Topological inference of manifolds with boundary

Yuan Wang [a,1], Bei Wang [b,*,2]

[a] *Northwestern University, Evanston, IL, USA*
[b] *University of Utah, Salt Lake City, UT, USA*

A B S T R A C T

Given a set of data points sampled from some underlying space, there are two important challenges in geometric and topological data analysis when dealing with sampled data: reconstruction – how to assemble discrete samples into global structures, and inference – how to extract geometric and topological information from data that are high-dimensional, incomplete and noisy. Niyogi et al. (2008) have shown that by constructing an offset of the samples using a suitable offset parameter could provide reconstructions that preserve homotopy types therefore homology for densely sampled smooth submanifolds of Euclidean space without boundary. Chazal et al. (2009) and Attali et al. (2013) have introduced a parameterized set of sampling conditions that extend the results of Niyogi et al. to a large class of compact subsets of Euclidean space. Our work tackles data problems that fill a gap between the work of Niyogi et al. and Chazal et al. In particular, we give a probabilistic notion of sampling conditions for manifolds with boundary that could not be handled by existing theories. We also give stronger results that relate topological equivalence between the offset and the manifold as a deformation retract.

© 2019 Published by Elsevier B.V.

1. Introduction

In *manifold learning*, a topic of high interest is to understand the structure of low-dimensional objects embedded in high-dimensional space. Such objects are typically assumed to be (sub)manifolds of Euclidean spaces. In recent years, it is becoming clear that the offsets of sampled data points on a manifold can reflect the geometric and topological structure of the manifold itself (e.g. [1–3]). In particular, given sampled points drawn from a probability distribution that has support on or near a submanifold without boundary, Niyogi et al. [3] have shown that one can learn the homology of the submanifold with high confidence. More specifically, for a compact manifold $\mathcal{M}$ embedded in Euclidean space $\mathbb{R}^N$ and a set of randomly sampled data points $\mathcal{X} = \{x_1, ..., x_n\}$ on $\mathcal{M}$, let $U = \bigcup_{x \in \mathcal{X}} B_\epsilon(x)$ be the offset of the data set $\mathcal{X}$, where $B_\epsilon(x)$ is a ball of radius ϵ centered at x and ϵ is chosen to be small relative to the minimum local feature size of $\mathcal{M}$. Then for any $\gamma \in (0, 1)$, there is a number m such that for all $n > m$, with probability γ, $\mathcal{M}$ is a deformation retract of U. Therefore the homology of U equals the homology of $\mathcal{M}$ (see [3, Theorem 3.1] for details).

Data that arise from smooth compact manifolds have been well-studied. However, the study of more complex spaces that are not necessarily manifolds via data samples seems much more difficult. When samples arise not from manifolds but from mixtures of manifolds with possible singularities, we are dealing with the notion of *stratification learning*. Roughly

speaking, a *stratified space* is a space that can be decomposed into manifold pieces (referred to as *strata*) that are glued together in a nice way. The study of stratified spaces is a classic topic in pure mathematics [4,5]. Statistical approaches rely on inferences of mixture models, local dimension estimation and subspace clustering [6–8]. In geometric and topological data analysis, progress has been made recently in the study of stratified spaces in the discrete and noisy settings [9–14], which draw inspirations from computational topology [15], intersection homology [16–18], graph theory and sheaf theory.

Among stratified spaces, manifolds with boundary is one of the simplest forms. A manifold with boundary is a stratified space: one stratum is its boundary, and the other stratum is the complement of such a boundary.

In this paper, we study the topology of offsets of data points on compact differentiable manifolds with boundary. We give a probabilistic notion of sampling conditions for manifolds with boundary that could not be handled by existing theories. In particular, we show that, with some care, a similar statement as [3, Theorem 3.1] holds for manifolds with boundary. We also demonstrate via simple examples how our sampling lower bounds could be derived in practice (Section 5).

The main result of this paper, Theorem 3.1, is proved by following the framework of [3]. First, we prove that the offset of data points deformation retracts to the manifold if the sample is sufficiently dense with respect to the local feature size of the manifold and the radius of the offset (Theorem 3.2). Second, we show that such density is achieved with high confidence when the data points are sufficiently abundant (Theorem 3.3). However, our proof in detail is very different from that of [3]: particular efforts have been made to overcome the complexity caused by the boundary. When a data point is near or on the boundary, the local geometry around it is more complicated, and the original ways [3] of performing deformation retract and estimating distances and volumes no longer work. We found that, through our arguments, such issues can be resolved by imposing concise and reasonable requirements regarding the minimum local feature size of *both* the manifold and its boundary and the (local) projection of the manifold to its tangent spaces (see Condition (1) and (2)).

It is worth noting that Chazal et al. [19] and Attali et al. [20] have extended the results of Niyogi et al. to a large class of compact subsets of Euclidean space. Specifically, [19] shows that the offset K^β of a compact set K is homotopy equivalent to the offset K'^α of another compact set K', for sufficiently small $\beta > 0$, if α satisfies some inequality involving the μ-reach of K' and the Hausdorff distance between K and K'; [20] shows that the offset X^η of a compact set X is homotopy equivalent to the Čech (respectively the Rips) complex of certain radius of a point cloud P in X if some inequality involving η, the radius and the μ-reach of X holds. Compared to the result of [19] and [20], the one in this paper only deals with compact manifolds with boundary. However, our result does have its own advantages. First of all, the topological equivalence between the offset of data points and the manifold we get here is a deformation retract, which is stronger than homotopy equivalence derived in [19] and [20]. Next, we get a probability estimate for the topological equivalence which is not provided in [19] and [20]. Last but not least, it appears that there are elementary cases of data on manifolds with boundary where the result in [19] does not apply, as the parameters associated with the data are completely outside the scope of [19, Theorem 4.6] and [20, Theorem 13 and 14]. A scenario is discussed in Section 3 (Example 3.4), where our Theorems 3.1 and 3.2 become applicable and work well. Finally, Turner [21] has also studied the reconstruction problem of compact sets with point clouds from the perspective of deformation retraction. Specifically, given two compact sets, one is a finite point cloud sampled from the other, [21, Theorem 4.3] gives a sufficient condition for when certain offsets of these two compact sets are homotopic in terms of the absence of μ-critical points in an annular region. We demonstrate in Example 3.4 that there exist scenarios that are outside the scope of [21, Theorem 4.3] where our results are applicable.

Result at a glance. In short, this paper improves our understanding of topological inference for manifolds with boundary and therefore enriches the toolbox for topological data analysis. Given a sample of n points from a differentiable manifold $\mathcal{M}$ with boundary in a high-dimensional Euclidean space, for a sufficiently large n, the ϵ-offset of the sample points is shown to have the same homotopy type as and deformation retracts to $\mathcal{M}$. The homotopy equivalence result has been proved by Niyogi et al. [3] for manifolds without boundary. Chazal et al. [19] have extended the result of Niyogi et al. to manifolds with boundary, however with weaker conclusions than the original paper. The current paper, instead, reaches the same conclusion as Niyogi et al. [3] for manifold with boundary, but under two mild conditions for the boundary (i.e., regarding the minimum feature size of the boundary and the uniform smoothness of the tangent bundle). A specific example shows that the new method is more powerful than that of Chazal et al. [19]. Therefore, our results fill a gap between the original work of Niyogi et al. [3] and the broader but weaker result of Chazal et al. [19]. Our proof techniques are nontrivial, although they frequently use the results of Niyogi et al. and make necessary adjustments. The theoretical results are complemented by experiments that confirm the theoretical findings.

2. Notations and preliminaries

2.1. Basics on manifolds

In this paper, for two points $p, q \in \mathbb{R}^N$, we use $|pq|$ to denote the line segment connecting p and q, $\overrightarrow{pq}$ to denote the vector from p to q, and $\|p - q\|$ to denote the Euclidean distance between p and q. For a set $K \subset \mathbb{R}^N$, $d(p, K) := \inf\{\|p - q\| : q \in K\}$ denotes the distance from $p \in \mathbb{R}^N$ to K. Moreover, for a non-negative real number $\alpha \in \mathbb{R}_{\geq 0}$, we use K^α to denote the offset of K with radius α, $K^\alpha = \{p \in \mathbb{R}^N \mid d(p, K) \leq \alpha\}$. $B_r(p)$ denotes the open ball with center p and radius r.

Let $\mathcal{M} \subseteq \mathbb{R}^N$ be a compact, *differentiable*, k-dimensional manifold possibly with boundary. Let $\partial\mathcal{M}$ denote the boundary of $\mathcal{M}$. Then $\partial\mathcal{M}$ is a compact manifold. Let $\mathcal{M}^\circ$ denote the interior of $\mathcal{M}$.

The *local feature size* of $\mathcal{M}$ is the function $\text{lfs}: \mathcal{M} \to \mathbb{R}_{\geq 0}$ defined by the distance from a point $x \in \mathcal{M}$ to its medial axis. The infimum of lfs is the reach of $\mathcal{M}$, $\text{reach}(\mathcal{M})$. For every number $0 < r < \text{reach}(\mathcal{M})$, the normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^N$. In the same way we define $\text{reach}(\partial \mathcal{M})$, and since $\partial \mathcal{M}$ is also a compact manifold, $\text{reach}(\partial \mathcal{M})$ is well-defined.

We use $\varphi_{p,\mathcal{M}}$ to denote the natural projection from $\mathcal{M}$ to $T_p(\mathcal{M})$, the tangent space to $\mathcal{M}$ at the point p. That is, $\varphi_{p,\mathcal{M}}: \mathcal{M} \to T_p(\mathcal{M})$. Conversely for any point $q \in T_p(\mathcal{M})$ we use $\varphi_{p,\mathcal{M}}^{-1}(q)$ to denote the set of points in $\mathcal{M}$ which maps to q via $\varphi_{p,\mathcal{M}}$.

We take $\delta = \delta(\mathcal{M}) \in \mathbb{R}_{\geq 0}$ to be any non-negative real number such that for any $p \in \mathcal{M}$, the following Condition (1) and Condition (2) are satisfied:

$$\delta < \min\{\text{reach}(\mathcal{M}), \text{reach}(\partial \mathcal{M})\}, \tag{1}$$

$$\varphi_{p,\mathcal{M}}|_{B_\delta(p) \cap \mathcal{M}} \text{ is a diffeomorphism onto its image.} \tag{2}$$

Finally, suppose $\mathcal{X} = \{x_1, ..., x_n\}$ is a set of sampled data points from a compact, differentiable manifold $\mathcal{M}$ with boundary. $\mathcal{X}$ is ϵ-*dense* if for any $p \in \mathcal{M}$, there is a point $x \in \mathcal{X}$ such that $x \in B_\epsilon(p)$. Let $U = \bigcup_{x \in \mathcal{X}} B_\epsilon(x)$ denote the offset of $\mathcal{X}$. We also define the canonical map $\pi: U \to \mathcal{M}$ by

$$\pi(y) := \arg \min_{p \in \mathcal{M}} \|y - p\|.$$

By Condition (1) and the definition of reach, π is well-defined when $\epsilon \leq \delta$.

2.2. Volume of a hyperspherical cap

Let S be a k-dimensional hypersphere of radius r. Let H be a hyperplane that divides S into two parts. We take the smaller part as a hyperspherical cap. Let a be the radius of the base of the cap, and $\phi := \arcsin(\frac{a}{r})$. Then by [22], the volume of the hypershperical cap is

$$V = \frac{1}{2} V_k(r) I_{\sin^2 \phi} \left(\frac{k+1}{2}, \frac{1}{2} \right),$$

where $\Gamma(x)$ is the Gamma function, $V_k(r)$ is the volume of the k-dimensional sphere with radius r and $I_x(a, b)$ is the regularized incomplete beta function.

3. The main results

The main theorem of this paper (Theorem 3.1) is centered around a probabilistic notion of sampling conditions for manifolds with boundary, which relates the topological equivalence between the offset (of samples) and the manifold as a deformation retract. To the best of our knowledge, such a result has not been addressed by existing theories.

Theorem 3.1. *Let $\mathcal{M} \subseteq \mathbb{R}^N$ be a compact differentiable k-dimensional manifold possibly with boundary. Let $\mathcal{X} = x_1, x_2, ..., x_n$ be drawn by sampling $\mathcal{M}$ in i.i.d fashion according to the uniform probability measure on $\mathcal{M}$. Let $\epsilon \in (0, \frac{1}{2}\delta(\mathcal{M}))$ and $U = \bigcup_{x \in \mathcal{X}} B_\epsilon(x)$. Then for all*

$$n > \beta(\epsilon)\left(\ln \beta\left(\frac{\epsilon}{2}\right) + \ln\left(\frac{1}{\gamma}\right)\right),$$

U deformation retracts to $\mathcal{M}$ with probability $> 1 - \gamma$. In particular, with such confidence, the homology of U is the same as that of $\mathcal{M}$. Here, $\beta(y) := \dfrac{\text{vol}(\mathcal{M})}{\frac{\cos^k \theta(y)}{2^{k+1}} I_{1 - \frac{y^2 \cos^2 \theta(y)}{16\delta^2}} \left(\frac{k+1}{2}, \frac{1}{2} \right) \text{vol}(B_y^k)}$ and $\theta(y) = \arcsin(\frac{y}{4\delta})$.

Theorem 3.1 is implied by combining Theorems 3.2 and 3.3 below.

Theorem 3.2. *Let $\mathcal{X}$ be any finite collection of points $x_1, ..., x_n \in \mathbb{R}^N$ such that it is $\frac{\epsilon}{2}$-dense in $\mathcal{M}$. Then for any $\epsilon < \frac{\delta}{2}$, we have that U deformation retracts to $\mathcal{M}$.*

Theorem 3.3. *Let $\mathcal{X}$ be drawn by sampling $\mathcal{M}$ in i.i.d fashion according to the uniform probability measure on $\mathcal{M}$. Then with probability $1 - \gamma$, we have that $\mathcal{X}$ is $\frac{\epsilon}{2}$-dense ($\epsilon < \frac{\delta}{2}$) in $\mathcal{M}$ provided*

$$|\mathcal{X}| \geq \beta(\epsilon)(\ln \beta(\frac{\epsilon}{2}) + \ln(\frac{1}{\gamma})).$$

As is mentioned in Section 1, compared to [19, Theorem 4.6] and [20, Theorem 13], Theorem 3.2 only deals with manifolds with boundary, but it establishes a criterion for deformation retract, which is stronger than homotopy equivalence as in [19]. The following is an example where neither [19, Theorem 4.6] or [20, Theorem 13] applies but Theorem 3.2 does.

Example 3.4. Let C be a semi-unit-circle. Then C is a manifold with boundary and its boundary consists of the two end points which are also the end points of a diameter. Let the data be the points $A_1, A_2, ..., A_8$ which divide the semi-circle evenly into 7 arcs, with A_1 and A_8 being the end points. Denote $A := \{A_1, ..., A_8\}$ and let $\epsilon = 0.48$.

We first treat this case with the result in [19]. We adopt the notations there. [19, Theorem 4.6] is the main reconstruction theorem, and it requires the inequality

$$\frac{4d_H(K, K')}{\mu^2} \leq \alpha < r_\mu(K') - 3d_H(K, K') \tag{3}$$

in order for the offset K^α to be homotopy equivalent to K'. Here we let $K' = C$ and $K = A$. It is easy to see that $d_H(K, K') = 2 \sin \frac{\pi}{28} \approx 0.223928$, and

$$r_\mu(K') = r_\mu(C) = \begin{cases} 0 & \mu > 1 \\ 1 & 0 < \mu \leq 1. \end{cases}$$

If $\mu > 1$, then as $r_\mu(K') = 0$, the right half of (3) does not make sense. If $0 < \mu \leq 1$, then

$$\frac{4d_H(K, K')}{\mu^2} \geq 4d_H(K, K') = 8 \sin \frac{\pi}{28} > 1 - 6 \sin \frac{\pi}{28} = r_\mu(K') - 3d_H(K, K').$$

This is a contradiction to (3). Therefore in this case, [19, Theorem 4.6] does not apply.

Next we try [20, Theorem 13 and 14]. [20, Theorem 13] requires that

$$d_H(A, C) \leq \epsilon < \lambda^{\text{cech}}(\mu) r_\mu(C),$$

where

$$\lambda^{\text{cech}}(\mu) = \frac{-3\mu + 3\mu^2 - 3 + \sqrt{-8\mu^2 + 4\mu^3 + 18\mu + 2\mu^4 + 9 + \mu^6 - 4\mu^5}}{-7\mu^2 + 22\mu + \mu^4 - 4\mu^3 + 1}$$

We have $d_H(K, K') = 0.223928$ as well as the value of $r_\mu(C)$ as is deduced above. On the other hand, from [20, Fig. 9] we see that $\lambda^{\text{cech}}(\mu)$ is increasing on (0,1], hence $\lambda^{\text{cech}}(\mu) \leq \lambda^{\text{cech}}(1) = \frac{-3 + \sqrt{22}}{13} \approx 0.130032$. So it is easy to see that $d_H(A, C) > \lambda^{\text{cech}}(\mu) r_\mu(C)$ for all $\mu \in \mathbb{R}_+$. Therefore [20, Theorem 13] does not apply. [20, Theorem 14] requires that

$$d_H(A, C) \leq \epsilon < \lambda_n^{\text{rips}}(\mu) r_\mu(C).$$

But by [20, Fig. 9], $\lambda_n^{\text{rips}}(\mu)$ is always smaller than $\lambda^{\text{cech}}(\mu)$, so [20, Theorem 14] does not apply either.

With some calculation, one can see that in order to apply [19, Theorem 4.6], at least 28 regularly sampled points along the semi-unit-circle are necessary; in order to apply [20, Theorem 13], at least 14 regularly sampled points are necessary. We also note that both [19] and [20] may handle some noisy sampling: they do not assume positive reach in general, and points can be sampled on or near the compact set.

We then try [21, Theorem 4.3]. Setting $K = C$ and $L = A$, using the notation of [21, Theorem 4.3], it requires condition (a): an inequality such that $(4 + \mu^2)\delta < \mu^2 r$, where $\delta > d_H(K, L) = d_H(A, C) = 2 \sin \frac{\pi}{28} \approx 0.223928$, r is the offset parameter of points in $L = A$, and $\mu \in (0, 1)$. [21, Theorem 4.3] also requires condition (b): $K_{[r-\delta, r-\delta+2\delta/\mu]}$ contains no μ-critical point. If both (a) and (b) are satisfied, then L_r deformation retracts to $K_{r-\delta}$. To match the results obtained by our Theorem 3.2, we would like $K_{r-\delta} = K_0 = K$; however, since $r > \delta$ due to condition (a), we would like $r - \delta$ to be minimized (such that it is as close as possible to 0). Set $\delta = 0.224$ (as close as possible to $d_H(A, C)$). To satisfy condition (a), we require $r > \frac{(4+\mu^2)\delta}{\mu^2} = \frac{4\delta}{\mu^2} + \delta > 5\delta = 1.12$. This means $r - \delta > 4\delta = 0.896$. Setting $r - \delta = 0.9$ ($r = 1.124$) and $\mu = 0.99$, the interval $[r-\delta, r-\delta+2\delta/\mu]$ however contains a μ-critical point (which is the midpoint connecting A_1 and A_8). This means condition (b) does not hold. In summary, the best possible result from [21, Theorem 4.3] applied to our example is that $L_{1.124}$ deformation retracts to $K_{0.9}$; however such a result does not hold because condition (b) is violated.

Finally we try to apply Theorem 3.2 to this case. It is easy to see that $\text{reach}(C) = \text{reach}(\partial C) = 1$, and it also satisfies Condition (2) to let $\delta(C) = 0.99$. Now recall that $\epsilon = 0.48 < \frac{\delta(C)}{2}$. Since $d_H(A, C) = 2 \sin \frac{\pi}{28} \approx 0.223928 < 0.24 = \frac{\epsilon}{2}$ as is calculated above, A is $\frac{\epsilon}{2}$-dense. Therefore Theorem 3.2 applies, which means that $U_{0.48}$ deformation retracts to C.
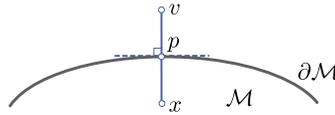
Fig. 1. The worse case for Lemma 4.3, where $||v - p|| = ||p - x|| = \frac{\epsilon}{2}$.

4. Proofs of the main results

To prove our main results, we begin with a series of lemmas. Recall the canonical map $\pi : U \to \mathcal{M}$ is defined by $\pi(x) := \arg\min_{p \in \mathcal{M}} \|x - p\|$. Lemmas 4.1-4.5 contribute to the proof of Theorem 3.2. Roughly speaking, they show that $\pi^{-1}(p)$ is star-shaped for every $p \in \mathcal{M}$, hence the deformation retract is well-defined. Lemmas 4.6-4.8 contribute to the proof of Theorem 3.3 by giving a lower bound to the volume of $\mathcal{M} \cap B_\epsilon(p)$ for every $p \in \mathcal{M}$ in terms of ϵ. Then, by directly applying results in [3], we get the estimation of the number of data points as in Theorem 3.3.

Lemma 4.1. *Choose $\delta = \delta(\mathcal{M}) > 0$ as in Section 2. Then for any $\lambda \in (0, \delta)$, any point q such that $\|q - \pi(q)\| < \lambda$ and any $q' \in |q\pi(q)|$, we have $\pi(q') = \pi(q)$.*

Proof. Suppose that this is not the case. Then

$$\|q - \pi(q)\| = \|q - q'\| + \|q' - \pi(q)\| > \|q - q'\| + \|q' - \pi(q')\| \geq \|q - \pi(q')\|.$$

This is a contradiction to Condition (1). ☐

Suppose a certain point $q \in U$ deformation retracts to a point $\pi(q) \in \mathcal{M}$. Let $|q\pi(q)|$ be the path of the deformation retract for q. Lemma 4.1 tells us that along the path $|q\pi(q)|$, all the points deformation retracts to $\pi(q)$.

From now on we set $\epsilon < \frac{\delta}{2}$. For a point $p \in \mathcal{M}$ and a point $p' \in \partial\mathcal{M}$ we define $\text{st}_\mathcal{M}(p)$ and $\text{st}_{\partial\mathcal{M}}(p')$ as

$$\text{st}_\mathcal{M}(p) := \bigcup_{x \in \mathcal{X}; x \in B_\epsilon(p)} (B_\epsilon(x) \cap T_p(\mathcal{M})^\perp).$$

$$\text{st}_{\partial\mathcal{M}}(p') := \bigcup_{x \in \mathcal{X}; x \in B_\epsilon(p')} (B_\epsilon(x) \cap T_{p'}(\partial\mathcal{M})^\perp).$$

For convenience, we also define $\text{st}_{\partial\mathcal{M}}(p') := \emptyset$ if $p' \in \mathcal{M}^\circ$. We present the next lemma, whose proof is exactly the same as those corresponding ones in [3], although the statements are somewhat different.

Lemma 4.2 ([3], Proposition 4.1)**.** $\text{st}_\mathcal{M}(p)$ *(resp. $\text{st}_{\partial\mathcal{M}}(p)$) is star-shaped for any $p \in \mathcal{M}$ (resp. $p \in \partial\mathcal{M}$).*

Lemma 4.3. $\pi^{-1}(p) \subseteq \text{st}_\mathcal{M}(p)$ *(resp. $\pi^{-1}(p) \subseteq \text{st}_{\partial\mathcal{M}}(p)$) for any $p \in \mathcal{M}$ (resp. $p \in \partial\mathcal{M}$).*

Proof. If $p \in \mathcal{M}^\circ$, this is already proven in [3, Proposition 4.2]. So for the rest of the proof we assume that $p \in \partial\mathcal{M}$. It is necessary to check that $\pi^{-1}(p) \subset T_p(\mathcal{M})^\perp$. This is true since the reach of $\mathcal{M}$ is assumed to be positive.

Let v be an arbitrary point in $\pi^{-1}(p)$. By the definition of $\text{st}_{\partial\mathcal{M}}$, we only need to consider the case where there is a point $q \in \mathcal{X}$ such that $q \notin B_\epsilon(p)$ and $v \in B_\epsilon(q)$. We also observe that by definition, our δ is less or equal to the τ in [3, Lemma 4.1], so all the conditions of [3] are satisfied. Therefore by [3], the distance between v and p is at most $\frac{\epsilon^2}{\delta}$. Now by the $\frac{\epsilon}{2}$-dense condition, there is a point $x \in \mathcal{X}$ such that $\|x - p\| \leq \frac{\epsilon}{2}$. Therefore

$$\|v - x\| \leq \|v - p\| + \|p - x\| \leq \frac{\epsilon^2}{\delta} + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \quad \square$$

Remark 4.4. We note that the worst case in Lemma 4.3 can happen when p lies on the boundary of $\mathcal{M}$. This is illustrated in Fig. 1, where M is a plane region with boundary, v is also inside the plane, and $|vp|$ and $|xv|$ are both perpendicular to $\partial\mathcal{M}$. This is why we require that $\epsilon < \frac{\delta}{2}$, which is more restrictive than the requirement $\epsilon < \sqrt{\frac{5}{3}}\delta$ in [3]. Note that this does not show that Theorem 3.2 is tight.

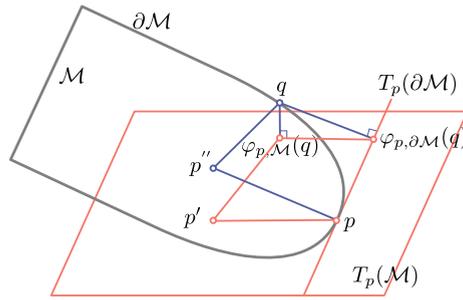**Lemma 4.5.** *For any $p \in \mathcal{M}$, $\pi^{-1}(p)$ is star-shaped with respect to p.*

Fig. 2. An illustration for Lemma 4.6 for the case when $q \in \partial\mathcal{M}$, where $\|p' - p\| = \delta$, $\|p'' - p\| = \delta$, $\|p'' - q\| > \delta$, $\|q - \varphi_{p,\partial\mathcal{M}}(q)\| \le \delta$.

Proof. If $p \in \mathcal{M}^\circ$, we have by Lemma 4.2 and 4.3 that $\pi^{-1}(p) = st_\mathcal{M}(p)$ which is star-shaped with respect to p. So for the rest of the proof we assume that $p \in \partial\mathcal{M}$. By Lemma 4.2, $st_{\partial\mathcal{M}}(p)$ is star-shaped; by Lemma 4.1, for any ray $l \subset T_p(\partial\mathcal{M})^\perp$ starting from p, either $\pi^{-1}(p) \cap l = \{p\}$ or $st_{\partial\mathcal{M}}(p) \cap l \subseteq \pi^{-1}(p) \cap l$; by Lemma 4.3 we know that $\pi^{-1}(p) \subseteq st_{\partial\mathcal{M}}(p)$. So if $q \in \pi^{-1}(p)$, then $|pq| \subseteq \pi^{-1}(p)$. □

Then Theorem 3.2 generalizes [3, Proposition 3.1] to compact manifolds with boundaries. Its proof is as follows.

Proof. We define the deformation retract $F(x, t) : U \times [0, 1] \to U$ as $F(x, t) = tx + (1 - t)\pi(x)$. By Lemma 4.1 and Lemma 4.5, this deformation retract is well-defined. Moreover since U is contained in $\mathcal{M}^\delta$, there is no critical point for distance functions, we get that U deformation retracts to $\mathcal{M}$. □

Lemma 4.6. *Let $p \in \partial\mathcal{M}$ and $A = \mathcal{M} \cap B_\epsilon(p)$, where $\epsilon \in (0, \delta)$. Then*

$$\text{vol}(A) > \frac{\cos^k \theta}{2} I_{1 - \frac{\epsilon^2 \cos^2 \theta}{4\delta^2}}\left(\frac{k+1}{2}, \frac{1}{2}\right)\text{vol}(B_\epsilon^k(p)),$$

where I is the regularized incomplete beta function, $B_\epsilon^k(p)$ is the k-dimensional ball in T_p centered at p, and $\theta = \arcsin(\frac{\epsilon}{2\delta})$.

Proof. Let $p' \in T_p(\mathcal{M})$ be the point such that $\overrightarrow{pp'}$ is perpendicular to $T_p(\partial\mathcal{M})$ and points to the inside of $\mathcal{M}$, and $\|p - p'\| = \delta$. We first want to show that

$$B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p') \subseteq \varphi_{p,\mathcal{M}}(A). \tag{4}$$

By Condition (2), $\varphi_{p,\mathcal{M}}|_A$ is a homeomorphism onto its image. In particular,

$$\varphi_{p,\mathcal{M}}(\partial A) = \partial(\varphi_{p,\mathcal{M}}(A)).$$

It is easy to see that $B_s^k(p) \cap B_\delta^k(p') \cap \varphi_{p,\mathcal{M}}(A) \ne \emptyset$ for any $s > 0$. Let $q \in \mathcal{M}$ be a point on the boundary of A. Then $q \in \partial\mathcal{M} \cup \partial B_\epsilon(p)$. We prove (4) by proving the *claim* that no matter whether $q \in \partial\mathcal{M}$ or $q \in \partial B_\epsilon(p)$, $\varphi_{p,\mathcal{M}}(q)$ is outside $B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p')$. Indeed, if there exists a point $q' \in B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p')$ such that $q' \notin \varphi_{p,\mathcal{M}}(A)$, we choose a point $o \in B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p') \cap \varphi_{p,\mathcal{M}}(A)$. Then the line segment connecting o and q' must intersect with the $\partial(\varphi_{p,\mathcal{M}}(A))$. But on the other hand, by the convexity of $B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p')$, any intersection point is inside $B_{\epsilon\cos\theta}^k(p) \cap B_\delta^k(p')$. This is a contradiction.

So far we have reduced (4) to a claim in the previous paragraph. To prove the claim, we need to consider two cases where $q \in \partial\mathcal{M}$ and $q \in \partial B_\epsilon(p)$ respectively. This is done in the next two paragraphs.

We first deal with the case where $q \in \partial\mathcal{M}$. Let p'' be the point where $\overrightarrow{pp''}$ is in the same direction with $\overrightarrow{\varphi_{p,\partial\mathcal{M}}(q)q}$ and $\|p'' - p\| = \delta$ (if $\varphi_{p,\partial\mathcal{M}}(q) = q$, then set $p'' = p'$). By Condition (1), q is outside $B_\delta(p'')$. Now q, $\varphi_{p,\mathcal{M}}(q)$ and $\varphi_{p,\partial\mathcal{M}}(q)$ form a right-angled triangle where $|q\varphi_{p,\partial\mathcal{M}}(q)|$ is the hypotenuse. Define q' to be the point on the half line $[\varphi_{p,\partial M}(q), q)$ at distance $\|\varphi_{p,\partial M}(q) - \varphi_{p,M}(q)\|$ from $\varphi_{p,\partial\mathcal{M}}(q)$. Since a rotation along an axis through p and $\varphi_{p,\partial\mathcal{M}}(q)$ that sends p' to p'' will send the segment $[p', \varphi_{p,M}(q)]$ to $[p'', q']$ and therefore $\|p' - \varphi_{p,M}(q)\| = \|p'' - q'\|$. We also have

$$\|q' - \varphi_{p,\partial M}(q)\| = \|\varphi_{p,\partial M}(q) - \varphi_{p,M}(q)\| \le \|q - \varphi_{p,\partial M}(q)\|$$
$$\le \|q - p\| = \epsilon < \delta = \|p'' - p\|.$$

These together with the fact that q is outside $B_\delta(p'')$ implies that $\varphi_{p,\mathcal{M}}(q)$ is outside $B_\delta^k(p')$. This case is illustrated in Fig. 2.

Next we consider the case where $q \in \partial B_\epsilon(p)$. This implies that $\|q - p\| = \epsilon$. We have that $\|\varphi_{p,\mathcal{M}}(q) - p\| = \epsilon\cos(\phi)$ where ϕ is the (smallest nonnegative) angle between $\overrightarrow{pq}$ and $\overrightarrow{p\varphi_{p,\mathcal{M}}(q)}$. Let p'' be the point where $\overrightarrow{pp''}$ is in the same
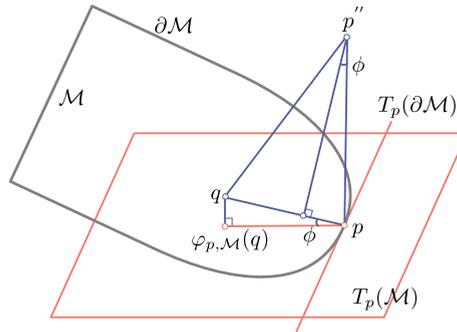
Fig. 3. An illustration for Lemma 4.6 for the case when $q \in \partial B_\epsilon(p)$, where $\|p - q\| = \epsilon$, $\phi < \theta$, $|p'' - p| = \delta$ and $\|p'' - q\| > \delta$.

direction with $\overrightarrow{\varphi_{p,\mathcal{M}}(q)q}$ (if $\varphi_{p,\mathcal{M}}(q) = q$, then choose an arbitrary p'' such that $pp'' \perp T_p(\mathcal{M})$) and $\|p'' - p\| = \delta$. Then by Condition (1), $\|p'' - q\| > \delta$. So by the definition of θ we see that $\phi < \theta$, and $\|\varphi_{p,\mathcal{M}}(q) - p\| = \epsilon \cos(\phi) > \epsilon \cos(\theta)$. Hence $\varphi_{p,\mathcal{M}}(q)$ is outside $B^k_{\epsilon \cos \theta}(p)$. This is illustrated in Fig. 3.

So we have proved (4) by proving the claim. (4) implies that $\text{vol}(A) \geq \text{vol}(B^k_{\epsilon \cos \theta}(p) \cap B^k_\delta(p'))$, the right-hand side of which consists of two hyperspherical caps. For convenience we choose the lower bound of the right side to be the hyperspherical cap that belongs to $B^k_{\epsilon \cos \theta}(p)$. By Section 2.2, the volume of this hyperspherical cap is $\frac{1}{2} I_{1 - \frac{\epsilon^2 \cos^2 \theta}{4\delta^2}}(\frac{k+1}{2}, \frac{1}{2}) \text{vol}(B^k_{\epsilon \cos \theta}(p))$. Moreover we know that $\text{vol}(B^k_{\epsilon \cos \theta}(p)) = \cos^k \theta \, \text{vol}(B^k_\epsilon(p))$, so we are done. □

Using the same argument as in the third paragraph of the proof of the last lemma, we actually have

Lemma 4.7. *Let $p \in \mathcal{M}$ and $\epsilon > 0$ such that $\partial \mathcal{M} \cap B_\epsilon(p) = \emptyset$. Let $A := \mathcal{M} \cap B_\epsilon(p)$. Then*

$$\text{vol}(A) \geq \text{vol}(B^k_{\epsilon \cos \theta}(p)) = \cos^k \theta \, \text{vol}(B^k_\epsilon(p)).$$

Combining Lemma 4.6 and 4.7, we have

Lemma 4.8. *Let $p \in \mathcal{M}$ and $A = \mathcal{M} \cap B_\epsilon(p)$, where $\epsilon \in (0, \delta)$. Then*

$$\text{vol}(A) \geq \frac{\cos^k \theta'}{2^{k+1}} I_{1 - \frac{\epsilon^2 \cos^2 \theta'}{16\delta^2}}(\frac{k+1}{2}, \frac{1}{2}) \text{vol}(B^k_\epsilon(p)),$$

where $\theta' := \arcsin(\frac{\epsilon}{4\delta})$.

Proof. If $d(p, \partial \mathcal{M}) > \frac{\epsilon}{2}$, then $\partial \mathcal{M} \cap B_{\frac{\epsilon}{2}}(p) = \emptyset$. So by Lemma 4.7,

$$\text{vol}(A) \geq \cos^k \theta' \text{vol}(B^k_{\frac{\epsilon}{2}}(p))$$

and we are done. If $d(p, \partial \mathcal{M}) \leq \frac{\epsilon}{2}$, let p' be a point on $\partial \mathcal{M}$ that has minimum distance from p. Then $B_{\frac{\epsilon}{2}}(p') \subset B_\epsilon(p)$. So

$$\text{vol}(A) \geq \text{vol}(B_{\frac{\epsilon}{2}}(p') \cap \mathcal{M}) \geq \frac{\cos^k \theta'}{2} I_{1 - \frac{\epsilon^2 \cos^2 \theta'}{16\delta^2}}(\frac{k+1}{2}, \frac{1}{2}) \text{vol}(B^k_{\frac{\epsilon}{2}}(p))$$

$$= \frac{\cos^k \theta'}{2^{k+1}} I_{1 - \frac{\epsilon^2 \cos^2 \theta'}{16\delta^2}}(\frac{k+1}{2}, \frac{1}{2}) \text{vol}(B^k_\epsilon(p)),$$

where the last inequality is by Lemma 4.6. □

We observe that the right side of the inequality in Lemma 4.8 is $\frac{\text{vol}(\mathcal{M})}{\beta(\epsilon)}$, where the function β is defined in Theorem 3.1 (note that the θ' here corresponds to the θ in Theorem 3.1). By [3, Lemma 5.1 and 5.2], a satisfactory number of points $|\mathcal{X}|$ as in Theorem 3.3 is of the form $\frac{1}{\alpha}(\ln l + \ln \frac{1}{\gamma})$, where α is any lower bound of $\frac{\text{vol}(A)}{\text{vol}(\mathcal{M})}$ and l is any upper bound of $\frac{\epsilon}{2}$-packing-number. So by Lemma 4.8, it suffices to take α and l to be $\frac{1}{\beta(\epsilon)}$ and $\beta(\frac{\epsilon}{2})$, respectively. Therefore we obtain Theorem 3.3. Finally, combining Theorems 3.2 and 3.3, we arrive at Theorem 3.1.
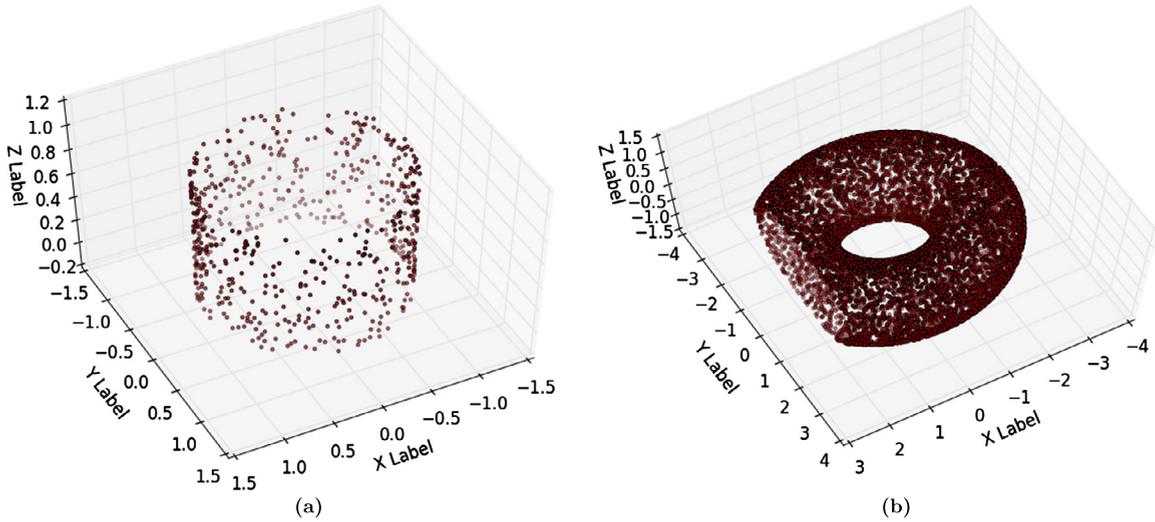
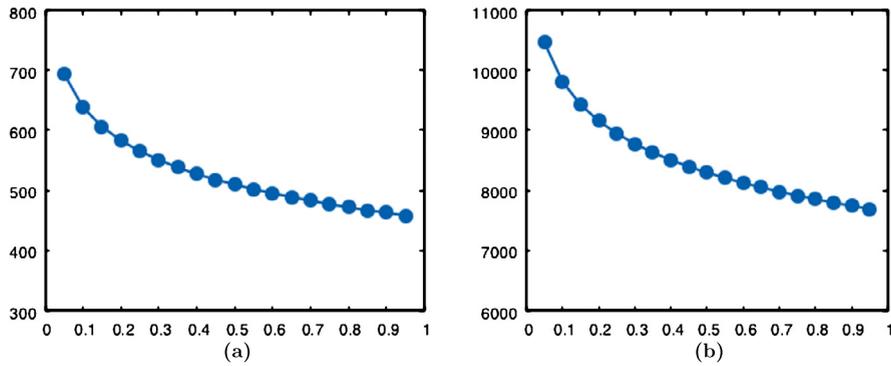**Fig. 4.** Point cloud samples for *cylinder* (**a**) and *torus* (**b**).

Fig. 5. Lower bounds for *cylinder* (**a**) and *torus* (**b**) for a fixed $\epsilon = 0.49$, x-axis corresponds to γ while y-axis corresponds to n^*.

5. Experimental illustration

In this section, we work on two typical examples of manifolds with boundary. The first example is a cylindrical surface, referred to as the *cylinder* dataset, which has radius 1 and height 1. More precisely, it can be expressed as

$$\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1, z \in [0, 1]\}.$$

The second example is a torus with a cap chopped off, referred to as the *torus* dataset. In $\mathbb{R}^3$, it can be expressed as the torus with inner circle $x^2 + y^2 = 1$ and the outer circle $x^2 + y^2 = 9$, and the part with $x \geq 2$ is chopped off.

Sampling parameters. As stated in the main Theorem 3.1, the lower bound of sampling that guarantees deformation retraction with probability $1 - \gamma$ can be expressed as

$$n^* = \beta(\epsilon)(\ln \beta(\frac{\epsilon}{2}) + \ln(\frac{1}{\gamma})),$$

where $\beta(x) := \dfrac{\mathrm{vol}(\mathcal{M})}{\frac{\cos^k \theta}{2^{k+1}} I_{1 - \frac{x^2 \cos^2 \theta}{16\delta^2}}(\frac{k+1}{2}, \frac{1}{2}) \mathrm{vol}(B_x^k)}$ and $\theta = \arcsin(\frac{x}{4\delta})$.

For *cylinder*, $\mathrm{vol}(\mathcal{M}) = 2\pi$, $k = 2$, $\mathrm{vol}(B_x^2) = \pi x^2$, $\delta = 1$. For instance, setting $\epsilon = 0.49$, $\gamma = 0.1$ gives rise to $n^* = 638$, as illustrated in Fig. 4 (a).

For *torus*, $\mathrm{vol}(\mathcal{M}) = (8 - 0.522) \cdot 2\pi$, $k = 2$, $\mathrm{vol}(B_x^2) = \pi x^2$, $\delta = 1$. For instance, setting $\epsilon = 0.49$, $\gamma = 0.1$ gives rise to $n^* = 9809$, as illustrated in Fig. 4 (b).

Distribution of lower bounds. For a fixed sample quality ϵ, we demonstrate the distribution of lower bounds n^* as γ increases from 0.05 to 0.95 (that is, confidence ranges from 95% to 5%). This is shown in Fig. 5. Intuitively, for a fixed sample quality, we need more point samples in order to obtain higher confidence in topological inference.
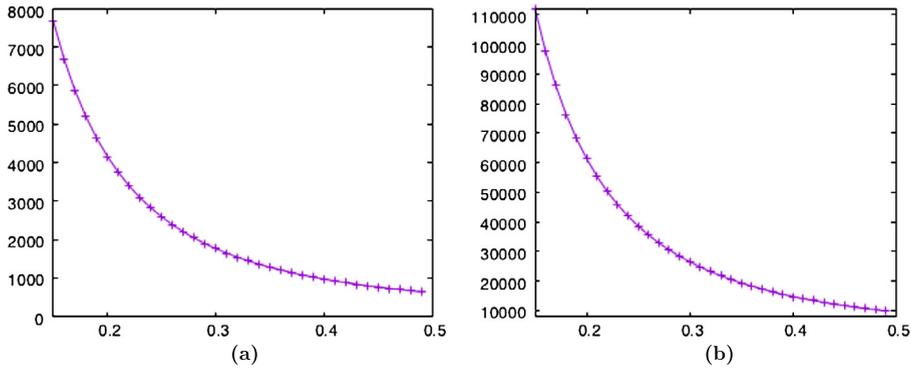
Fig. 6. Lower bounds for *cylinder* (**a**) and *torus* (**b**) for a fixed $\gamma = 0.1$, x-axis corresponds to the ϵ while y-axis corresponds to n^*.
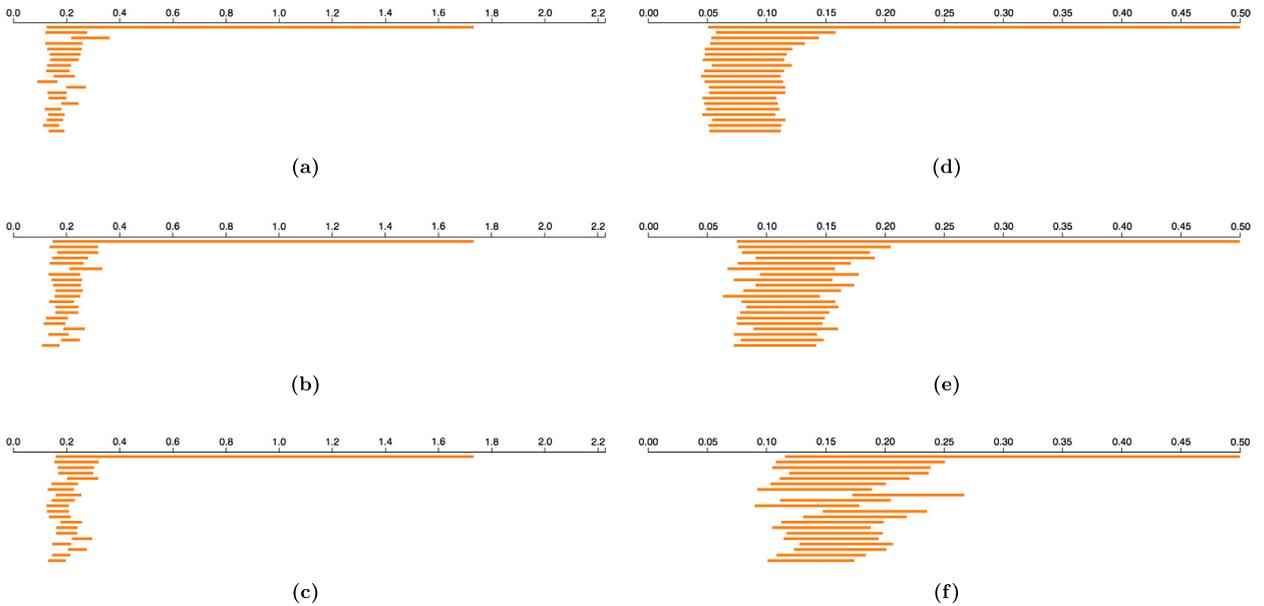

Fig. 7. Persistent barcodes for *cylinder*. Each plot only shows the top 20 most persistent (longest) cycles. For a fixed $\epsilon = 0.49$, $r \in [0, 2.2]$: (**a**) $\gamma = 0.1$, $n^* = 638$; (**b**) $\gamma = 0.2$, $n^* = 583$; and (**c**) $\gamma = 0.3$, $n^* = 551$. For a fixed $\gamma = 0.1$ (90% confidence), $r \in [0, 0.5]$: (**d**) $\epsilon = 0.2$, $n^* = 4160$; (**e**) $\epsilon = 0.3$, $n^* = 1763$; and (**f**) $\epsilon = 0.4$, $n^* = 967$. Notice different scales between the plots on the left and on the right.

Meanwhile, for a fixed $\gamma = 0.1$, which corresponds to a confidence of 90%, we illustrate the distribution of lower bounds n^* as ϵ increases from 0.15 to 0.5. By Theorem 3.1, it is rather obvious that we need more points to have higher quality samples for a fixed confidence level. This is shown in Fig. 6.

Homology computation. Finally we can perform homology computation on the above point clouds; in particular, for a given sample $\mathcal{X}$ and its corresponding U, we show that the homology of U equals the homology of $\mathcal{M}$. Admittedly, homology is a very weak verification of our main sampling theorem. In fact, if one's goal is only to recover the same homology of a manifold with point samples, our estimation from Theorem 3.1 is an obvious overestimation. In other words, our estimation of the lower bound n^* has to account for the boundary condition and to guarantee deformation retract (not just homology or homotopy equivalence).

Nevertheless, we show the results of homology inference across multiple γ with a fixed ϵ, as well as the results across multiple ϵ with a fixed γ. We rely on the computation of persistent homology to recover the homological information of a point cloud sample. Persistent homology, roughly speaking, operates on a point cloud sample $\mathcal{X}$ and tracks how the homology of $U(r) = \bigcup_{x \in \mathcal{X}} B_r(x)$ changes as r increases (where typically $r \in [r_0 = 0, r_k]$, for some positive real value r_k). Specifically, it applies the homology functor H to a sequence of topological spaces connected by inclusions,

$$U(r_0) \to \cdots \to U(r_i) \to U(r_{i+1}) \to \cdots \to U(r_k),$$

and studies a multi-scale notion of homology,

$$\mathsf{H}(U(r_0)) \to \cdots \to \mathsf{H}(U(r_i)) \to \mathsf{H}(U(r_{i+1})) \to \cdots \to \mathsf{H}(U(r_k)),$$
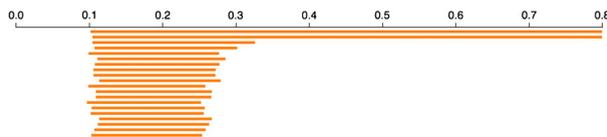
Fig. 8. Persistent barcodes for *torus*, for $\epsilon = 0.49$, $\gamma = 0.2$, $n^* = 9157$, $r \in [0, 0.8]$. Only the top 20 most persistent (longest) cycles are shown.

see [15,23,24] for introduction to persistent homology. We use the software package Ripser [25] for the computation of persistent homology. Given a point cloud sample $\mathcal{X}$, Ripser computes its persistent homology using Vietoris–Rips complexes formed on $\mathcal{X}$ and encodes the homological information using persistence barcodes. In a nutshell, each bar in the persistence barcodes captures the time when a homology class appears and disappears as r increases. As the homology of a union of balls is guaranteed (by the Nerve Lemma) to be the one of a Čech complex, the results of [20] could be utilized to deduce results on a Vietoris–Rips complex from a Čech complex. Note that using the Vietoris–Rips complexes to approximate Čech complexes adds an additional approximation factor [26, Theorem 2.5]. This approximation factor, however, is not an issue for the results described here.

For *cylinder* dataset, the 1-dimensional homology of its underlying manifold should be of rank one, as the dataset contains one significant cycle (tunnel). For a fixed $\epsilon = 0.49$, we compute the 1-dimensional persistent homology of the point clouds at parameter $\gamma = 0.1, 0.2, 0.3$ respectively. Their persistent barcodes are shown in Fig. 7(a)-(c) respectively. For each plot, the longest bar corresponds to the most significant 1-dimensional cycle, which clearly corresponds to the true homological feature of the underlying manifold.

Meanwhile, the 1-dimensional homology of the manifold underlying the *torus* dataset (with a cap chopped off) should be of rank two, as the dataset contains two significant cycles (same as the classic torus dataset). We have similar results as in the case of *cylinder* dataset. For simplicity, we give the persistent barcodes for $\epsilon = 0.49$, $\gamma = 0.2$, $n^* = 9157$ in Fig. 8. Here, the first two longest bars correspond to the two most significant 1-dimensional cycles, which again clearly correspond to the true homological features of the underlying manifold.

6. Discussions

Given a point cloud sample of a compact, differentiable manifold with boundary, we give a probabilistic notion of sampling condition that is not handled by existing theories. Our main results relate topological equivalence between the offset and the manifold as a deformation retract, which is stronger than homological or homotopy equivalence. Many interesting questions remain.

First, while our sampling condition considers differentiable manifolds with boundary, it cannot be trivially extended to handle manifolds with corners. The fundamental difficulty arises because the reach becomes zero in the case of manifolds with corners. We suspect that deriving practical sampling conditions for manifolds with corners, and in general, for stratified spaces, is challenging and requires new ways of thinking.

Second, we have conducted experiments that verify homological equivalence between the offset of samples and the underlying manifold. However, such an experiment is a very weak verification of our main inference theorem. Experimentally computing or verifying deformation retract in the point cloud setting (as stated in Theorem 3.1), possibly via the study of discrete gradient fields, remains an open question.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Amenta, M. Bern, Surface reconstruction by Voronoi filtering, Discrete Comput. Geom. 22 (4) (1999) 481–504.
[2] S.-W. Cheng, T.K. Dey, E.A. Ramos, Manifold reconstruction from point samples, in: Proc. Annu. ACM-SIAM Symp. Discrete Algorithms, 2005, pp. 1018–1027.
[3] P. Niyogi, S. Smale, S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, Discrete Comput. Geom. 39 (1) (2008) 419–441.
[4] M. Goresky, R. MacPherson, Stratified Morse Theory, Springer-Verlag, 1988.
[5] S. Weinberger, The Topological Classification of Stratified Spaces, University of Chicago Press, Chicago, IL, 1994.
[6] G. Haro, G. Randall, G. Sapiro, Stratification learning: detecting mixed density and dimensionality in high dimensional point clouds, in: Advances in Neural Information Processing Systems (NIPS), vol. 17.
[7] G. Lerman, T. Zhang, Probabilistic recovery of multiple subspaces in point clouds by geometric lp minimization, Ann. Stat. 39 (5) (2010) 2686–2715.
[8] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1945–1959.
[9] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, D. Morozov, Inferring local homology from sampled stratified spaces, in: IEEE Symposium on Foundations of Computer Science, 2007, pp. 536–546.
[10] P. Bendich, B. Wang, S. Mukherjee, Local homology transfer and stratification learning, in: ACM-SIAM Symposium on Discrete Algorithms, 2012, pp. 1355–1370.
[11] P. Skraba, B. Wang, Approximating local homology from samples, in: ACM-SIAM Symposium on Discrete Algorithms (SODA), 2014, pp. 174–192.

[12] M. Belkin, Q. Que, Y. Wang, X. Zhou, Graph Laplacians on singular manifolds: toward understanding complex spaces: graph Laplacians on manifolds with singularities and boundaries, in: Conference on Learning Theory.

[13] V. Nanda, Local cohomology and stratification, arXiv:1707.00354, 2017.

[14] A. Brown, B. Wang, Sheaf-theoretic stratification learning, in: International Symposium on Computational Geometry.

[15] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, American Mathematical Society, 2010.

[16] P. Bendich, Analyzing stratified spaces using persistent versions of intersection and local homology, PhD thesis, Duke University, 2008.

[17] P. Bendich, J. Harer, Persistent intersection homology, Found. Comput. Math. 11 (2011) 305–336.

[18] M. Goresky, R. MacPherson, Intersection homology I, Topology 19 (1982) 135–162.

[19] F. Chazal, D. Cohen-Steiner, A. Lieutier, A sampling theory for compact sets in Euclidean space, Discrete Comput. Geom. 41 (3) (2009) 461–479.

[20] D. Attali, A. Lieutier, D. Salinas, Vietoris-Rips complexes also provide topologically correct reconstructions of sampled shapes, Comput. Geom. 46 (4) (2013) 448–465.

[21] K. Turner, Cone fields and topological sampling in manifolds with bounded curvature, Found. Comput. Math. 13 (6) (2013) 913–933.

[22] S. Li, Concise formulas for the area and volume of a hyperspherical cap, Asian J. Math. Stat. 4 (1) (2011) 66–70.

[23] H. Edelsbrunner, J. Harer, Persistent homology – a survey, Contemp. Math. 453 (2008) 257–282.

[24] R. Ghrist, Barcodes: the persistent topology of data, Bull. Am. Math. Soc. 45 (1) (2008) 61–75.

[25] U. Bauer, Ripser, https://github.com/Ripser/ripser, 2016.

[26] V. de Silva, R. Ghrist, Coverage in sensor networks via persistent homology, Algebraic Geom. Topol. 7 (2007) 339–358.