Topology from Data via Geodesic Complexes*

Tamal K. Dey† Kuiyu Li‡

March 17, 2009

Abstract

Recently several types of complexes have been proposed for topological analysis of data lying on a manifold in a high dimensional space. The effectiveness of the method in practice surely depends on the computational costs of constructing these complexes. The complexes such as restricted Delaunay, alpha complex, Čech and witness complex are difficult to compute in high dimensions. As an alternative, Rips complex, a well known structure in algebraic topology, has been proposed for computing homological information. While their computations are easy, their size tends to be large. We propose a Rips-like complex called *geodesic complex* which has smaller size than the standard Rips complex. The gain in size results from the fact that a geodesic complex is built by approximating intrinsic distances on the embedded manifold whereas a Rips complex is built with extrinsic distances in the embedding space. In the course of the development, we connect among various existing results which may find further use in topological analysis of data.

*Research supported by NSF grant CCF-0635008.

†Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: tamaldey@cse.ohio-state.edu

‡Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: tamaldey@cse.ohio-state.edu

1 Introduction

A considerable amount of interest has been generated recently in applying geometric and topological techniques to data analysis in high dimensional spaces. Assuming that the data is sampled from a low dimensional manifold lying in a high dimensional space, algorithms that 'learn' different properties of the manifold are the focus of these works. We are specifically interested in extracting the topology (homological information) of the manifold from its point data.

The low dimensional version of the problem known as curve and surface reconstruction in two and three dimensions have been studied vigorously in the past decade. Many concepts and techniques that ensure topological and geometric guarantees for the output have resulted from this endeavor, see e.g. [1, 2, 3, 8, 13]. This line of research also got extended to high dimensional space to the manifold learning problem [7, 8, 12]. These extensions are theoretically sound but are not practical mainly because they dwell on data structures such as Delaunay triangulations and alpha shapes that have impractically high computational cost in large dimensions. This is why alternative data structures such as witness complex, Čech complex, and Rips complexes have been suggested very recently [10, 22]. Among them it appears that Rips complex is an attractive choice as it can be computed more easily than the others. Taking this view point, Chazal and Oudot [10] show how one can build a hierarchy of Rips complexes from a point cloud data to compute the homology of the sampled manifold with topological persistence [15, 24]. Our work is motivated by this development.

Given a point data P sampled from a manifold $M \subset \mathbb{R}^d$, a Rips complex of P is computed by collecting all simplices whose edges have lengths less than an input parameter. The metric used for calculating lengths is taken as the metric of the embedding space which is the Euclidean space $\mathbb{R}^d$ here. We propose to replace this extrinsic metric with the intrinsic metric of the manifold M. The reason is that, the Rips complex, with the intrinsic metric is lighter in size than the one computed with the extrinsic metric; see Table 1. Unfortunately, it is not possible to compute lengths with the intrinsic metric since M is not given. We circumvent this problem by computing a graph connecting points in P that allows approximation of geodesic distances in M. A complex which we call *geodesic complex* is built using these approximate geodesic distances. We show that, geodesic complexes are interleaved with geodesic Čech complexes allowing computations of the homological ranks of M as in [10].

One of our main departures from the earlier methods is that we consider intrinsic metric of the manifold which has not been studied very closely in the context of topology detection in high dimensions. It has been used for other related problems in data analysis. We name a few. Tenenbaum et al. used intrinsic metric in their well known multidimensional scaling technique for dimensionality reduction [23]. Gao et al. [16] consider extracting topological information using complexes that use intrinsic metric but their study is restricted to two dimensional domains. Clarkson [6] presents several results that connect Riemannian geometry with various strategies of sampling manifolds in high dimensional spaces. In our case, we are not concerned with the sampling of the manifold but with deciphering topological properties of the manifold from a *given* sample. Nevertheless, concepts from Riemannian geometry play a key role in both cases.

Geodesics and some of their properties are essential for developing our algorithm. In section 2 we present these concepts and justify why a sampling condition defined via intrinsic metric is not too restrictive compared to a standard sampling condition with an extrinsic metric. In section 3 we introduce geodesic complex and their properties. In particular, we show that it interleaves with the Čech complex giving us an interleaved homology sequence from

which the homology of M can be derived by persistence. Following Chazal and Oudot [10], we build a sequence of subsamples of increasing size. This allows us to approximate the true geodesic distances at least for a range of scale. The algorithm and its justification is described in section 4. We conclude with a discussion in section 5 which alludes to possible extensions and future research.

2 Geodesics

Let $M \subset \mathbb{R}^d$ be a compact, smooth manifold without boundary. Assume that the metric in M is induced by the scalar product $< \cdot, \cdot >$ in $\mathbb{R}^d$. A curve $\gamma: I \subset \mathbb{R} \to M$ is a *geodesic* if the vector representing the rate of change of the tangent $\gamma'(t)$ has no component along M for all $t \in I$. More formally, the covariant derivative (defined by Riemannian connection) $\frac{D}{dt}(\gamma'(t))$ is 0 for all $t \in I$. Given a vector u in the tangent space TM_p at a point $p \in M$, there is a geodesic $\gamma(t)$ where $\gamma(0) = p$ and $\gamma'(0) = u/||u||$. We denote this geodesic as $\gamma(t, p, u)$. Notice that any two points p and q in M may have multiple geodesics between them. Among them, the ones minimizing the length(if they exist) are called the *minimizing geodesics* between p and q. Since M is compact, it is geodesically complete meaning that any two points admit a minimizing geodesic. One can define a distance metric $d_M: M \times M \to \mathbb{R}$ where $d_M(p, q)$ is the length of a minimizing geodesic between p and q in M. The usual Euclidean distance metric $d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfies $d(p, q) \leqslant d_M(p, q)$ for any $p, q \in M \subset \mathbb{R}^d$.

2.1 Geodesic radii

We will deal with geodesic balls that are counterpart of Euclidean balls defined with Euclidean metric. A *geodesic ball* $\mathsf{B}(p, r)$ of radius r centered at point $p \in M$ is the union of all points $x \in M$ so that $d_M(p, x) < r$. Notice that geodesic balls are open in M. The *exponential map* $\exp_p: TM_p \to M$ is defined as $\exp_p(u) = \gamma(||u||, p, u)$. This map projects an Euclidean ball $B(0, r)$ centered at $p = 0$ in TM_p with radius r to a geodesic ball $\mathsf{B}(p, r) \subset M$ if r is sufficiently small. In fact, this defines a well known intrinsic quantity $\rho_i(M)$ for M called *injectivity radius*. The injectivity radius $\rho_i(p)$ at $p \in M$ is the supremum of r so that the restriction $\exp_p: (B(0, r) \subset TM_p) \to (\mathsf{B}(p, r) \subset M)$ is a diffeomorphism. Define $\rho_i(M) = \inf_{p \in M}\{\rho_i(p)\}$.

We are interested in geodesic balls centered at p that have an additional property similar to the convexity property of the Euclidean balls. A set $X \subset M$ is *convex* if for any two points $p, q \in X$, there is a unique minimizing geodesic γ_{pq} between p and q and γ_{pq} is contained in X. We can define the *convexity radius* $\rho_c(p)$ at $p \in M$ as the supremum of r where $\mathsf{B}(p, r)$ is convex. Extending we define the convexity radius $\rho_c(M) = \inf_{p \in M}\{\rho_c(p)\}$.

For our results we will need that the input data is a dense sample of a smooth, compact, manifold $M \subset \mathbb{R}^d$ without boundary. In earlier works, this density is measured relative to an extrinsic distance such as *local feature size* [1, 2], *reach* [21], *weak feature size* [9] and other variants [8]. Here we will define density relative to the convexity radius, an intrinsic distance. A natural question is how are the two quantities related. Specifically, can the convexity radius be too small requiring a large number of samples to satisfy the density condition?

We find out that the convexity radius of M is not much smaller than the *reach* of M. Let $A(M)$ denote the medial axis of M. The reach $\rho(M)$ is $\inf_{x \in M, y \in A(M)}\{d(x, y)\}$. The relation between the convexity radius $\rho_c(M)$ and the reach $\rho(M)$ is derived via *sectional curvature* of M. Skipping a formal definition of sectional curvature (available in any standard Riemannian geometry book, e.g. [14]), we only mention that for a point $p \in M$, and two vectors $u, v \in TM_p$, the sectional curvature $\kappa_p(u, v)$ measures the Gaussian curvature of the surface formed by the

geodesics going through p and being tangent to the plane spanned by u and v. Let $\kappa(M) = \sup_{p,u,v}\{|\kappa_p(u,v)|\}$. The following result connects sectional curvature and reach.

Proposition 2.1 *For a smooth compact manifold $M \subset \mathbb{R}^d$ without boundary, $\kappa(M) \leqslant \frac{2}{\rho(M)^2}$.*

Proof. We use some of the concepts from Riemannian geometry and a result from [21] to prove this claim. Given two vectors $u, v \in TM_p$ from two vector fields on M, one has a bilinear symmetric form $B{:}TM_p \times TM_p \rightarrow TM_p^{\perp}$ that maps u, v to a vector in the normal space $TM_p^{\perp}$. Let η be any unit vector in the normal space $TM_p^{\perp}$ and u, v be any two unit vectors in the tangent space TM_p. It is known that there exists (see [21]) a linear self-adjoint operator $L_\eta{:}TM_p \rightarrow TM_p$ so that

$$< \eta, B(u,v) >=< u, L_\eta v > \text{ and } ||L_\eta|| \leqslant \frac{1}{\rho(M)}. \tag{1}$$

We claim that $|B(u,v)| \leqslant \frac{1}{\rho(M)}$. Indeed, for $\eta = B(u,v)/||B(u,v)||$

$$||B(u,v)|| =< \eta, B(u,v) >=< u, L_\eta v >\leqslant ||u||.||L_\eta||.||v|| \leqslant \frac{1}{\rho(M)}. \tag{2}$$

We extract another result from Riemannian geometry (see do Carmo [14], Theorem 2.5) which is due to Gauss.

$$\kappa_p(u,v) = \bar{\kappa}_p(u,v) + < B(u,u), B(v,v) > -||B(u,v)||^2. \tag{3}$$

where $\bar{\kappa}_p(u,v)$ is the sectional curvature of the embedding space containing M. Since $\mathbb{R}^d$ has zero sectional curvature at all points, we have $\bar{\kappa}_p(u,v) = 0$. Applying the inequality of 2 we get

$$|\kappa_p(u,v)| \leqslant | < B(u,u), B(v,v) > | + ||B(u,v)||^2 \leqslant ||B(u,u)||.||B(v,v)|| + ||B(u,v)||^2 \leqslant \frac{2}{\rho(M)^2}.$$

The claim of the proposition follows since $\kappa_p(u,v)$ is independent of the choice of the two vectors u, v for the plane spanned by u and v. ◻

The above proposition with a result in [11] provide a lower bound on the convexity radius.

Theorem 2.1 *For a smooth compact manifold $M \subset \mathbb{R}^d$, $\rho_c(M) \geqslant \min\{\frac{\rho_i}{2}, \frac{\pi\rho(M)}{\sqrt{2}}\}$.*

Proof. We extract from Chavel [11] that $\rho_c(M) \geqslant \min\{\frac{\rho_i(M)}{2}, \frac{\pi}{\sqrt{\kappa(M)}}\}$. This relation in combination with Proposition 2.1 gives the desired bound. ◻

The injectivity radius $\rho_i(M)$ is one of the fundamental intrinsic quantities of M and various lower bounds have been derived for different classes of M. We collect the following result from standard sources in Riemannian geometry [14, 19].

Proposition 2.2 *Let $M \subset \mathbb{R}^d$ be a compact, smooth manifold with positive sectional curvature $\kappa(M)$. Then, $\rho_i(M) \geqslant \min\{\frac{\pi}{\sqrt{\kappa(M)}}, \frac{\ell(M)}{2}\}$ where $\ell(M)$ is the length of the shortest non-trivial geodesic loop in M. In particular,*

- *if M is not simply connected, $\rho_i(M) \geqslant \frac{\pi}{2\sqrt{\kappa(M)}}$.*
- *if M is even dimensional and simply connected, $\rho_i(M) \geqslant \frac{\pi}{\sqrt{\kappa(M)}}$.*
- *if M is odd dimensional, simply connected, and $\kappa(M)/4 \leqslant \kappa_p(u,v) \leqslant \kappa(M)$, $\rho_i(M) \geqslant \frac{\pi}{\sqrt{\kappa(M)}}$.*

2.2 Čech complex and homology

Let P be a discrete subset of M. Our plan is to form a sequence of simplicial complexes using the points in P and compute the homology of M from the homology of these complexes. We will build a sequence of complexes called *geodesic complexes* out of the data points which we will show interleave with a sequence of geodesic Čech complexes. Then, applying an approach of Chazal and Oudot [10], we will be able to show how the homology of M can be computed from a pair of geodesic complexes. For this result, we need that the geodesic Čech complexes capture the topology of M. The use of geodesic balls and convexity radius play a key role in this respect as Lemma 2.1 and the discussion afterward show.

We are interested in convex geodesic balls. An useful property of these balls is that they intersect in convex sets. Let $\cap X_i$ denote the intersection of sets $\{X_i\}$.

Proposition 2.3 *The set $\cap \mathsf{B}(p_i, r_i)$, $i = 1, .., k$, is either empty or convex if $r_i \leqslant \rho_c(p_i)$.*

Proof. Let $x, y \in M$ be any two points in $\cap \mathsf{B}(p_i, r_i)$ if it is not empty. Since $r_i \leqslant \rho_c(p_i)$, the ball $\mathsf{B}(p_i, r_i)$ is convex. It follows that the unique minimizing geodesic between x and y in M belongs to each $\mathsf{B}(p_i, r_i)$ for $i = 1, ..., k$. Therefore, this geodesic also belongs to $\cap \mathsf{B}(p_i, r_i)$. ◻

Lemma 2.1 *The set $\cap \mathsf{B}(p_i, r_i)$, $i = 1, ..., k$, is either empty or contractible if $r_i \leqslant \rho_c(p_i)$.*

Proof. Let $X = \cap \mathsf{B}(p_i, r_i)$. If X is not empty, consider a point $p \in X$. Since X is convex by Proposition 2.3, the unique minimizing geodesic connecting p and any point $x \in X$ lies within X. This property of X makes it a subset of $\mathsf{B}(p, \rho_i(p))$ by a standard result on geodesics (Corollary to Proposition 2.2 in [14]). Then, by definition of injectivity radius $\rho_i(p)$, there is $W = \exp_p^{-1}(X)$ where the restriction of $\exp_p$ on W is a diffeomorphism. The set $W \subset TM_p$ has the property that, for any $u \in W$, $tu \in W$ where $0 \leqslant t \leqslant 1$. This is because if $\gamma(||u||, p, u) \in X$, so is $\gamma(t||u||, p, tu)$ by the convexity of X. But, then W being a star of p in TM_p is contractible. Then, by diffeomorphism, X is contractible. ◻

Lemma 2.1 is the basis of our claim that the geodesic Čech complexes built with appropriate parameters capture the topology of M. For a set of balls $\cup B = \{B(p_i, r_i)\}$, defined with any metric at $\{p_i \in P\}$, one has a natural simplicial complex given by the intersection pattern of these balls. The *Čech complex* $\mathcal{C}(P, \cup B)$ is defined by the collection of simplices $\{\sigma = [p_{i_1}, p_{i_2}, ..., p_{i_k}]\}$ where $\cap B(p_{i_j}, r_{i_j})$ is non-empty. The well known *Nerve Theorem* of Leray states that $\mathcal{C}(P, \cup B)$ is homotopy equivalent to $\cup B$ if each nonempty intersection $\cap B(p_{i_j}, r_{i_j})$ is contractible [20]. If the metric is Euclidean, $\cap B(p_{i_j}, r_{i_j})$ is convex since a set of Euclidean balls may intersect only in a convex set. Lemma 2.1 allows us to apply Nerve Theorem to geodesic balls with radii smaller than the convexity radius. The corresponding geodesic Čech complex becomes homotopy equivalent to M if the union of balls cover M.

The observation above could be the basis of an algorithm that could compute a geodesic Čech complex with an appropriate parameter and obtain the homology of M from it. However, there are two main difficulties in applying this idea. First, it is impossible to compute whether two geodesic balls intersect or not if M is not given. Second, determining the appropriate radii of the geodesic balls is nontrivial. For the first, we resort to the geodesic complex as defined later. For the second, we adopt a technique of interleaving complexes and hence homology groups for a range of radii as proposed in [10].

The homology groups of topological spaces are invariants under topological equivalence. We refer the reader to Hatcher [17] for definitions of homology groups. For a topological space

T, we denote its kth homology group by $\mathsf{H}_k(T)$. Assume that the coefficient ring over which homology is defined is a field so that $\mathsf{H}_k(T)$ is a vector space. The dimension of $\mathsf{H}_k(T)$ is the kth Betti number of T. A continuous map $f\colon T \to T'$ between two spaces T and T' induces a homomorphism $f^*\colon \mathsf{H}_k(T) \to \mathsf{H}_k(T')$ between their homology groups. In our case, f will be the inclusion map $i\colon T \subseteq T'$ which will induce a homomorphism $i^*\colon \mathsf{H}_k(T) \to \mathsf{H}_k(T')$.

Consider a set of geodesic balls $\mathsf{B}_i^\varepsilon = \mathsf{B}(p_i, \varepsilon)$ for a discrete set $P = \{p_i \in M\}$ where $\cup \mathsf{B}_i^\varepsilon = M$. Then, if $\varepsilon \leqslant \rho_c(M)$, each non-empty intersection $\mathsf{B}_{i_1}^\varepsilon \cap ... \cap \mathsf{B}_{i_j}^\varepsilon$ is contractible according to Lemma 2.1. Applying Nerve Theorem, we obtain that $\mathcal{C}(P, \cup \mathsf{B}_i^\varepsilon)$ is homotopy equivalent to M. Writing $\mathcal{C}^\varepsilon(P) = \mathcal{C}(P, \cup \mathsf{B}_i^\varepsilon)$ we get that $\mathcal{C}^\varepsilon(P)$ is homotopy equivalent to M if $\varepsilon \leqslant \rho_c(M)$.

The condition that $\cup \mathsf{B}_i^\varepsilon = M$ is fulfilled if P is dense in M. We define the density of P as follows.

Definition 1 *A discrete set $P \subset M$ is an ε-sample of M if each closed geodesic ball $\overline{\mathsf{B}(x, \varepsilon)}$, $x \in M$, contains at least one point in P. We say P is a tight ε-sample if P is an ε-sample and there is a $x \in M$ so that $\mathsf{B}(x, \varepsilon) \cap P = \emptyset$ but $\overline{\mathsf{B}(x, \varepsilon)} \cap P \neq \emptyset$.*

Clearly, if P is an ε-sample of M, $\cup \mathsf{B}(p_i, \alpha) = M$ for $\alpha > \varepsilon$. Since the homotopy equivalence translates to isomorphism at the homology levels, we have the following lemma.

Lemma 2.2 *Let P be an ε-sample of M. Then, $\mathsf{H}_k(\mathcal{C}^\alpha(P))$ is isomorphic to $\mathsf{H}_k(M)$ for $\varepsilon < \alpha \leqslant \rho_c(M)$.*

3 Geodesic complex

We cannot compute $\mathcal{C}^\alpha(P)$ since we do not know M. To overcome this difficulty, we propose a new complex called *geodesic complex* that interleaves the Čech complexes when parameterized by the density of P. The geodesic complex is built using an approximation of the geodesic distances with the Euclidean distances since we have no way of computing exact geodesic distances.

Let $\mathsf{G}^\delta(P)$ denote a graph with the vertex set in P where any two points $p, q \in P$ are joined by an edge if and only if $d(p, q) \leqslant \delta$. The *discrete geodesic distance* $d_{\mathsf{G}^\delta}(p, q)$ is defined as the shortest path distance between p and q in G^δ.

Definition 2 *Given two positive reals α, δ, a geodesic complex, $\mathcal{G}_\delta^\alpha(P)$, is a collection of simplices with vertices in P where a simplex σ is in $\mathcal{G}_\delta^\alpha(P)$ if and only if each edge pq of σ satisfies $d_{\mathsf{G}^\delta}(p, q) \leqslant \alpha$.*

Notice that $\mathcal{G}_\delta^\alpha(P)$ differs from the usual Rips complex in measuring the edge lengths which are given by the discrete geodesic distance metric d_{G^δ} instead of the Euclidean metric d.

Discrete geodesic distances approximate true geodesic distances in M if P samples M adequately. Consequently, the complex $\mathcal{G}_\delta^\alpha(P)$ captures topological information of M only if P is a dense sample of M, and α and δ are chosen appropriately. Before we establish an approximation of the geodesic distance d_M with the discrete geodesic distance d_{G^δ}, we need a result on approximating the geodesic distance $d_M(p, q)$ between two points $p, q \in M$ with the Euclidean distance $d(p, q)$. We use the following result of Bernstein et. al [4]. Let $\frac{1}{r_0} = \max_{\gamma, t}\{\|\ddot{\gamma}(t)\|\}$ where γ varies over all unit speed geodesics in M and $t \in \mathbb{R}$.

Proposition 3.1 *For any two points $p, q \in M$, if $d_M(p, q) \leqslant \pi r_0$, then $d(p, q) \geqslant 2r_0 \sin(\frac{d_M(p,q)}{2r_0})$.*

Lemma 3.1 *For any two points $p, q \in M$, if $d_M(p,q) \leqslant \rho(M)/2$, then $d(p,q) \geqslant \frac{9}{10} d_M(p,q)$.*

Proof. First, we observe that r_0 is at least $\rho(M)$. Recall the definition of η and $B(\cdot,\cdot)$ from the proof of Proposition 2.1. For any point $\gamma(t)$ on a geodesic γ one has

$$< \eta, B(\dot{\gamma}(t), \dot{\gamma}(t)) >=< \eta, \ddot{\gamma}(t) >\leqslant \frac{1}{\rho(M)}$$

which implies $||\ddot{\gamma}(t)|| \leqslant \frac{1}{\rho(M)}$. The claim follows.

Second, $\sin(t) \geqslant t - t^3/6$ for $t \geqslant 0$. Plugging this into the bound given by Proposition 3.1 and writing $\ell = d_M(p,q)$, we get

$$d(p,q) \geqslant (1 - \frac{\ell^2}{24 r_0^2})\ell \geqslant (1 - \frac{\ell^2}{24\rho(M)^2})\ell.$$

Since $\ell \leqslant \rho(M)/2$, we have

$$d(p,q) \geqslant (1 - \frac{1}{96})\ell \geqslant \frac{9}{10} d_M(p,q).$$

□

Notice that the choice of the factor $\frac{9}{10}$ is a little arbitrary. We could have taken the factor $\frac{95}{96}$ which would tighten other constants slightly. In approximating exact geodesic distances with the discrete geodesic distances we need both a lower and an upper bound on the approximation. For the lower bound we use Lemma 3.1 and a result of Niyogi et al. [21]. For the upper bound we use a result of Bernstein et al. [4] directly.

Lemma 3.2 *If P is an ε-sample of M, then $\frac{9}{10} d_M(p,q) \leqslant d_{\mathsf{G}^\delta}(p,q) \leqslant (1 + \frac{4\varepsilon}{\delta}) d_M(p,q)$ for any $p, q \in P$ and $4\varepsilon \leqslant \delta \leqslant \frac{\rho(M)}{4}$.*

Proof. Let $p = p_0, p_1, ..., p_k = q$ be the sequence of vertices on the shortest path between p and q in $\mathsf{G}^\delta(P)$. We have $d_{\mathsf{G}^\delta}(p,q) = \Sigma_{i=0}^{k-1} d(p_i, p_{i+1})$. To obtain a lower bound on $d_{\mathsf{G}^\delta}(p,q)$, we need a lower bound on $d(p_i, p_{i+1})$ for each i. We could apply Lemma 3.1 only if $d_M(p_i, p_{i+1})$ is at most $\rho(M)/2$. However, the assumption of the lemma puts an upper bound of $\rho(M)/4$ only on $d(p_i, p_{i+1})$ but not on $d_M(p_i, p_{i+1})$. To circumvent this difficulty, we apply a result Niyogi et al. [21]. It says that

$$d_M(x,y) \leqslant \rho(M) - \rho(M)\sqrt{1 - \frac{2d(x,y)}{\rho(M)}} \text{ if } d(x,y) \leqslant \rho(M)/2.$$

We can apply the above result since $d(p_i, p_{i+1}) \leqslant \delta \leqslant \rho(M)/4 \leqslant \rho(M)/2$ by assumption. Therefore, for each i,

$$d_M(p_i, p_{i+1}) \leqslant \rho(M) - \rho(M)\sqrt{1 - \frac{2d(p_i, p_{i+1})}{\rho(M)}} \leqslant 2d(p_i, p_{i+1}) \leqslant \rho(M)/2.$$

Now applying Corollary 3.1, we get

$$d_{\mathsf{G}^\delta}(p,q) = \Sigma_{i=0}^{k-1} d(p_i, p_{i+1}) \geqslant \Sigma_{i=0}^{k-1} \frac{9}{10} d_M(p_i, p_{i+1}) \geqslant \frac{9}{10} d_M(p,q).$$

The upper bound on $d_{\mathsf{G}^\delta}(p,q)$ follows directly from Theorem 2 in [4]. This theorem requires that, P is an ε-sample of M, $\delta \geqslant 4\varepsilon$, and $\mathsf{G}^\delta(P)$ contains all edges pq for which $d_M(p,q) \leqslant \delta$. All these conditions are satisfied here. ◻

Our strategy will be to construct a graph $\mathsf{G}^\delta(P)$ and hence $\mathcal{G}_\delta^\alpha(P)$ with $\delta \geqslant 4\varepsilon$ where P is an ε-sample. By appealing to Lemma 3.2, we show that this complex is sandwiched between two Čech complexes.

Lemma 3.3 *Let P be an ε-sample of M where $4\varepsilon \leqslant \delta \leqslant \frac{\rho(M)}{4}$. Following inclusions hold:*

$$\mathcal{C}^{\frac{\alpha}{4}}(P) \subseteq \mathcal{G}_\delta^\alpha(P) \subseteq \mathcal{C}^{\frac{10\alpha}{9}}(P).$$

Proof. The geodesic complex $\mathcal{G}_\delta^\alpha(P)$ connects any two $p, q \in P$ where $d_{\mathsf{G}^\delta}(p,q) \leqslant \alpha$. It follows from Lemma 3.2 that $d_M(p,q) \leqslant \frac{10}{9}\alpha$. Consider a simplex $[p_0, p_1, .., p_k]$ in $\mathcal{G}_\delta^\alpha(P)$. For each $p_i \in \{p_0, .., p_k\}$, the geodesic ball $\mathsf{B}(p_i, \frac{10}{9}\alpha)$ contains all points in $\{p_0, ..., p_k\}$ since $d_M(p_i, p_j)$ is at most $\frac{10}{9}\alpha$ for all $j \in \{0, .., k\}$. Hence $\mathcal{G}_\delta^\alpha(P) \subseteq \mathcal{C}^{\frac{10\alpha}{9}}(P)$.

To show the other inclusion, consider a simplex $\sigma \in \mathcal{C}^{\frac{\alpha}{4}}(P)$. Any edge pq of this simplex satisfies $d_M(p,q) \leqslant \frac{\alpha}{2}$. Therefore, this simplex appears in $\mathcal{G}_\delta^\alpha(P)$ since $d_{\mathsf{G}^\delta}(p,q) \leqslant 2d_M(p,q)$ for $\delta \geqslant 4\varepsilon$ according to the right inequality in Lemma 3.2. ◻

Lemma 3.4 *Let P be an ε-sample of M where $4\varepsilon \leqslant \delta \leqslant \frac{\rho(M)}{4}$. Following sequence of homomorphisms between homology groups is induced by inclusions:*

$$\mathsf{H}_k(\mathcal{C}^{\frac{\alpha}{4}}(P)) \to \mathsf{H}_k(\mathcal{G}_\delta^\alpha(P)) \to \mathsf{H}_k(\mathcal{C}^{\frac{10\alpha}{9}}(P)) \to \mathsf{H}_k(\mathcal{G}_\delta^{\frac{40\alpha}{9}}(P)) \to \mathsf{H}_k(\mathcal{C}^{\frac{400\alpha}{81}}(P)) \tag{4}$$

Furthermore, if $4\varepsilon \leqslant \alpha \leqslant \frac{81}{400}\rho_c$, and $i^*\colon \mathsf{H}_k(\mathcal{G}_\delta^\alpha(P)) \to \mathsf{H}_k(\mathcal{G}_\delta^{\frac{40\alpha}{9}}(P))$, then (image i^*) *is isomorphic to $\mathsf{H}_k(M)$ written as* (image i^*) $\approx \mathsf{H}_k(M)$.

Proof. The sequence of homomorphisms is induced by inclusions in the respective complexes which is asserted by Lemma 3.3. Furthermore, if $4\varepsilon \leqslant \alpha \leqslant \frac{81}{400}\rho_c$, Lemma 2.2 implies

$$\mathsf{H}_k(\mathcal{C}^{\frac{\alpha}{4}}(P)) \approx \mathsf{H}_k(\mathcal{C}^{\frac{10\alpha}{9}}(P)) \approx \mathsf{H}_k(\mathcal{C}^{\frac{400\alpha}{81}}(P)) \approx \mathsf{H}_k(M).$$

So, in the sequence 4, $\mathsf{H}_k(\mathcal{G}_\delta^\alpha(P))$ and $\mathsf{H}_k(\mathcal{G}_\delta^{\frac{40\alpha}{9}}(P))$ are sandwiched between three homology groups whose ranks are equal to the rank of $\mathsf{H}_k(M)$. Following the approach of [10] one can show that $\operatorname{rank}(\mathsf{H}_k(\mathcal{G}_\delta^\alpha(P)) \to \mathsf{H}_k(\mathcal{G}_\delta^{\frac{40\alpha}{9}}(P))) = \operatorname{rank}(\mathsf{H}_k(M))$. Since we are working on homology groups that are vector spaces (coefficient ring is a field), the equality in their ranks implies isomorphism between them. ◻

4 Algorithm

We wish to apply Lemma 3.4 to determine the homology group of M from $P \subset M$. For this, we need $\delta \geqslant 4\varepsilon$. Also, δ should not be too large compared to ε. For otherwise α, which we will take as a factor of δ, will not satisfy the upper bound of $\frac{81}{400}\rho_c$ as needed in Lemma 3.4.

In essence, we need an estimation of ε where P is an ε-sample. Often ε is estimated with k-nearest neighbor distances where the choice of k is somewhat arbitrary. This is recognized as one of the main problems in well known data analysis algorithms such as MDS or PCA [18, 23]. We propose to subsample P to estimate the right scale for computing the geodesic graph. We follow a strategy of Chazal and Oudot [10] for building a series of geodesic complexes from the given data. Subsamples are used both for building a series of geodesic complexes and also for computing an appropriate geodesic graph.

We compute a nested sequence of subsamples $\{p_0\} = L_0 \subset L_1 \subset ... \subset L_k = P$ where $L_{i+1} = L_i \cup \{p_{i+1}\}$ with p_{i+1} being the furthest point in $P \setminus L_i$ from L_i. Next lemma is the key to estimating δ for building $\mathsf{G}^\delta(L_i)$. Let Q_p be the set of points in $P \setminus L_i$ which have $p \in L_i$ as the closest point among all points in L_i. Let q be the furthest from p among all points in Q_p. Let $\delta_p = d(p, q)$. If Q_p is empty, take δ_p to be 0. Define $\delta = \max_p\{\delta_p\}$.

Lemma 4.1 *If $L_i \subset P$ is a tight ε_i-sample and P is an ε-sample of M respectively, then for $\varepsilon < \varepsilon_i \leqslant \rho(M)/2$, one has $\frac{9}{10}(\varepsilon_i - \varepsilon) \leqslant \delta \leqslant \varepsilon_i$.*

Proof. Consider a point $x \in M$ so that $d_M(x, L_i) = \varepsilon_i$. Since L_i is a tight ε_i-sample such a point exists. Let w be the closest point to x in $P \setminus L_i$. We claim that w is also the closest point to x in P. If not, there is a point in L_i which is closest to x in P. Then, $d_M(x, L_i) \leqslant \varepsilon$ contradicting that $\varepsilon < \varepsilon_i$.

We have $d_M(w, x) \leqslant \varepsilon$ since P is an ε-sample. Let p be the closest point to w in L_i. Then,

$$d_M(w, p) \geqslant d_M(x, p) - \varepsilon \geqslant d_M(x, L_i) - \varepsilon = \varepsilon_i - \varepsilon.$$

Since L_i is an ε_i-sample, $d_M(w, p) \leqslant \varepsilon_i \leqslant \frac{\rho(M)}{2}$. We can apply Lemma 3.1 to claim $d(w, p) \geqslant \frac{9}{10} d_M(w, p)$. Then, we have

$$\delta \geqslant d(w, p) \geqslant \frac{9 d_M(w, p)}{10} \geqslant \frac{9}{10}(\varepsilon_i - \varepsilon).$$

This proves the lower bound on δ.

To prove the upper bound, consider the pair (u, p), $p \in L_i$, $u \in P \setminus L_i$ which realizes the distance $\delta = \delta_p$. Since L_i is an ε-sample and $u \in M \setminus L_i$,

$$\delta = \delta_p \leqslant d_M(u, p) \leqslant \varepsilon_i.$$

□

The algorithm TOPODATA as delineated below computes the rank of the persistent homology group of a pair of geodesic complexes built hierarchically. For a large range (range of ε_i), this rank of the persistent homology coincides with the rank of M as Theorem 4.1 shows. This means one can run TOPODATA on a data set and check the range of i for which persistent homology between pairs of geodesic complexes built from L_i remains stable. The computed persistent Betti number in this range is the rank of $\mathsf{H}_k(M)$. Persistent Betti numbers can be computed by the standard algorithm; see [15, 24].

TOPODATA(P, k)

1. Initialize $L = \emptyset$;
2. While $L \neq P$ do
   (a) compute $p := argmax_{q \in P} \min_{r \in L} d(q, r)$;
   (b) $L := L \cup \{p\}$; $P := P \setminus \{p\}$; compute $\delta := \max_{q \in P} \min_{r \in L} d(q, r)$;
   (c) Compute $\mathsf{G}^{5\delta}$;
   (d) Compute persistence between $\mathcal{G}^{5\delta}_{5\delta}(L)$ and $\mathcal{G}^{\frac{200\delta}{9}}_{5\delta}(L)$;
3. endwhile

Theorem 4.1 *Given an ε-sample P of a smooth compact manifold $M \subset \mathbb{R}^d$ without boundary,* TOPODATA*(P) computes the rank of $\mathsf{H}_k(M)$ when $L \subseteq P$ is an ε_i-sample for $\frac{14}{9}\varepsilon \leqslant \varepsilon_i \leqslant \frac{81}{2000}\min\{\rho_c, \rho(M)\}$.*

Proof. By Lemma 4.1, we have $\frac{9}{10}(\varepsilon_i - \varepsilon) \leqslant \delta \leqslant \varepsilon_i$ from which we get $4\varepsilon_i < 5\delta \leqslant 5\varepsilon_i$ for $\varepsilon \leqslant \frac{9}{14}\varepsilon_i$. Since $5\varepsilon_i \leqslant \frac{81}{400}\rho(M) \leqslant \frac{\rho(M)}{4}$ by assumption, the graph $\mathsf{G}^{5\delta}$ satisfies the conditions of Lemma 3.2 which enables us to approximate the geodesic distance d_M by its discrete approximation with $d_{\mathsf{G}^{5\delta}}$. Therefore, we can apply Lemma 3.4 to claim the following sequence of homomorphisms:

$$\mathsf{H}_k(\mathcal{C}^{\frac{5\delta}{4}}(L)) \to \mathsf{H}_k(\mathcal{G}^{5\delta}_{5\delta}(P)) \to \mathsf{H}_k(\mathcal{C}^{\frac{50\delta}{9}}(L)) \to \mathsf{H}_k(\mathcal{G}^{\frac{200\delta}{9}}_{5\delta}(L)) \to \mathsf{H}_k(\mathcal{C}^{\frac{2000\delta}{81}}(L))$$

Then, by Lemma 3.4, $rank(\mathsf{H}_k(\mathcal{G}^{5\delta}_{5\delta}(L)) \to \mathsf{H}_k(\mathcal{G}^{\frac{200\delta}{9}}_{5\delta}(L)) = rank(\mathsf{H}_k(M))$ proving the conclusion of the lemma only if $\frac{4\varepsilon_i}{5} \leqslant \delta \leqslant \frac{81}{2000}\rho_c$. The lower bound is satisfied since $\delta \geqslant \frac{9}{10}(\varepsilon_i - \varepsilon) > \frac{4\varepsilon_i}{5}$ for $\varepsilon \leqslant \frac{9}{14}\varepsilon_i$. Since $\delta \leqslant \varepsilon_i$, the upper bound is satisfied if $\varepsilon_i \leqslant \frac{81}{2000}\rho_c$ which is an upper limit of the stated range of ε_i. ∎

Notice that we can tighten the constants so that one needs to consider the persistent homology between $\mathcal{G}^{(4+\tau)\delta}_{\delta}(L)$ and $\mathcal{G}^{(16+\tau')\delta}_{\delta}(L)$ for some small constants $\tau < \tau' < 1$. This can be achieved by considering a graph $\mathsf{G}^{(4+\tau)\delta}$ and requiring ε to be small enough to satisfy $(4 + \tau)\delta \geqslant 4\varepsilon_i$.

5 Discussions

In this section we discuss various issues related to this work.

Complexity: Since geodesic complexes have smaller size than the Rips complexes (with the same parameter), the time complexity analysis in [10] carries over here. Therefore, the computations of the geodesic complexes including the persistence take $O(c(m)|L|^4)$ time for each iteration as was shown in [10]. Here $c(m)$ is a quantity that depends solely on the dimension m of the manifold M. The computation of the geodesic graph cannot take more than $O(|L|^2)$. Thus, the overall complexity of the algorithm is same as that of the algorithm in [10] which is $O(c(m)n^5)$ where n is the number of given data points. However, the difference in size between geodesic and Rips complexes (see Table 1) will have an impact on the running time in practice. We expect that this difference is accentuated further in high dimensions.

Noise: We assumed that the point data is noise-free, that is, they lie on M. If not, we can assume that P lies within a small distance from M. Let $p^\perp$ be the closest point (in Euclidean metric) of M for a point $p \in P$ and let $P^\perp = \{p^\perp : p \in P\}$. We say P is an ε-sample of M if $P^\perp$ is an ε-sample of M and $d(p, p^\perp) \leqslant \varepsilon$ for each $p \in P$. Clearly, the inclusions stated in Lemma 3.3 hold with the point set $P^\perp$. It can be shown that there exist constants c_1, c_2 so that $c_1 d(p, q) \leqslant d(p^\perp, q^\perp) \leqslant c_2 d(p, q)$. This leads to a sequence of inclusions $\mathcal{C}^{c_1'\alpha}(P^\perp) \subseteq \mathcal{G}_\delta^\alpha(P) \subseteq \mathcal{C}^{c_2'\alpha}(P^\perp)$ for some appropriate constants c_1', c_2' and δ where we assume that p and $p^\perp$ represent the same vertex at the complex level. It is not hard to observe that this sequence can lead to a version of Lemma 3.4 for noisy data. Therefore, TOPODATA still works with noisy data albeit with the constants and constraints being adjusted appropriately.

Model	**Vertices**	$\frac{\alpha}{\delta}$	**Geodesic Complex**	**Euclidean Rips Complex**
Double-torus	12286	3	1126429	1504055
		5	3193262	4705078
		8	8475571	13314353
Genus3	18633	3	1838652	2627464
		5	5160445	7830444
		8	13489960	21371696
Botijo	23607	3	2235874	2773734
		5	6371507	8440593
		8	16809920	21778008

Table 1: Difference in number of edges between geodesic and Rips complexes for point data on surfaces in three dimensions. As the ratio $\frac{\alpha}{\delta}$ increases, the size difference becomes more prominent.

Extensions: One obvious extension of our method would be to accommodate larger class of inputs such as manifolds with boundary, piecewise smooth manifolds, and even larger class such as metric spaces. The theory of geodesics for smooth manifolds with boundaries exist and therefore may be applied in our approach. However, extending these theories to other classes with computational methods remains a challenging open problem.

We assumed that the manifold is embedded in Euclidean space. How about other embedding spaces? If the metric of the embedding space is specified by some mechanism such as a distance matrix for the data, one can apply the method of this paper assuming that the embedding space induces a metric on the manifold. However, it would be a non-trivial exercise to extend all required results such as approximating geodesic distances with distances in the embedding spaces.

The geodesic complex clearly reduces the size of the Rips complex built on the extrinsic distances. Table 1 shows some comparison data between the number of edges in the two complexes in three dimensions. It would be still interesting to reduce the size even further by considering some other complex. Is there a sub-complex of the geodesic complex that still captures the topology of the manifold? We hope to address these questions in future research.

References

[1] N. Amenta and M. Bern. Surface reconstruction by Voronoi filtering. *Discr. Comput. Geom.* **22** (1999), 481–504.

[2] N. Amenta, M. Bern, and D. Eppstein. The crust and the β-skeleton: combinatorial curve reconstruction. *Graphic. Models Image Process.* (1998), 125–135.

[3] N. Amenta, S. Choi, T. K. Dey and N. Leekha. A simple algorithm for homeomorphic surface reconstruction. *Internat. J. Comput. Geom. Applications* **12** (2002), 125–141.

[4] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Tech Report*, Dept. Psychology, Stanford University, USA, 2000. Available at http://isomap.stanford.edu/BdSLT.pdf

[5] J.-D. Boissonnat, L. J. Guibas, and S. Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Proc. 23rd Ann. Sympos. Comput. Geom.* (2007), 194–203.

[6] C. Clarkson. Building triangulations using ε-nets. *Proc. 38th Sympos. Theory Comput.* (2006).

[7] F. Chazal and A. Lieutier. Topology guaranteeing manifold reconstruction using distance function to noisy data. *Proc. 22nd Ann. Sympos. Comput. Geom.* (2006), 112–118.

[8] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in Euclidean space. *Proc. 22nd Ann. Sympos. Comput. Geom.* (2006), 319–326.

[9] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discr. Comput. Geom.* **37** (2007), 601–617.

[10] F. Chazal and S. Oudot. Towards persistence-based reconstruction in Euclidean spaces. *Proc. 24th Ann. Sympos. Comput. Geom.* (2008), 232–241.

[11] I. Chavel. Riemannian Geometry: A Modern Introduction. Cambridge U. Press, New York, 1994.

[12] S.-W. Cheng, T. K. Dey, and E. A. Ramos. Manifold reconstruction from point samples. *Proc. 16th Sympos. Discrete Algorithms* (2005), 1018–1027.

[13] T. K. Dey. Curve and Surface Reconstruction : Algorithms with Mathematical Analysis. Cambridge University Press, New York, 2007.

[14] M. P. do Carmo. Riemannian geometry. Birkhäuser, Boston, 1992.

[15] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511–533.

[16] J. Gao, L. J. Guibas, S. Y. Oudot, and Y. Wang. Geodesic Delaunay triangulations and witness complexes in the plane. *Proc. ACM-SIAM Sympos. Discrete Algorithms* (2008), 571–580.

[17] A. Hatcher. Algebraic Topology. Cambridge U. Press, New York, 2002.

[18] I. T. Jollife. Principal Component Analysis. Springer series in statistics, Springer, NY, 2002.

[19] W. P. A. Klingenberg. Riemannian Geometry. Walter de Gruyter, Berlin, 1995.

[20] J. Leray. Sur la forme des espaces topologiques et sur les points fixes des représentations. *J. Math. Pure Appl.* **24** (1945), 95–167.

[21] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* (2006).

[22] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Proc. Sympos. Point-Based Graph.* (2004), 157–166.

[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science* **290** (2000).

[24] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discr. Comput. Geom.* **33** (2005), 249–274.