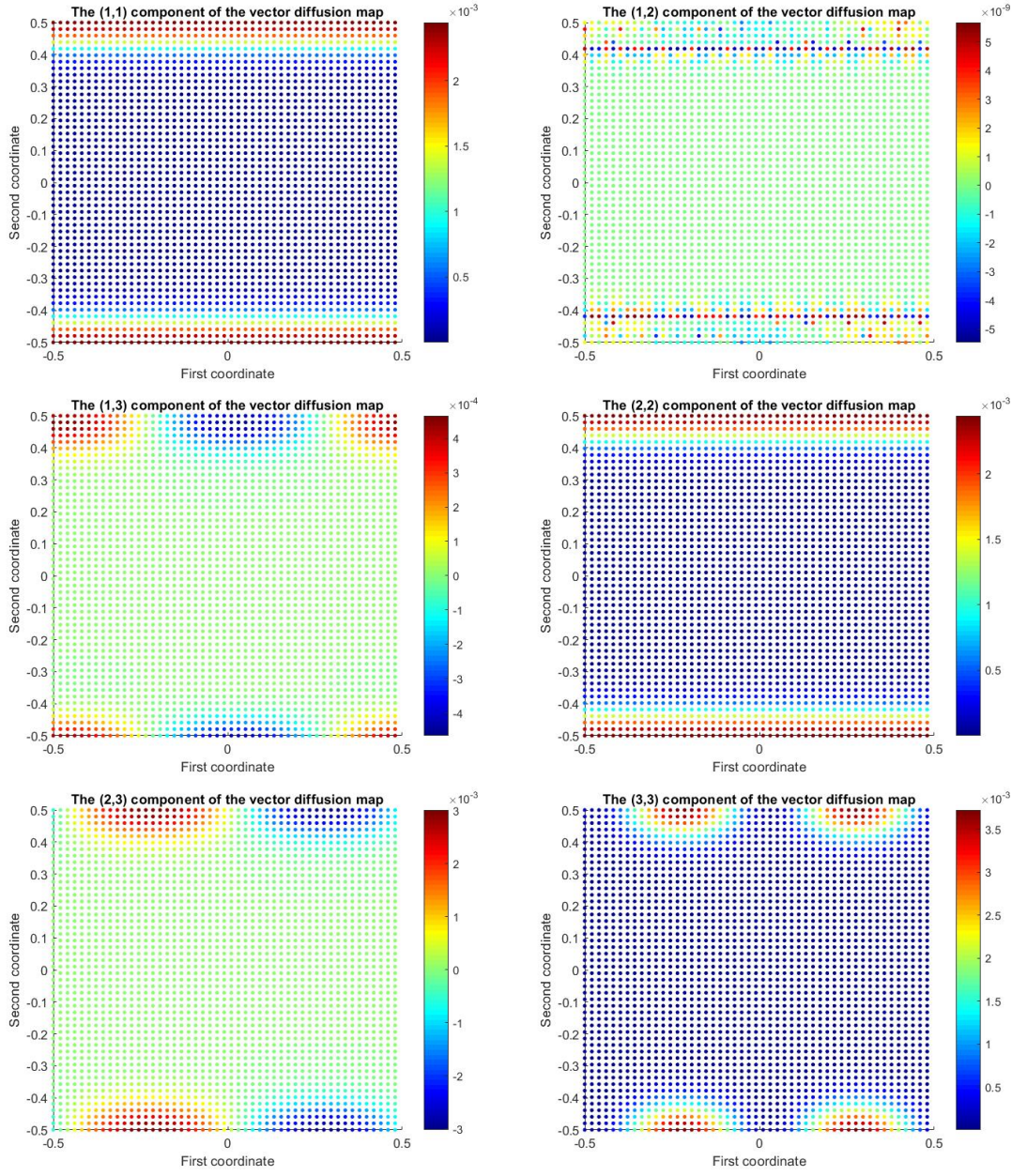


Figure 5: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of vector diffusion maps.

We compare the proposed methodology with the t-SNE, PCA, and diffusion map algorithms applied to the dataset X , as shown in Figure 7. The colorbar and dataset organization are the same as in Figure 1.

The Hodge diffusion map, for both first and second order ($k = 1$ and $k = 2$), map the vertical sections of the torus, corresponding where the first coordinate u_i is constant and assigned the same color, to approximations of straight lines in the two and three dimensional space. In contrast, vector diffusion maps represent the entire data set as several parallel straight lines, without distinguishing the vertical sections. The t-SNE algorithm transforms these vertical sections into nonlinear curves, while PCA algorithm projects the torus onto a two-dimensional plane, where the vertical sections collapse into overlapping ellipses. Diffusion Maps, on the other hand, arrange the vertical sections into lines forming a circular pattern.

The results show that both diffusion maps and Hodge diffusion maps are capable of identifying and classifying vertical sections by mapping them to approximate straight lines in two- or three-dimensional spaces. This suggests that these algorithms extract topological features by mapping points from the same vertical section onto a straight line in the embedded space. As a result, linear classifiers can be used as a postprocessing step to efficiently classify data points based on their topological features.

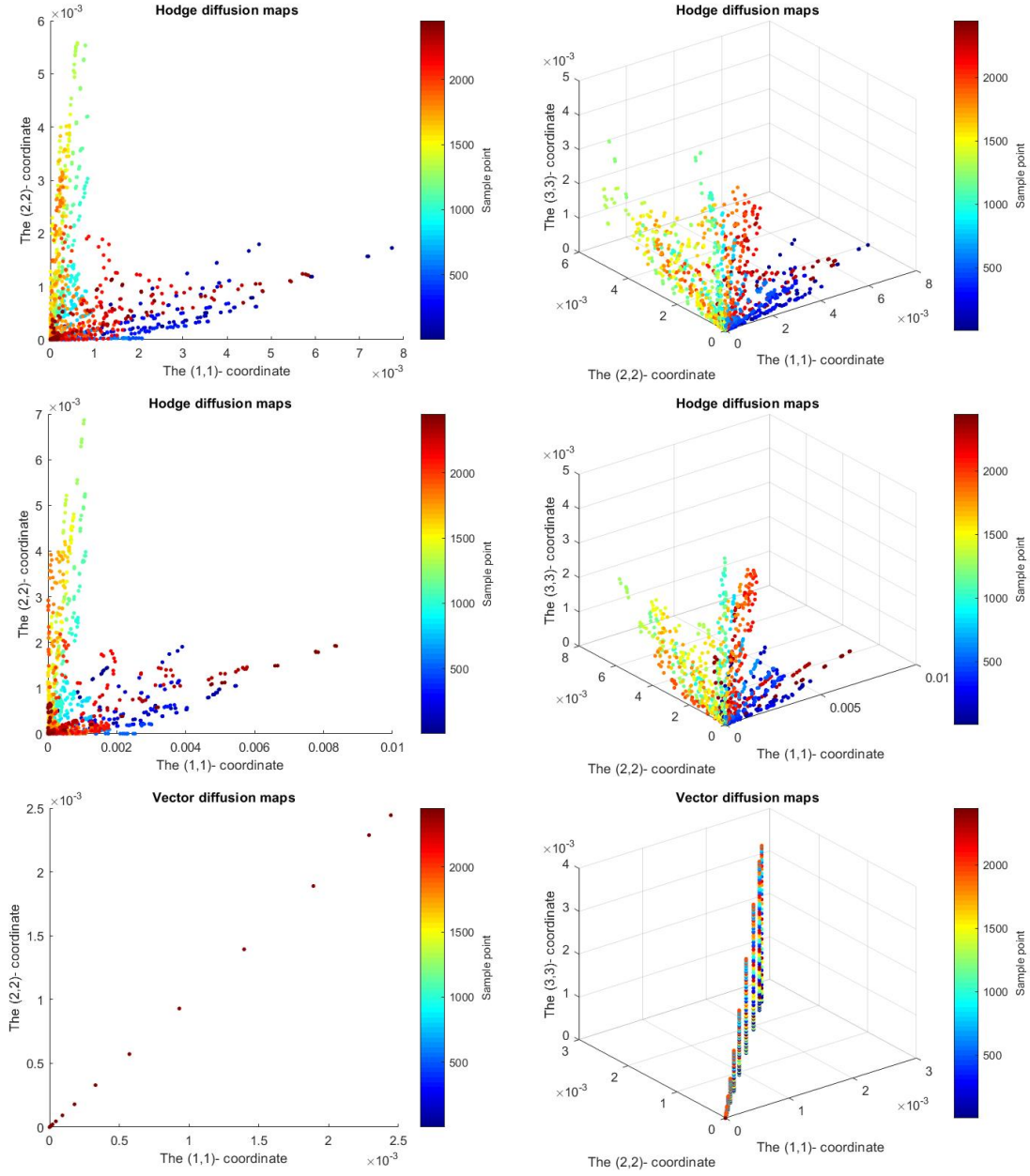


Figure 6: Plot of the diagonal coordinates of the Hodge diffusion maps and vector diffusion maps. The first row shows the Hodge diffusion maps of the first order ($k = 1$). The second row shows the second order ($k = 2$) Hodge diffusion maps, and the third row shows the vector diffusion map. In the left column, we plot the first two diagonal coordinates, $(1,1)$ and $(2,2)$, and in the right column, we plot the first three diagonal coordinates, $(1,1)$, $(2,2)$, and $(3,3)$. The colorbar indicates the order of the points, matching the colorbar in Figure 1.

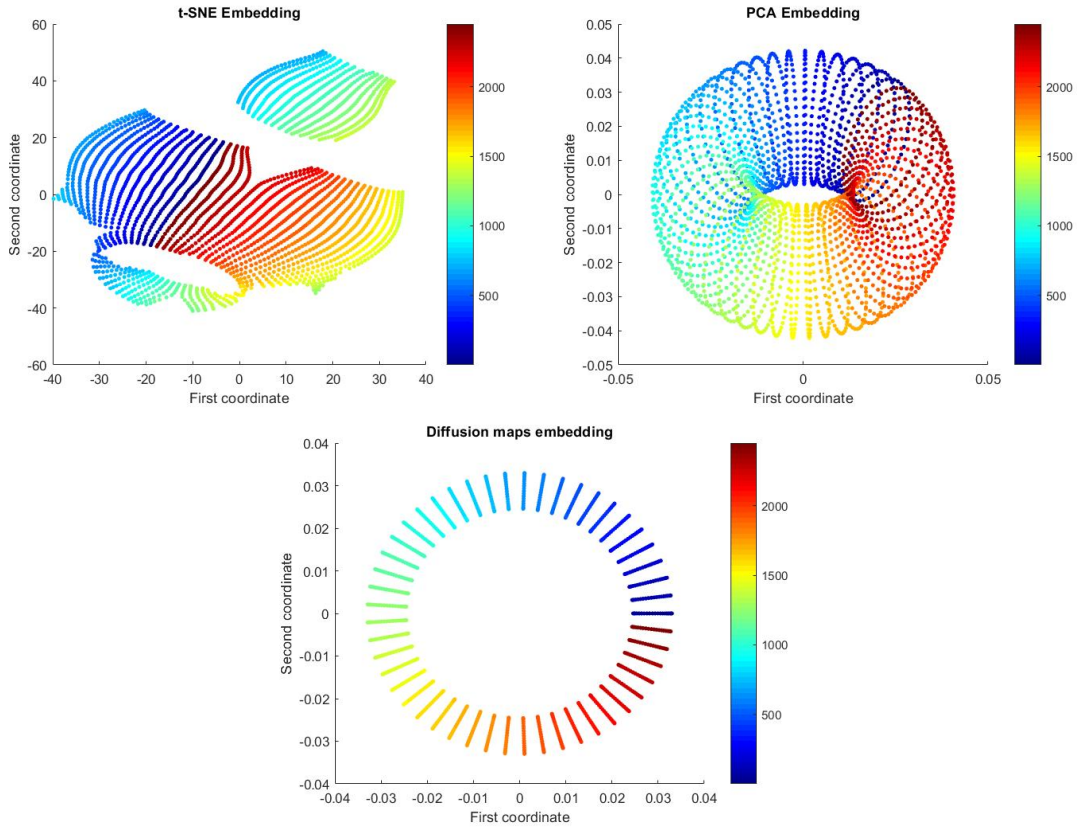


Figure 7: Plot of the t-SNE, PCA, and Diffusion Maps algorithms applied to the dataset sampled on the torus.

6.2 Results over the two-dimensional sphere

The first-order normalized Hodge diffusion maps embedding ($k = 1$) is shown in Figure 8, while the second-order embedding ($k = 2$) is presented in Figure 9. Additionally, the vector diffusion maps embedding is displayed in Figure 10. The computation times were 87.98 seconds for the first-order Hodge diffusion maps and 13.86 seconds for the second-order.

The results show that both the first- and second-order Hodge embeddings reveal two regions with distinct characteristics: one where the second coordinate u_2 is close to 0.5, and another outside this range. Within each region, the values of the (i, j) components exhibit similar patterns, as reflected in the distinct color patterns unique to each region. In contrast, the vector diffusion maps also identify two regions—one where u_2 is near 0 and another outside this region. Similar to the Torus case, both algorithms detect two separate regions, which are linked by the Weitzenböck identity, connecting the Hodge Laplacian and the Connection Laplacian operator.

Additionally, in Figure 11, we display the diagonals of the normalized Hodge and vector diffusion maps embeddings. The first and second rows correspond to the first-order ($k = 1$) and second-order ($k = 2$) normalized Hodge diffusion maps, respectively, while the third row shows the vector diffusion maps. The left column contains the first two components, and the right column contains the first three components of the respective diagonals.

To evaluate the proposed methodology, we compare it with t-SNE, PCA, and diffusion maps algorithms applied to the dataset X as shown in Figure 12. The colorbar and dataset organization follow the same convention as in Figure 2. As illustrated in Figure 11, the two- and three-dimensional embeddings produced by the first- and second-order Hodge diffusion maps ($k = 1$ and $k = 2$), respectively, map the vertical sections of the dataset X , corresponding to points where the first coordinate u_i is constant and assigned the same color, into distinct regions. These regions can then be separated by linear classifiers, providing a method to divide the dataset based on its vertical sections. Thus, the proposed methodology provides a useful toolbox for classifying points in the dataset based on topological patterns.

In contrast, vector diffusion maps embed the dataset into two straight lines, failing to differentiate between distinct vertical sections. Similarly, both the t-SNE and PCA algorithms transform the vertical sections into nonlinear curves, making it difficult to apply linear classifiers to separate the data based on these sections. Additionally, the diffusion map algorithm struggles to differentiate between these vertical sections.

Among all the algorithms tested, Hodge diffusion maps is the only method that enables the use of linear classifiers in the embeddings to categorize the dataset based on points with similar topological structures defined by the vertical sections.

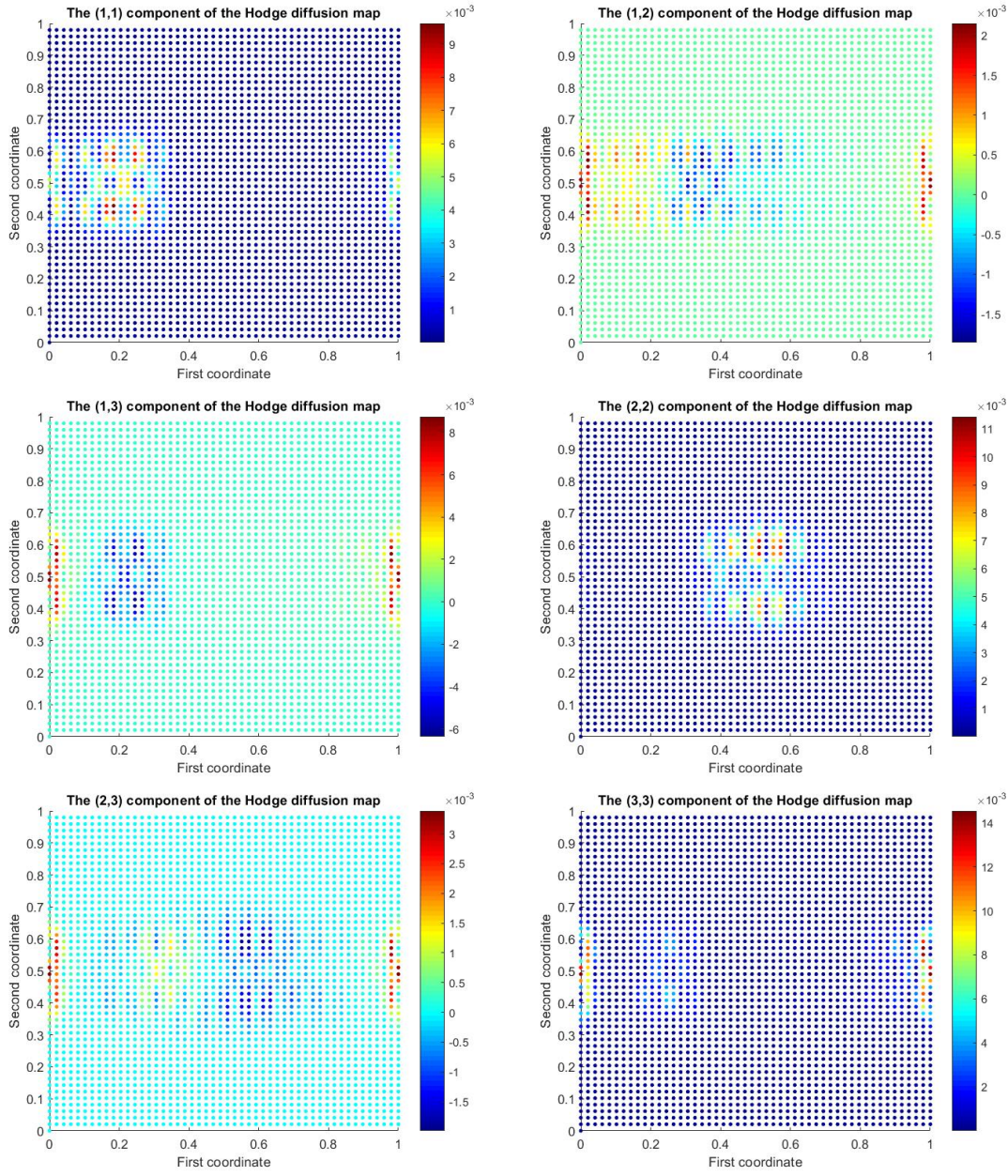


Figure 8: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of the first order normalized Hodge Diffusion Maps $\eta_{1,3}^1$ as given in Eq. (32), for sampled points on the sphere.

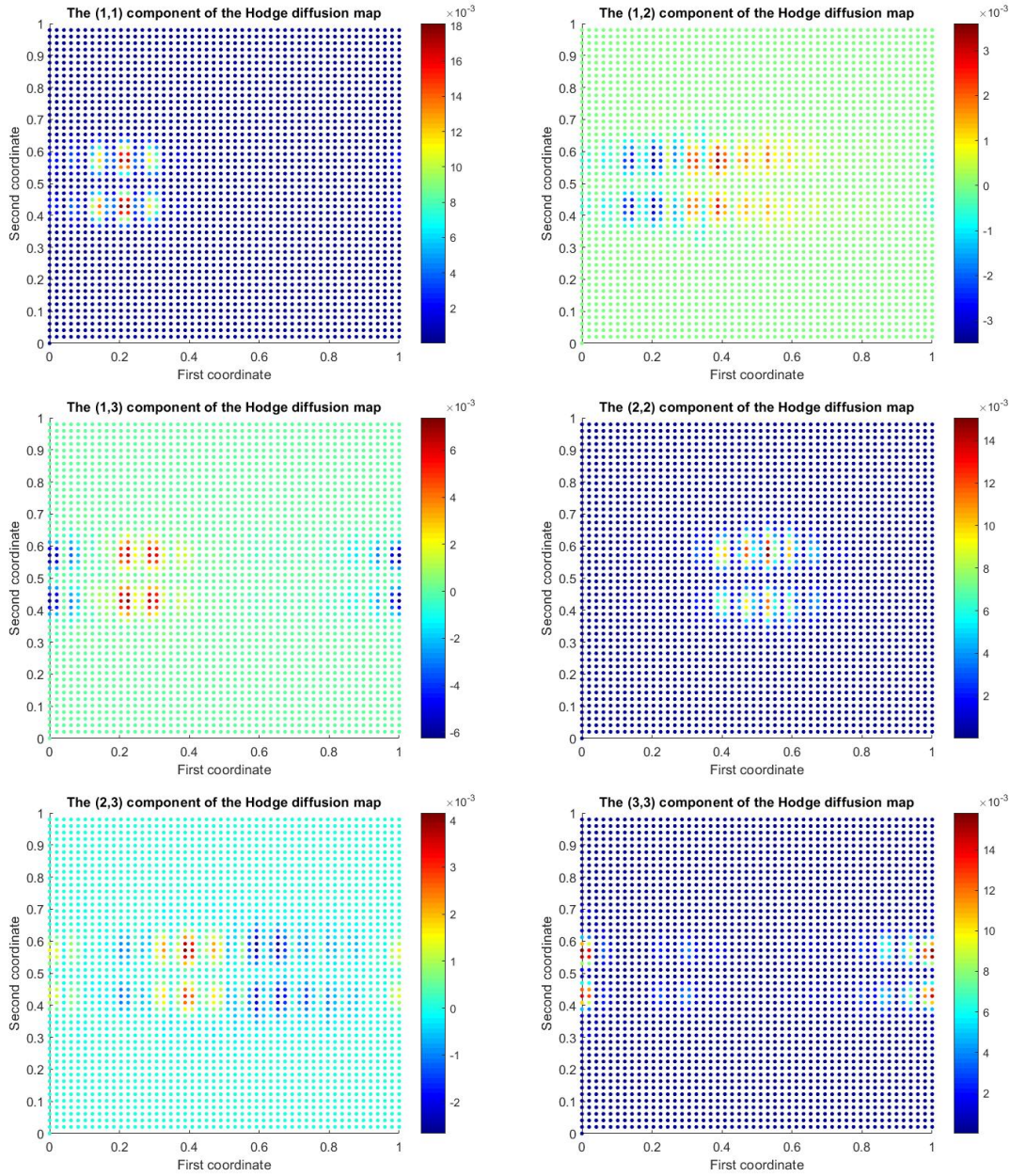


Figure 9: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of the second order normalized Hodge Diffusion Maps $\eta_{2,3}^1$ as given in Eq. (32), for sampled points on the sphere.

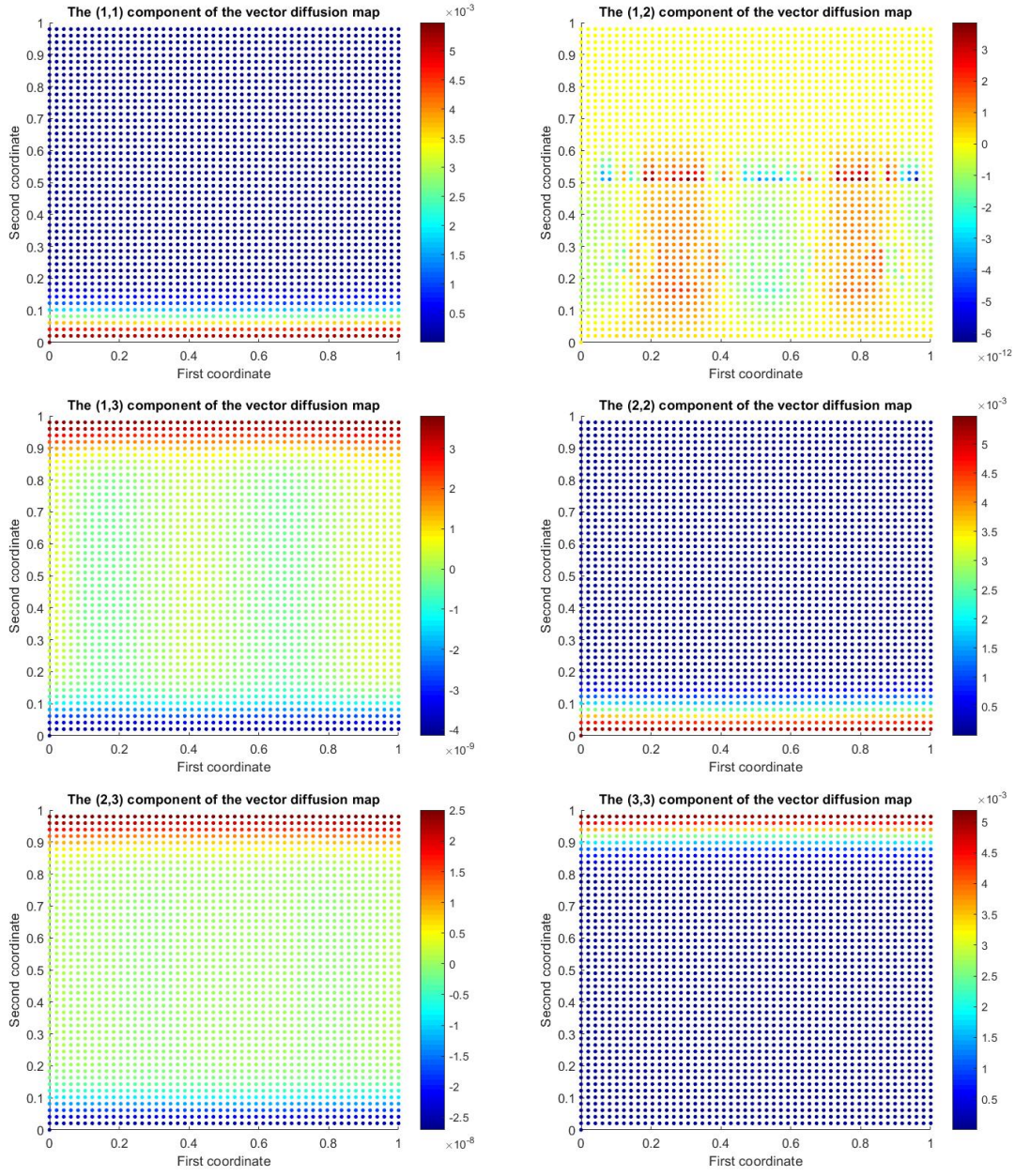


Figure 10: Plot of the (c_1, c_2) components, where $1 \leq c_1 \leq c_2 \leq 3$, of vector diffusion maps for sampled points on the sphere.

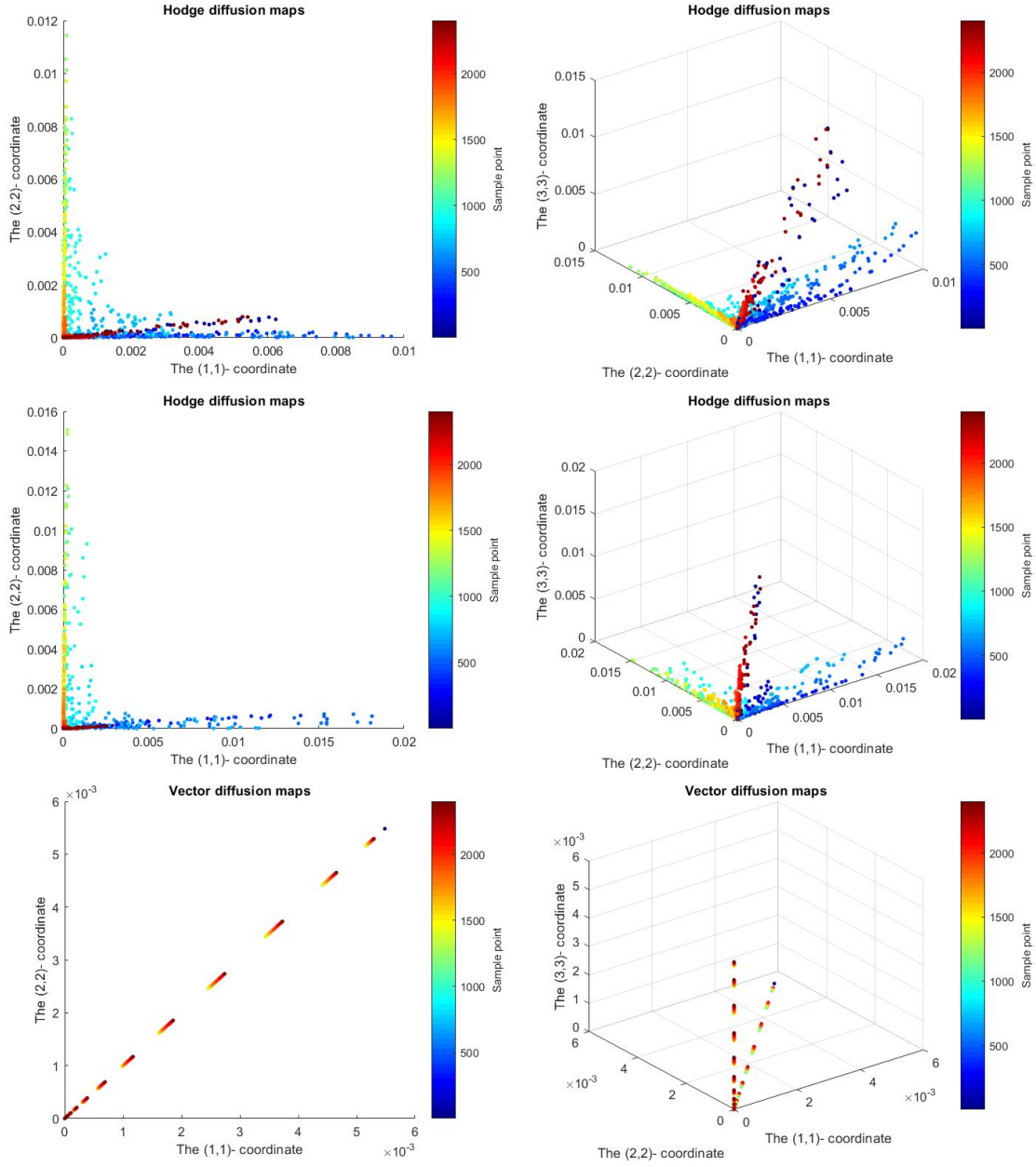


Figure 11: Plot of the diagonal coordinates of the Hodge diffusion maps and vector diffusion maps. The first row shows the Hodge diffusion maps of the first order ($k = 1$). The second row shows the second order ($k = 2$) Hodge diffusion maps, and the third row shows the vector diffusion map. In the left column, we plot the first two diagonal coordinates, (1,1) and (2,2), and in the right column, we plot the first three diagonal coordinates, (1,1), (2,2), and (3,3). The colorbar indicates the order of the points, matching the colorbar in Figure 1. The dataset consists of points sampled over the sphere.

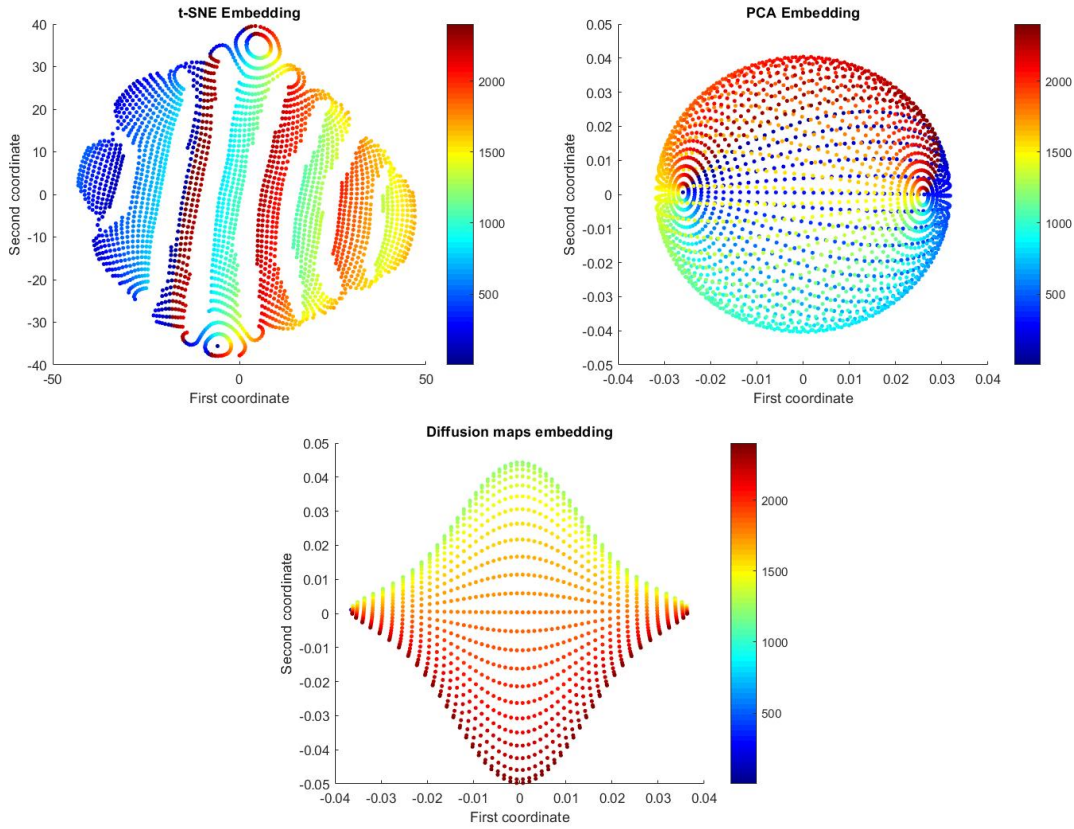


Figure 12: Plot of the t-SNE, PCA, and Diffusion Maps algorithms applied to the dataset sampled on the sphere S^2 .

7 Conclusions and future directions

In this paper, we introduce Hodge diffusion maps, a generalization of both vector diffusion maps and classical diffusion maps. Assuming the dataset lies on a manifold, our algorithm leverages the k -th Hodge Laplacian—closely connected to the k -th cohomology group via Hodge theory—to extract topological features. While classical diffusion maps correspond to the computation of the zero-order Hodge Laplacian, and vector diffusion maps to the connection Laplacian (equivalent to the first-order Hodge Laplacian), our approach extends naturally to compute the Hodge Laplacian of any order $k \geq 1$. This enables the extraction of richer topological information from the dataset.

We validate our approach through two numerical experiments on datasets sampled from a torus and a sphere. In the first experiment (torus), both the proposed Hodge diffusion maps and classical diffusion maps successfully embed the vertical sections of the dataset as straight lines in both 2D and 3D Euclidean space, whereas the other methods fail to capture this structure. In the second experiment (sphere), only Hodge diffusion maps successfully separate each vertical section into distinct regions in both 2D and 3D embeddings, while classical diffusion maps do not achieve this distinction. Additionally, the embeddings produced by t-SNE and PCA map points with the similar topological structure -defined by the vertical sections- onto nonlinear curves. This complicates the use of linear classifiers to separate the data based on these vertical sections, further highlighting the superior performance of the proposed algorithm over these methods.

These experiments demonstrate that Hodge diffusion maps, as a dimensionality reduction technique, more effectively capture the topological structure of a dataset by mapping points with similar topological features to nearby regions in Euclidean space. This facilitates the use of linear classifiers to categorize the embedded points, underscoring Hodge diffusion maps as a valuable complementary tool for extracting additional topological information from the dataset.

Based on the connection between diffusion map theory and vector diffusion maps with cryo-electron microscopy (Cryo-EM), as in Refs. [SS11, SW12, CSSS08], we plan to explore the use of Hodge diffusion maps as a pre-processing tool for heterogeneous particles in future work. Another potential direction is to incorporate the Hodge-Laplacian matrix, as outlined in Equation (28), as a penalty term for regularizing neural networks. This approach could enhance the network’s ability to extract topological features from the dataset, thereby improving its robustness during training.

Acknowledgements. The first author was supported by Centro de Modelamiento Matemático (CMM) BASAL fund FB210005 for center of excellence from ANID-Chile. The second author was supported by Fondecyt ANID postdoctoral grant 3220631.

References

- [CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.
- [CSSS08] Ronald R Coifman, Yoel Shkolnisky, Fred J Sigworth, and Amit Singer. Graph laplacian tomography from unknown random projections. *IEEE Transactions on Image Processing*, 17(10):1891–1899, 2008.
- [DC92] Manfredo P Do Carmo. *Riemannian Geometry*. Mathematics (Boston, Mass.). Birkhäuser, 1992.
- [DC16] Manfredo P Do Carmo. *Differential geometry of curves and surfaces*. Courier Dover Publications, 2016.
- [GF25] Alvaro Almeida Gomez and Jorge Duque Franco. *Codes and numerical implementations of Hodge Diffusion Maps*, Apr. 3, 2025. <https://github.com/alvaroalmeidagomez/HDM>.
- [GNZ21] Alvaro Almeida Gomez, Antônio J Silva Neto, and Jorge P Zubelli. Diffusion representation for asymmetric kernels. *Applied Numerical Mathematics*, 166:208–226, 2021.
- [GNZ23] Alvaro Almeida Gomez, Antônio J. Silva Neto, and Jorge P. Zubelli. A diffusion-map-based algorithm for gradient computation on manifolds and applications. *IEEE Access*, 11:90622–90640, 2023.
- [HHS⁺23] Mingzhen He, Fan He, Lei Shi, Xiaolin Huang, and Johan AK Suykens. Learning with asymmetric kernels: Least squares and feature interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10044–10054, 2023.
- [HHYH23] Mingzhen He, Fan He, Ruikai Yang, and Xiaolin Huang. Diffusion representation for asymmetric kernels via magnetic transform. *Advances in Neural Information Processing Systems*, 36:53742–53761, 2023.
- [JLYY11] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.
- [Laf04] Stéphane S Lafon. *Diffusion maps and geometric harmonics*. Yale University, 2004.
- [Lee12] J. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2012.

- [Lim20] Lek-Heng Lim. Hodge laplacians on graphs. *Siam Review*, 62(3):685–715, 2020.
- [RGWC⁺24] Emily Ribando-Gros, Rui Wang, Jiahui Chen, Yiying Tong, and Guo-Wei Wei. Combinatorial and hodge laplacians: Similarities and differences. *SIAM Review*, 66(3):575–601, 2024.
- [SS11] Amit Singer and Yoel Shkolnisky. Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming. *SIAM journal on imaging sciences*, 4(2):543–572, 2011.
- [SW11] Amit Singer and Hau-tieng Wu. Orientability and diffusion maps. *Applied and computational harmonic analysis*, 31(1):44–58, 2011.
- [SW12] Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012.
- [TSL23] Bogdan Toader, Fred J. Sigworth, and Roy R. Lederman. Methods for cryo-em single particle reconstruction of macromolecules having continuous heterogeneity. *Journal of Molecular Biology*, 435(9):168020, 2023. New Frontier of Cryo-Electron Microscopy Technology.
- [War83] Frank W Warner. *Foundations of differentiable manifolds and Lie groups*, volume 94. Springer Science & Business Media, 1983.
- [WWLX22] Ronald Koh Joon Wei, Junjie Wee, Valerie Evangelin Laurent, and Kelin Xia. Hodge theory-based biomolecular data analysis. *Scientific Reports*, 12(1):9699, 2022.

A Alternating forms and alternating arrays

A vector subspace $V \subseteq \mathbb{R}^n$ is (non-canonically) isomorphic to its dual V^* . Fixing an inner product on V induces a natural isomorphism $V \simeq V^*$, which in turn establishes corresponding isomorphisms between various spaces constructed from V and V^* . For instance, this yield isomorphism between

$$\underbrace{V \otimes \cdots \otimes V}_{k\text{-times}} \text{ and } \underbrace{V^* \otimes \cdots \otimes V^*}_{k\text{-times}},$$

as well as between the exterior powers

$$\bigwedge^k(V) \text{ and } \bigwedge^k(V^*).$$

Since we are dealing with discrete data, a more concrete representation of the vector subspace V and its associated constructions is needed for numerical implementation, rather than relying solely on its abstract definition. To address this, we introduce the notion of a k -dimensional real array of size (n_1, n_2, \dots, n_k) . This approach provides a framework for defining alternating forms and, subsequently, differential forms in a manner that is better suited for numerical computations.

For every natural number n , we denote $I_n = \{1, 2, \dots, n\}$ and S_n be the set of all the permutations of I_n . A k -dimensional real array of size (n_1, n_2, \dots, n_k) is defined as a function

$$f : I_{n_1} \times \dots \times I_{n_k} \rightarrow \mathbb{R}.$$

In particular, a 1-dimensional array of size (n) corresponds to a vector of \mathbb{R}^n , while 2-dimensional array of size (n_1, n_2) corresponds to an $n_1 \times n_2$ matrix. In this sense, k -dimensional array naturally generalize the notions of vectors and matrices.

We denote the set of k -dimensional arrays of size (n_1, n_2, \dots, n_k) by $M(n_1, n_2, \dots, n_k)$. Given two arrays, one of dimension k (denoted $f \in M(n_1, n_2, \dots, n_k)$) and the other of dimension l (denoted $g \in M(m_1, m_2, \dots, m_l)$), we define their tensor product $f \otimes g$ as a $k + l$ -dimensional array in $M(n_1, n_2, \dots, n_k, m_1, m_2, \dots, m_l)$, given by

$$f \otimes g(i_1, \dots, i_k, j_1, \dots, j_l) = f(i_1, \dots, i_k)g(j_1, \dots, j_l)$$

with $i_s \in I_{n_s}$ and $j_{\hat{s}} \in I_{m_{\hat{s}}}$ with $1 \leq s \leq k$ and $1 \leq \hat{s} \leq l$. We endow the space $M(n_1, n_2, \dots, n_k)$ with the Frobenius inner product, defined as

$$\langle A, B \rangle_F = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_k=1}^{n_k} A(i_1, i_2, \dots, i_k) B(i_1, i_2, \dots, i_k) \quad (33)$$

for all $A, B \in M(n_1, n_2, \dots, n_k)$. Now observe that $M(n)$ corresponds to \mathbb{R}^n , so any linear subspace V of \mathbb{R}^n naturally defines a subspace of $M(n)$. Given an d -dimensional linear subspace $V \subseteq \mathbb{R}^n$, the tensor product space

$$\underbrace{V \otimes \dots \otimes V}_{k\text{-times}}$$

is identified with the linear subspace of $M(\underbrace{n, \dots, n}_{k\text{-times}})$ spanned by the elements

$$\{v_1 \otimes v_2 \otimes \dots \otimes v_k | v_i \in V.\}$$

Here, V is identified with its corresponding subspace in $M(n)$.

Definition 1. A k -alternating form $\omega : \underbrace{V \times V \times \dots \times V}_{k\text{-times}} \rightarrow \mathbb{R}$ defined over a vector space V is a multilinear map such that is alternating, that is, if for all vectors v_1, v_2, \dots, v_k and any permutation $\sigma \in S_k$

$$\omega^\sigma(v_1, \dots, v_k) := \omega(v_{\sigma(1)}, \dots, v_{\sigma(k)}) = (\text{sign } \sigma) \omega(v_1, \dots, v_k).$$

We denote the set of k -alternating forms as $\wedge^k(V^*)$

Since we have the Frobenius product, it induces an isomorphism between

$$\underbrace{V \otimes V \otimes \cdots \otimes V}_{k\text{-times}} \text{ and } \underbrace{(V \otimes V \otimes \cdots \otimes V)^*}_{k\text{-times}}.$$

As a result, each alternating form ω corresponds uniquely to a k -dimensional array W , allowing us to switch between ω and W using this isomorphism. We introduce the following definition:

Definition 2. A k -dimensional array $f \in \underbrace{V \otimes V \otimes \cdots \otimes V}_{k\text{-times}} \subseteq M(\underbrace{n, \dots, n}_{k\text{-times}})$ is called k -alternating array in V if it satisfies the following condition:

- **C1.** For all indices $i_1, i_2, \dots, i_k \in I_n$ and all permutation $\sigma \in S_k$, we have:
 $f(i_{\sigma(1)}, i_{\sigma(2)}, \dots, i_{\sigma(k)}) = (\text{sign } \sigma) f(i_1, i_2, \dots, i_k).$

We denote the set of k -alternating arrays as $\Theta^k(V)$.

An important property is that any k -dimensional array $f \in \underbrace{V \otimes \cdots \otimes V}_{k\text{-times}}$, even if it is not alternating, satisfies the following property, whose proof is straightforward

Property 1. For every permutation $\sigma \in S_k$, we let $f^\sigma(i_1, \dots, i_k) := f(i_{\sigma(1)}, \dots, i_{\sigma(k)})$, then for any vectors $v_1, \dots, v_k \in V$ we have:

$$< f^\sigma, v_1 \otimes v_2 \cdots \otimes v_k >_F = < f, v_{\sigma(1)} \otimes v_{\sigma(2)} \cdots \otimes v_{\sigma(k)} >_F$$

Proposition A.1. Let V a d -dimensional linear subspace of \mathbb{R}^n and $\omega \in \wedge^k(V^*)$ an alternating k -form. Then, there exists an unique k -alternating array $W \in \Theta^k(V)$ such that:

$$\omega(v_1, v_2, \dots, v_k) = < W, v_1 \otimes v_2 \cdots \otimes v_k >_F \quad (34)$$

where $< \cdot, \cdot >$ is the Frobenius inner product for arrays defined as in [Equation \(33\)](#).

Proof This follows from the fact that the isomorphism between V and V^* induced by an inner product (\cdot, \cdot) is given explicitly by the mapping

$$v \in V \rightarrow v^* \in V^*, \text{ where } v^*(t) := (v, t).$$

More precisely, using the formalism introduced so far: By the universal property of the tensor product ([Lee12](#), Proposition 12.7), there exists a unique linear map

$$L : V \otimes \cdots \otimes V \rightarrow \mathbb{R}$$

such that the following commutative diagram holds:

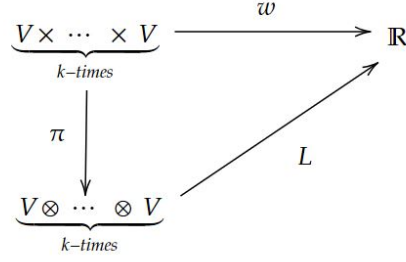


Figure 13: Commutative diagram for the functions w , L and π .

Here, the map

$$\pi : \underbrace{V \times \cdots \times V}_{k\text{-times}} \rightarrow \underbrace{V \otimes \cdots \otimes V}_{k\text{-times}}$$

is defined by

$$\pi(v_1, \dots, v_k) = v_1 \otimes \cdots \otimes v_k.$$

By the Riesz representation theorem, there exists a unique k -dimensional array $W \in \underbrace{V \otimes \cdots \otimes V}_{k\text{-times}}$ such that

$$L(v_1 \otimes v_2 \cdots \otimes v_k) = \langle W, v_1 \otimes v_2 \cdots \otimes v_k \rangle_F$$

for all $v_1, v_2, \dots, v_k \in V$. This together with the commutative diagram proves [Equation \(34\)](#). To show that W is alternating, let $\sigma \in S_k$ be a permutation and consider the k -dimensional array W^σ as in [Property 1](#). Then by the same property, we obtain

$$\begin{aligned}
\langle W^\sigma, v_1 \otimes v_2 \cdots \otimes v_k \rangle_F &= \langle W, v_{\sigma(1)} \otimes v_{\sigma(2)} \cdots \otimes v_{\sigma(k)} \rangle_F \\
&= w(v_{\sigma(1)}, v_{\sigma(2)}, \dots, v_{\sigma(k)}) \\
&= (\text{sgn } \sigma) w(v_1, \dots, v_k)
\end{aligned}$$

By the uniqueness of W , it follows that $W^\sigma = (\text{sgn } \sigma)W$, which completes the proof. ■

Example 1. As an example illustrating [Proposition A.1](#), consider the determinant as alternating form on \mathbb{R}^n . For any vector v_1, v_2, \dots, v_n , we have

$$\begin{aligned}
\det(v_1, v_2, \dots, v_n) &= \sum_{\sigma \in S_n} (\text{sgn } \sigma) v_1(\sigma(1)) v_2(\sigma(2)) \cdots v_n(\sigma(n)) \\
&= \langle W, v_1 \otimes v_2 \cdots \otimes v_n \rangle_F
\end{aligned}$$

where the array $W(i_1, \dots, i_n)$ is defined as $\text{sgn } \sigma$ if there exist a permutation σ with $\sigma(s) = i_s$ for $1 \leq s \leq n$ and 0 otherwise. In the two dimensional case \mathbb{R}^2 the 2-alternating array W is given by

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Remark A.1. *Proposition A.1 establishes an isomorphism between k -alternating forms $\Lambda^k(V^*)$ and k -alternating arrays $\Theta^k(V)$, which $I: \Lambda^k(V^*) \rightarrow \Theta^k(V)$.*

Remark A.2. *Given an alternating form $\omega \in \Lambda^k(\mathbb{R}^n)^*$ and a linear subspace V of \mathbb{R}^n , the restriction $\omega|_{V \times \dots \times V}$ induces an alternating form in $\Lambda^k(V)$. In this case, it is straightforward to show that $I(\omega|_{V \times \dots \times V}) = \mathcal{P}_{V \otimes \dots \otimes V}(I(\omega))$, where $\mathcal{P}_{V \otimes \dots \otimes V}$ is the orthogonal projection onto the subspace $V \otimes \dots \otimes V$. Therefore for any k -alternating array $W \in \Theta^k(\mathbb{R}^n)$ the projection onto the tensor space $V \otimes \dots \otimes V$ induces an k -alternating array on the linear subspace V , which is given by $\mathcal{P}_{V \otimes \dots \otimes V}(W) \in \Theta^k(V)$.*

Via the isomorphism I , we can compute the wedge product of k -alternating forms using k -alternating arrays. Consequently, all possible (discrete) computations of differential k -forms will inherently rely on k -alternating arrays, if this isomorphism is not explicitly mentioned. For instance, let ω_1 be a k_1 -alternating form and ω_2 a k_2 -alternating form. Recall that their wedge product $\omega_1 \wedge \omega_2$ is a $(k_1 + k_2)$ -alternating form given by

$$\omega_1 \wedge \omega_2 = \frac{1}{k_1!k_2!} \sum_{\sigma \in S_{k_1+k_2}} (\text{sgn } \sigma)(\omega_1 \otimes \omega_2)^\sigma,$$

where $\omega_1 \otimes \omega_2(u, v) = \omega_1(u)\omega_2(v)$, and ω^σ is defined as in [Definition 1](#). In this framework, we obtain the compatibility relation

$$I(\omega_1 \wedge \omega_2) = I(\omega_1) \wedge I(\omega_2),$$

where the wedge product on the right-hand side is defined in the same manner as in differential forms. Naturally, this identity can be verified directly using the isomorphism I and the previously established properties. We illustrate this in the following proposition:

Proposition A.2. *Let $\omega_1 \in \Lambda^{k_1}(V^*)$ and $\omega_2 \in \Lambda^{k_2}(V^*)$ then:*

$$I(\omega_1 \wedge \omega_2) = I(\omega_1) \wedge I(\omega_2)$$

Proof Observe that for all $v_1, \dots, v_{k_1}, v_{k_1+1}, \dots, v_{k_1+k_2} \in V$, we have, by [Property 1](#)

$$\begin{aligned} & \langle I(\omega_1) \wedge I(\omega_2), v_1 \otimes \dots \otimes v_{k_1} \otimes v_{k_1+1} \otimes \dots \otimes v_{k_1+k_2} \rangle_F \\ &= \frac{1}{k_1!k_2!} \sum_{\sigma \in S_{k_1+k_2}} (\text{sgn } \sigma) \langle I(\omega_1) \otimes I(\omega_2), v_{\sigma(1)} \otimes \dots \otimes v_{\sigma(k_1)} \otimes v_{\sigma(k_1+1)} \otimes \dots \otimes v_{\sigma(k_1+k_2)} \rangle_F \\ &= \frac{1}{k_1!k_2!} \sum_{\sigma \in S_{k_1+k_2}} (\text{sgn } \sigma) \langle I(\omega_1), v_{\sigma(1)} \otimes \dots \otimes v_{\sigma(k_1)} \rangle \langle I(\omega_2), v_{\sigma(k_1+1)} \otimes \dots \otimes v_{\sigma(k_1+k_2)} \rangle_F \\ &= \frac{1}{k_1!k_2!} \sum_{\sigma \in S_{k_1+k_2}} (\text{sgn } \sigma)(\omega_1 \otimes \omega_2)^\sigma(v_1, \dots, v_{k_1}, v_{k_1+1}, \dots, v_{k_1+k_2}) \\ &= (\omega_1 \wedge \omega_2)(v_1, \dots, v_{k_1}, v_{k_1+1}, \dots, v_{k_1+k_2}). \end{aligned}$$

Since $I(\omega_1) \wedge I(\omega_2)$ belongs to $\Theta^{k_1+k_2}(V)$, the uniqueness of Equation (34) guarantees that

$$I(\omega_1) \wedge I(\omega_2) = I(w_1 \wedge w_2).$$

■

We conclude this section by recalling the following result, which will be used in various calculations: If v_1, v_2, \dots, v_d form an orthonormal basis for V , then the set of wedge products

$$\left\{ \frac{1}{\sqrt{k!}} v_{i_1}^* \wedge v_{i_2}^* \wedge \dots \wedge v_{i_k}^* \mid i_1 < i_2 < \dots < i_k \right\} \quad (35)$$

constitutes an orthonormal basis for $\Lambda^k(V^*)$, where the inner product on $\Lambda^k(V^*)$ is given by

$$\langle \omega_1, \omega_2 \rangle_{\Lambda^k(V^*)} = \langle I(\omega_1), I(\omega_2) \rangle_F$$

B Proof of Theorem 3.1

In this section, we present the technical details supporting Theorem 3.1. The proof builds upon the framework developed in [GNZ23], with several components adapted to fit our setting. For additional background and a more comprehensive treatment of the underlying concepts, we refer the reader to [GNZ23], as well as to [DC92, DC16] for a thorough introduction to differential geometry.

Recall that \mathcal{M} is a closed (i.e., compact without boundary) Riemannian manifold and let $x \in \mathcal{M}$. For a small positive real number ε , consider the map $\psi = \exp_x \circ T : B(0, \varepsilon) \subset \mathbb{R}^d \rightarrow \mathcal{M}$ which defines a normal coordinate system around the point x . Here, \exp_x denotes the exponential map at x , and $T : \mathbb{R}^d \rightarrow T_x \mathcal{M}$ is a rotation from \mathbb{R}^d onto the tangent space $T_x \mathcal{M}$, which is considered as subset of \mathbb{R}^n . Note that $\psi(0) = x$. We now recall some estimates in normal coordinates system that are useful for approximating differential operators. The Taylor expansion of ψ around the point 0 is given by

$$\psi(v) = x + T(v) + \frac{1}{2} D^2 \psi_0(v, v) + O(\|v\|^3), \quad (36)$$

where $D^2 \psi_0$ denotes the second order differential (also known as the Hessian) of ψ at 0. Let $v \in B(0, \varepsilon) \subset \mathbb{R}^d$, and consider the geodesic $\gamma_{T(v)}$, with initial tangent vector $T(v) \in T_x \mathcal{M}$. Then, the expansion in Equation (36) can be rephrased in terms of the geodesic as

$$\gamma_{T(v)}(t) = x + T(v)t + \frac{1}{2} D^2 \psi_0(v, v)t^2 + O(\|v\|^3)t^3, \quad (37)$$

for $t \in \mathbb{R}$. Since the covariant derivative of a geodesic vanishes, we have that $\gamma_{T(v)}''$ is orthogonal to $T_x \mathcal{M}$. Therefore, from Equations (36) and (37), we obtain the following estimates

$$\|\psi(v) - x\|^2 = \|T(v)\|^2 + O(\|v\|^4), \quad (38)$$

and

$$\mathcal{P}_{T_x\mathcal{M}}(\psi(v) - x) = T(v) + O(\|v\|^3), \quad (39)$$

where $\mathcal{P}_{T_x\mathcal{M}}$ denotes the orthogonal projection onto the tangent space $T_x\mathcal{M}$. Here we have taken advantage of the fact that the manifold \mathcal{M} is embedded in \mathbb{R}^n . Moreover, letting e_1, \dots, e_d be the standard basis in \mathbb{R}^d , then by differentiating Equation (39) with respect to the variable v_i we obtain:

$$\mathcal{P}_{T_x\mathcal{M}}\left(\frac{\partial\psi}{\partial v_i}(v)\right) = T(e_i) + O(\|v\|^2), \quad (40)$$

Using Estimates (38) and (39), we conclude that there exist positive constants M_1 and M_2 such that, for $\|v\|$ small

$$\|v\| - M_2\|v\|^3 \leq \|\psi(v) - x\| \leq M_1\|v\|.$$

In particular, if $\|v\|^2 \leq \frac{1}{2M_2}$, then

$$\frac{1}{2}\|v\| \leq \|\psi(v) - x\| \leq M_1\|v\|.$$

This implies that, for small t , we have the following inclusion:

$$B(0, t/M_1) \subseteq \psi^{-1}(U(x, t^\delta)) \subseteq B(0, 2t). \quad (41)$$

where $U(x, t^\delta)$ denotes the ball in \mathcal{M} centered at x with radius t^δ , that is

$$U(x, t^\delta) := \{y \in \mathcal{M} \mid \|y - x\| \leq t^\delta\}.$$

B.1 Expansion of the Operator in Equation (4)

In this section, we continue with the technical development of the proof of Theorem 3.1. The central idea is to apply the Taylor expansion of the differential form w around the point p . To this end, we present a sequence of lemmas that progressively build toward the main result, which will be established at the end of the section.

Lemma B.1. *Assume that $\frac{1}{2} < \delta < 1$, and let $K : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}^m$ be a vector value kernel. Define*

$$P_{t,\delta}(x) = \int_{U(x, t^\delta)} K(x, y) e^{-\frac{\|y-x\|^2}{2t^2}} dVol(y),$$

where the integration is performed componentwise for the vector-valued function. Suppose that for small t , the function $\psi : B(0, 2t^\delta) \rightarrow \mathcal{M}$ defines a normal coordinate system in a neighborhood of x . Let $S : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a vector value function such that

$$K(x, \psi(v)) - S(v) = O(\|v\|^r),$$

and

$$K(x, y) = O(\|x - y\|^s).$$

Then, the following estimate holds:

$$P_{t,\delta}(x) = O((e^{C_2 t^{4\delta-2}} - 1)t^{s+d} + t^{r+d}) + \int_{\psi^{-1}(U(x,t^\delta))} S(v) e^{\frac{-\|T(v)\|^2}{2t^2}} dv.$$

where T is a rotation from \mathbb{R}^d onto the tangent space $T_x \mathcal{M}$.

Proof Using Equation (41), we assume that for small t , the set $U(x,t^\delta)$ lies within the image of a normal chart $\psi : B(0, 2t^\delta) \rightarrow \mathcal{M}$ centered in x . Therefore, we can write:

$$\begin{aligned} \int_{U(x,t^\delta)} K(x,y) e^{\frac{-\|y-x\|^2}{2t^2}} dVol(y) &= \int_{\psi^{-1}(U(x,t^\delta))} K(x, \psi(v)) e^{\frac{-\|\psi(v)-x\|^2}{2t^2}} dv \\ &= \int_{\psi^{-1}(U(x,t^\delta))} K(x, \psi(v)) (e^{\frac{-\|\psi(v)-x\|^2}{2t^2}} - e^{\frac{-\|T(v)\|^2}{2t^2}}) dv \\ &\quad + \int_{\psi^{-1}(U(x,t^\delta))} (K(x, \psi(v)) - S(v)) e^{\frac{-\|T(v)\|^2}{2t^2}} dv \\ &\quad + \int_{\psi^{-1}(U(x,t^\delta))} S(v) e^{\frac{-\|T(v)\|^2}{2t^2}} dv. \end{aligned}$$

We now estimate the first term, which we denote by

$$A := \int_{\psi^{-1}(U(x,t^\delta))} K(x, \psi(v)) (e^{\frac{-\|\psi(v)-x\|^2}{2t^2}} - e^{\frac{-\|T(v)\|^2}{2t^2}}) dv.$$

Using Equation (38), and the inequality $|e^x - 1| \leq e^{|x|} - 1$, we obtain

$$\begin{aligned} \left| e^{\frac{-\|\psi(v)-x\|^2}{2t^2}} - e^{\frac{-\|T(v)\|^2}{2t^2}} \right| &= e^{\frac{-\|T(v)\|^2}{2t^2}} \left| e^{\frac{O(\|v\|^4)}{2t^2}} - 1 \right| \\ &\leq e^{\frac{-\|T(v)\|^2}{2t^2}} (e^{\frac{C_1 \|v\|^4}{2t^2}} - 1). \end{aligned}$$

Therefore, by Equation (41) we obtain

$$\begin{aligned} \|A\| &\leq C_3 t^s (e^{C_2 t^{4\delta-2}} - 1) t^d \int_{\mathbb{R}^d} \|v\|^s e^{-\|v\|^2/2} dv \\ &= O((e^{C_2 t^{4\delta-2}} - 1) t^{s+d}). \end{aligned}$$

On the other hand, by assumption we have

$$\int_{\psi^{-1}(U(x,t^\delta))} (K(x, \psi(v)) - S(v)) e^{\frac{-\|T(v)\|^2}{2t^2}} dv = O(t^{r+d}).$$

■

Lemma B.2. Under the same assumptions as in Lemma B.1, consider the integral

$$E := \int_{\psi^{-1}(U(x,t^\delta))} Q(v) e^{\frac{-\|T(v)\|^2}{2t^2}} g(v) dv,$$

where g is a smooth function at 0 and Q is a homogeneous polynomial of degree l . Then, we have the following estimate:

$$E = \int_{\mathbb{R}^d} Q(v) e^{\frac{-\|T(v)\|^2}{2t^2}} \left(g(0) + \sum_{i=1}^d \frac{\partial g}{\partial v_i}(0) v_i \right) dv + O(t^{d+l} e^{-M_2 t^{2(\delta-1)}} + t^{d+2+l}).$$

Proof Using the Taylor expansion of g around 0 we have

$$E = \int_{\psi^{-1}(U(x, t^\delta))} Q(v) e^{-\frac{\|T(v)\|^2}{2t^2}} \left(g(0) + \sum_{i=1}^d \frac{\partial g}{\partial v_i}(0) v_i + O(\|v\|^2) \right) dv.$$

Next, define

$$B := \left\| \int_{\mathbb{R}^d \setminus \psi^{-1}(U(x, t^\delta))} Q(v) e^{-\frac{\|T(v)\|^2}{2t^2}} \left(g(0) + \sum_{i=1}^d \frac{\partial g}{\partial v_i}(0) v_i \right) dv \right\|.$$

Using Equation (41) and the rapid decay of the exponential function, we obtain the estimate

$$B \leq C_4 t^{d+l} e^{-M_2 t^{2(\delta-1)}} \int_{\mathbb{R}^d \setminus B(0, t^{\delta-1}/M_1)} P(\|v\|) e^{-\frac{\|T(v)\|^2}{4}} dv.$$

for some polynomial P . Hence

$$B = O(t^{d+l} e^{-M_2 t^{2(\delta-1)}}),$$

for an appropriate constant $M_2 > 0$. Finally, we observe that the contribution of the remainder term in the Taylor expansion satisfies

$$\int_{\psi^{-1}(U(x, t^\delta))} Q(v) e^{-\frac{\|T(v)\|^2}{2t^2}} O(\|v\|^2) dv = O(t^{d+2+l}).$$

■

Lemma B.3. *Under the same assumptions as in Lemma B.1, we obtain the following result*

$$\int_{\mathcal{M}} K(x, y) e^{-\frac{\|y-x\|^2}{2t^2}} dVol(y) = P_{t,\delta}(x) + O(t^{s+2(1-\delta)(d+2)}),$$

Proof By assumption, the expression

$$\left\| \int_{\mathcal{M} \setminus U(x, t^\delta)} K(x, y) e^{-\frac{\|y-x\|^2}{2t^2}} dVol(y) \right\| = \left\| \int_{\mathcal{M}} K(x, y) e^{-\frac{\|y-x\|^2}{2t^2}} dVol(y) - P_{t,\delta}(x) \right\|$$

is bounded from above by

$$F_1 \int_{\mathcal{M} \setminus U(x, t^\delta)} \|x - y\|^s e^{-\frac{\|y-x\|^2}{2t^2}} dVol(y) \tag{42}$$

for some constant $F_1 > 0$. Since the exponential decay dominates the polynomial growth at infinity, there exists a constant $F_2 > 0$ such that for all $z \in \mathbb{R}^n$

$$\|z\|^{s+2(d+2)} e^{-\frac{\|z\|^2}{2}} \leq F_2$$

Therefore, the expression in Equation (42) is bounded from above by

$$\begin{aligned} & F_1 F_2 \int_{\mathcal{M} \setminus U(x, t^\delta)} \frac{t^{s+2(d+2)}}{\|x - y\|^{2(d+2)}} dVol(y) \\ & \leq F_1 F_2 \int_{\mathcal{M} \setminus U(x, t^\delta)} t^{s+2(1-\delta)(d+2)} dVol(y) \\ & \leq F_1 F_2 t^{s+2(1-\delta)(d+2)} Vol(\mathcal{M}). \end{aligned}$$

■

We recall some standard computations involving the moments of the Gaussian distribution, which will be useful in the proof of Theorem 3.1. For all index i , we have

$$\int_{\mathbb{R}^d} v_i e^{-\frac{\|T(v)\|^2}{2t^2}} dv = 0,$$

and

$$\int_{\mathbb{R}^d} v_i^2 e^{-\frac{\|T(v)\|^2}{2t^2}} dv = (2\pi)^{\frac{d}{2}} t^{d+2}.$$

Moreover, if $i \neq j$,

$$\int_{\mathbb{R}^d} v_i v_j e^{-\frac{\|T(v)\|^2}{2t^2}} dv = 0.$$

These identities show that all odd moments vanish, and only the even-order moments contribute significantly. Consequently, we will focus on the even moments of the Gaussian distribution in what follows.

Lemma B.4. *Let $x \in \mathcal{M}$, and suppose $h : \mathcal{M} \rightarrow \mathbb{R}$ is a smooth function in x . Then*

$$\int_{\mathcal{M}} e^{-\frac{\|y-x\|^2}{2t^2}} h(y) dVol(y) = (2\pi)^{\frac{d}{2}} t^d h(x) + O(t^{d+4\delta-2}) + O(t^{2(1-\delta)(d+2)}),$$

Proof Let ψ be the map that defines normal coordinates at the point $x \in \mathcal{M}$. We apply Lemmas B.1, B.2, and B.3 to the functions $K(x, y) = h(y)$, $S(v) = h(\psi(v))$, $Q(v) = 1$, and $g(v) = h(\psi(v))$, using the parameters $r = 2$, $s = 0$ and $l = 0$. For any $\frac{1}{2} < \delta < 1$, Lemma B.3 guarantees:

$$\int_{\mathcal{M}} e^{-\frac{\|y-x\|^2}{2t^2}} h(y) dVol(y) = P_{t,\delta}(x) + O(t^{2(1-\delta)(d+2)})$$

On the other hand, by applying Lemmas B.1 and B.2, and using the rapid decay of the exponential function, we find that

$$P_{t,\delta}(x) = h(x) \int_{\mathbb{R}^d} e^{-\frac{\|T(v)\|^2}{2t^2}} dv + O(t^{d+4\delta-2})$$

We remark that in Lemma B.2, the integral involving the first-order partial derivatives of $g(v)$ vanishes due that the odd moments of the Gaussian are zero. Additionally, note that in the proof of this lemma, we used the fact that $g(0) = h(x)$. ■

Lemma B.5. *Under the same assumptions as in Lemmas B.1 and B.2, we obtain the following estimate for any $\frac{1}{2} < \delta < 1$:*

$$d_t(x) = (2\pi)^{\frac{d}{2}} t^d q(x) + O(t^{d+4\delta-2}) + O(t^{2(1-\delta)(d+2)}),$$

where d_t is defined in Equation (2).

Proof This follows directly from Lemma B.4 applied to the function $h(y) = q(y)$. ■

With these lemmas established, we are now ready to prove Theorem 3.1. The proof is presented below.

Proof of Theorem 3.1 Let e_1, \dots, e_d be the standard basis of \mathbb{R}^d , and let ψ be the map that defines the normal coordinates at the point $x \in \mathcal{M}$. Using the normal coordinate system, the k -differential form w can locally be written as:

$$w(\psi(v)) = \sum_I a_I(v) \frac{\partial \psi}{\partial v_{i_1}}(v) \wedge \dots \wedge \frac{\partial \psi}{\partial v_{i_k}}(v)$$

Moreover, since $\frac{\partial \psi}{\partial v_{i_j}}(v) = \mathcal{P}_{T_x \mathcal{M}} \left(\frac{\partial \psi}{\partial v_{i_j}}(v) \right) + \mathcal{P}_{T_x \mathcal{M}^\perp} \left(\frac{\partial \psi}{\partial v_{i_j}}(v) \right)$, we can expand the previous expression as

$$w(\psi(v)) = \sum_I a_I(v) \mathcal{P}_{T_x \mathcal{M}} \left(\frac{\partial \psi}{\partial v_{i_1}}(v) \right) \wedge \dots \wedge \mathcal{P}_{T_x \mathcal{M}} \left(\frac{\partial \psi}{\partial v_{i_k}}(v) \right) + L, \quad (43)$$

where L is the remaining term which involves the wedge product of some term of the orthogonal complement $\mathcal{P}_{T_x \mathcal{M}^\perp} \left(\frac{\partial \psi}{\partial v_{i_j}}(v) \right)$. Next, we use Equation (40) to further expand Equation (43):

$$w(\psi(v)) = \sum_I a_I(v) T(e_{i_1}) \wedge \dots \wedge T(e_{i_k}) + L + O(\|v\|^2).$$

Thus, the difference between the two forms $w(\psi(v)) - w(x)$, both viewed as multi-dimensional arrays in \mathbb{R}^n can be expanded as follows:

$$w(\psi(v)) - w(x) = \sum_I (a_I(v) - a_I(0)) T(e_{i_1}) \wedge \dots \wedge T(e_{i_k}) + L + O(\|v\|^2). \quad (44)$$

Since the term L involves the wedge product of elements in the orthogonal complement $T_x \mathcal{M}^\perp$, the orthogonal projection onto $\wedge^k T_x \mathcal{M}$ vanishes:

$$\mathcal{P}_{T_x \mathcal{M} \wedge \dots \wedge T_x \mathcal{M}}((\psi(v) - x) \wedge L) = 0. \quad (45)$$

By a similar argument, the following projection also vanishes:

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}((\mathcal{P}_{T_x \mathcal{M}^\perp}(\psi(v) - x)) \wedge (w(\psi(v)) - w(x))) = 0. \quad (46)$$

Using Equation (39), we have

$$\begin{aligned} \psi(v) - x &= \mathcal{P}_{T_x \mathcal{M}}(\psi(v) - x) + \mathcal{P}_{T_x \mathcal{M}^\perp}(\psi(v) - x) \\ &= T(v) + \mathcal{P}_{T_x \mathcal{M}^\perp}(\psi(v) - x) + O(\|v\|^3). \end{aligned} \quad (47)$$

Next, combining Equations (44) to (47), we obtain the following expression:

$$\begin{aligned} &\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}((\psi(v) - x) \wedge (w(\psi(v)) - w(x))q(\psi(v))) \\ &= \sum_I \sum_j v_j (a_I(v) - a_I(0)) T(e_j) \wedge T(e_{i_1}) \wedge \cdots \wedge T(e_{i_k}) q(\psi(v)) + O(\|v\|^3) \end{aligned} \quad (48)$$

Furthermore, since $\psi(v) - x = O(\|v\|^1)$ and a_I is smooth, Equation (44) shows that $w(\psi(v)) - w(x) = O(\|v\|^1)$. Thus, we have

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}((\psi(v) - x) \wedge (w(\psi(v)) - w(x))q(\psi(v))) = O(\|v\|^2) \quad (49)$$

We now apply the previous equations in conjunction with Lemmas B.1, B.2, and B.3 to complete the proof. Specifically, we use these lemmas for the following functions:

$$K(x, y) = \mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}((y - x) \wedge (w(y) - w(x))q(y)), \quad (50)$$

$$S(v) = \sum_I \sum_j v_j (a_I(v) - a_I(0)) T(e_j) \wedge T(e_{i_1}) \wedge \cdots \wedge T(e_{i_k}) q(\psi(v)),$$

$Q(v) = v_i$ and $g(v) = (a_I(v) - a_I(0))q(\psi(v))$. Note that $g(0) = 0$ and $\frac{\partial g}{\partial v_i}(0) = \frac{\partial a_I}{\partial v_i}(0)q(x)$. In this context, according to Equations (48) and (49), the parameters appearing in the hypotheses of the lemmas for $K(x, y)$ and $S(v)$ are $r = 3$, $s = 2$, and $l = 1$.

To prove the result, we first apply Lemma B.3 with the kernel K defined in Equation (50), which allows us to decompose the integral as follows:

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_t w(x)) = \frac{1}{d_t(x)} (P_{t,\delta}(x) + O(t^{2+2(1-\delta)(d+2)})) \quad (51)$$

for all $\frac{1}{2} < \delta < 1$, where $\mathbf{P}_t w(x)$ is defined in Equation (4). Note that the kernel $K(x, y)$ defined above is not to be confused with $K_t(x, y)$ in Equation (4).

Next, applying Lemmas B.1 and B.2 to the previous functions K and S , and using the rapid decay of the exponential function, we obtain the following expression:

$$\begin{aligned} P_{t,\delta}(x) &= \sum_I \sum_{j_1} \sum_{j_2} \left(\int_{\mathbb{R}^d} v_{j_1} v_{j_2} e^{-\frac{\|T(v)\|^2}{2t^2}} \frac{\partial a_I}{\partial v_{j_2}}(0) q(x) dv \right) (T(e_{j_1}) \wedge T(e_{i_1}) \wedge \cdots \wedge T(e_{i_k})) \\ &\quad + O(t^{d+2+(4\delta-2)}). \end{aligned} \quad (52)$$

Since the odd moments of the Gaussian are zero, the terms in Equation (52) for which $j_1 \neq j_2$ are vanish. Therefore, Equation (52) simplifies to:

$$\begin{aligned}
P_{t,\delta}(x) &= (2\pi)^{\frac{d}{2}} t^{d+2} q(x) \sum_I \sum_j \frac{\partial a_I}{\partial v_j}(0) (T(e_j) \wedge T(e_{i_1}) \wedge \cdots \wedge T(e_{i_k})) \\
&\quad + O(t^{d+2+(4\delta-2)}) \\
&= (2\pi)^{\frac{d}{2}} t^{d+2} q(x) \mathbf{d}(w)(x) + O(t^{d+2+(4\delta-2)}).
\end{aligned} \tag{53}$$

Now, combining Equation (51), Equation (53) and Lemma B.5, we obtain the following expression:

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_{\mathbf{t}} w(x)) = \frac{(2\pi)^{\frac{d}{2}} t^{d+2} q(x) \mathbf{d}(w)(x) + O(t^{d+2+(4\delta-2)}) + O(t^{2+2(1-\delta)(d+2)})}{(2\pi)^{\frac{d}{2}} t^d q(x) + O(t^{d+4\delta-2}) + O(t^{2(1-\delta)(d+2)})}. \tag{54}$$

This estimate holds for all $\frac{1}{2} < \delta < 1$. In particular, it holds for all δ satisfying the condition in Equation (5). For any such δ , the exponents in Equation (54) satisfy $2(1-\delta)(d+2) > d$ and $0 < 4\delta - 2 < 2$. Consequently, Equation (54) simplifies to:

$$\mathcal{P}_{\wedge^{k+1} T_x \mathcal{M}}(\mathbf{P}_{\mathbf{t}} w(x)) = t^2 (\mathbf{d}(w)(x) + O(t^f)),$$

where the exponent f is defined as $f = \min(4\delta - 2, 2(1-\delta)(d+2))$. ■