Parameter-free Topology Inference and Sparsification for Data on Manifolds

Tamal K. Dey* Zhe Dong* Yusu Wang*

Abstract

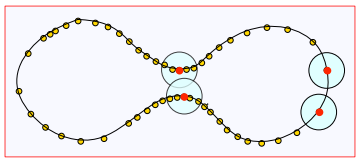
In topology inference from data, current approaches face two major problems. One concerns the selection of a correct parameter to build an appropriate complex on top of the data points; the other is the typical large size of this complex. We address these two issues in the context of inferring homology from sample points of a smooth manifold of known dimension sitting in a Euclidean space $\mathbb{R}^k$. We show that, for a sample of n points, we can identify a set of $O(n)$ points (as opposed to $O(n^{\lceil \frac{k}{2} \rceil})$ Voronoi vertices) approximating a subset of the medial axis that suffices to compute a distance sandwiched between the well known *local feature size* and the local *weak feature size*. This distance, called the *lean feature size*, helps to prune the input set at least to the level of local feature size while making the data locally uniform. The local uniformity in turn helps to build a complex of linear size for homology inference on top of the sparsified data without requiring any user-supplied distance threshold. Unlike most topology inference results, ours does not require that the input is dense relative to a *global* feature such as *reach* or *weak feature size*; instead it can be adaptive with respect to the local feature size. We present some empirical evidence in support of our theoretical claims.

1 Introduction

In recent years, considerable progress has been made in analyzing data for inferring the topology of a space from which the data is sampled. Often this process involves building a complex on top of the data points, and then analyzing the complex using various mathematical and computational tools developed in computational topology. There are two main issues that need attention to make this approach viable in practice. The first one stems from the requirement of choosing appropriate parameters to build the complexes so that the provable guarantees align with the computations. The other one arises from the unmanageable size of the complex—a problem compounded by the fact that the input can be large and usual complexes such as Vietoris-Rips built on top of it can be huge (exponential in its dimension).

In this paper, we address both of the above two issues with a technique for data sparsification. The data points are assumed to be sampled from a smooth manifold of known dimension sitting in some Euclidean space. We sparsify the data so that the resulting set is locally uniform and is still good for homology inference. Observe that, with a sample whose density varies with respect to a local feature size (such as the lfs proposed for surface reconstruction [2, 3]), no global parameter for building an appropriate complex can be found. The figure in the next paragraph illustrates this difficulty.

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



For the non-uniform sample shown in the figure, there is no single radius that can be chosen to construct, for example, Rips or Čech complexes. To connect points in the sparsely sampled part on right, the radius needs to be bigger than the feature size at the small neck in the middle. If chosen, this radius destroys the neck in the middle thus creating spurious topology. Our solution to this problem is a sparsification strategy so that the sample becomes locally uniform [13, 17] while guaranteeing that no topological information is lost. The sparsification is carried out without requiring any extra parameter and the resulting local uniformity eventually helps to construct the appropriate complex on top of the sparsified set.

The sparsification also addresses the problem of 'size' because it produces a sub-sample of the original input. The technique of subsampling has been suggested in some recent work. The witness complex builds on the idea of subsampling and subsequently constructing a complex using nearest neighbors [23]. Unfortunately, guarantees about topological inference cannot be achieved with witness complexes unless some non-trivial modifications are made and parameters are tuned. Sparsified Rips complexes proposed by Sheehy [22] also uses subsampling to summarize the topological information contained in a Rips filtration (a nested sequence). The graph induced complex proposed in [15] alleviates the 'size' problem even further by replacing the Rips complexes with a more sparsified complex. Both approaches, however, only approximate the true persistence diagram and hence to infer homology exactly require a user-supplied parameter to find the 'sweet spot' in the filtration range. Furthermore, none of these sparsifications are designed to work with a non-uniform input that is adaptive to a *local* as opposed to a *global* feature size.

Our algorithm computes a *lean set* of only $O(n^2)$ points using which the distance to a certain subset of the medial axis (similar to λ-medial axis) can be approximated. This approximation differs from polar approximation [4, 16] or λ-medial axis [10], which can also be used in principle for our purpose. But, these existing methods compute Voronoi diagrams which, for n points in $\mathbb{R}^k$, require $\Omega(n^{\lceil \frac{k}{2} \rceil})$ Voronoi vertices in the worst-case. In contrast, our lean set does not directly approximate the medial axis or its variants, but rather, the distance to it. The size of this lean set can be further brought down to $O(n)$ as shown in Section 2.3. More specifically, the distance to this lean set, which we call the *lean feature size*, is shown to lie between the local feature size lfs and the weak local feature size wlfs. Sparsifying the input with respect to this lean feature size allows the data to be decimated at least to the level of lfs, but at the same time keeps it dense enough with respect to the weak local feature size, which eventually leads to topological fidelity. This roughly means that the data is sparsified adaptively as much as possible without sacrificing the topological information (see experimental results in Figure 1).

The sparsified points are connected in a Rips-like complex using the lean feature size computed for each sample point. Following the approach in [12], the guarantee for topological fidelity is obtained by interleaving the union of a set of balls with the offsets of the manifold. To account for the adaptivity of the sample density, the distance for offsets are scaled appropriately (similar to a scaling in [14]) by the lean feature size, and the approach in [12] is adapted to this framework. To the best of our knowledge, this is the first sparsification strategy that handles adaptive input samples, produces an adaptive as well as a locally uniform sparsified sample, and infers homology without requiring a threshold parameter.

2 Sparsification

Let X be a smooth compact manifold embedded in a k-dimensional ambient Euclidean space $\mathbb{R}^k$. Our goal is to sparsify a dense and possibly adaptive sample of X and still be able to recover homology of X from it.

Distance function, feature size, and sample density. Let $d(x, A)$ denote the distance between a point $x \in \mathbb{R}^k$ and its closest point in a compact set $A \subset \mathbb{R}^k$. Consider the *distance function* $d_{\mathsf{X}} : \mathbb{R}^k \to \mathbb{R}$

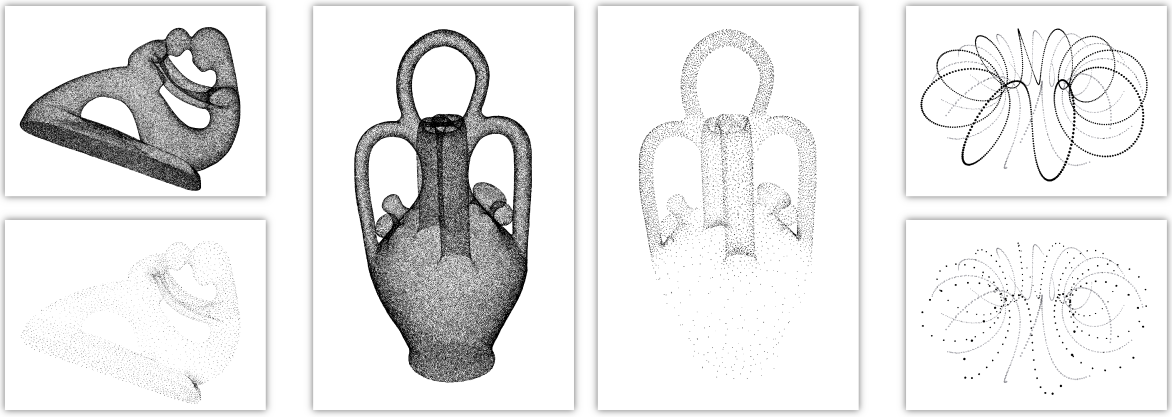


Figure 1: Sparsification: Original samples of 126500 and 101529 points of MOTHERCHILD and BOTIJO are decimated to 6016 and 8622 points respectively. Betti numbers are computed correctly by our algorithm (Section 4). The rightmost picture shows a 3D curve sample (top) and the *lean set* (bottom) approximating a relevant subset of the medial axis which otherwise spans a much larger subspace of $\mathbb{R}^3$.

defined as $d_{\mathsf{X}}(x) = d(x, \mathsf{X})$. Let $\Pi(x) = \{y \in \mathsf{X} \mid d(x, y) = d_{\mathsf{X}}(x)\}$ be the set of closest points of $x \in \mathbb{R}^k$ in X. Notice that, for any $y \in \Pi(x)$, the segment xy is contained in the normal space $\mathbf{N}_x$ of X at x. The *medial axis* M of X is the closure of the set of points with at least two closest points in X, and thus $M := Closure\{m \in \mathbb{R}^k \mid |\Pi(m)| \geq 2\}$.

The *local feature size* at a point $x \in \mathsf{X}$, denoted by $\mathrm{lfs}(x)$, is defined as the smallest distance between x and the medial axis M; that is, $\mathrm{lfs}(x) = d(x, M)$ [2]. There is another feature size definition that is particularly useful for inferring homological information [11]. This feature size is defined as the distance to the critical points of the distance function d_{X}, which is not differentiable everywhere. However, one can still define the following vector which extends the concept of gradient to d_{X} [20]. Specifically, given any point $x \in \mathbb{R}^k \setminus \mathsf{X}$, let $c(x)$ be the center of the unique minimal ball B_x enclosing $\Pi(x)$. Define the *gradient vector* at x: $\nabla_d(x) = \frac{x - c(x)}{d(x, \mathsf{X})}$ and the critical points $C := \{x \in \mathbb{R}^k \,|\, \nabla_d(x) = 0\}$. The *weak local feature size* at a point $x \in \mathsf{X}$, denoted by $\mathrm{wlfs}(x)$, is defined as $\mathrm{wlfs}(x) = d(x, C)$. Given an ε-dense sample w.r.t. the lfs which is known as the ε-sample in the literature [16], we would like to sparsify it to a locally uniform sample w.r.t. some function, ideally lfs, or wlfs. This motivates the following definition.

Definition 2.1 *A discrete sample $P \subset \mathsf{X}$ is called c-dense w.r.t. a function $\phi : \mathsf{X} \to \mathbb{R}$ if $\forall x \in \mathsf{X}$, $d(x, P) \leq c \cdot \phi(x)$. It is c-sparse if each pair of distinct points $p, q \in P$ satisfies $d(p, q) \geq c \cdot \phi(p)$. The sample P is called (c_1, c_2)-uniform w.r.t. ϕ if it is c_1-dense and c_2-sparse w.r.t. ϕ.*

To produce a (c_1, c_2)-uniform sample w.r.t. lfs or wlfs one needs to compute lfs or wlfs or their approximations. This in turn needs the computation of at least a subset of the medial axis or its approximation. One option is to approximate this set using the Voronoi poles as in [2, 4]. This proposition faces two difficulties. First of all, it needs computing the Voronoi diagram in high dimensions. Second, approximating the medial axis may require a large number of samples when a manifold of a low co-dimension is embedded in a high dimensional Euclidean space. To overcome this difficulty we propose to compute a discrete set L near M of small cardinality which helps to estimate the distance to a subset of M (See the curve sample in Figure 1 for an example). The set L called the *lean set* allows us to define an easily computable feature size which we call *lean feature size*. We show that this feature size is sandwiched between the lfs and wlfs thereby enabling us to sparsify an arbitrarily dense sample to a (c_1, c_2)-uniform sample w.r.t. a function bracketed

by lfs and wlfs. The constants c_1, c_2 are universal which ultimately leads to a parameter-free inference of the homology.

From now on, we assume that the input P is a dense sample of X in the following adaptive sense [2]. Each point is also equipped with a normal information as stated in Assumption 2.2. We will see later how this normal information can be computed.

Assumption 2.2 *The input point set P is ε-dense w.r.t.* lfs *function on a compact smooth manifold $\mathsf{X} \subset \mathbb{R}^k$ of known dimension without boundary. Also, every point $p \in P$ has an estimated normal space $\tilde{\mathbf{N}}_p$ where $\angle(\tilde{\mathbf{N}}_p, \mathbf{N}_p) \leq \nu_\varepsilon = O(\varepsilon)$ [1] (see Section 2.2 for computations of $\tilde{\mathbf{N}}_p$).*

Notice that while we assume the input to be ε-dense w.r.t. lfs, we do not need to know lfs and, locally, the sample can be much denser and non-uniform. Now we define the *lean set* with respect to which we define the lean feature size.

2.1 Lean set

Definition 2.3 *A pair $(p, q) \in P \times P$ is* β-good *for $0 < \beta < \frac{\pi}{2}$ if the following two conditions hold:*

1. $\max\{\angle(\tilde{\mathbf{N}}_p, pq), \angle(\tilde{\mathbf{N}}_q, pq)\} \leq \frac{\pi}{2} - \beta$.
2. *Let $v = \frac{p+q}{2}$ be the midpoint of pq. The ball $B(v, c_\beta d(p, q))$ does not contain any point of P where $c_\beta = \frac{1}{3}\tan\frac{\beta}{2}$.*

Definition 2.4 *The* β-lean set *L_β is defined as:*

$$L_\beta = \{v |\ v = \tfrac{p+q}{2} \text{ is the mid point of } pq \text{ where } (p, q) \text{ is a } \beta\text{-good pair}\}.$$

The β-lean feature size is defined as $\mathrm{lnfs}_\beta(x) = d(x, L_\beta)$.

One of our main results is the following property of the lean feature size (recall the definition of ν_ε in Assumption 2.2).

Theorem 2.5 *Let θ, β be two positive constants so that $\frac{\pi}{4} \geq \theta \geq \beta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon$ for a sufficiently small $\varepsilon \leq \frac{1}{8}\sin^2\theta$. Then,*

1. $\mathrm{lnfs}_\beta(x) \leq c_1 \cdot \mathrm{wlfs}(x)$ *for any point x in X,*
2. $\mathrm{lnfs}_\beta(p) \geq c_2 \cdot \mathrm{lfs}(p)$ *for every point $p \in P$*

where $c_1 = 1 + \cos\theta + \varepsilon$, $c_2 = \frac{2c_0c_\beta}{1+c_0+2c_0c_\beta}$, and $c_0 = \sin(\beta - \nu_\varepsilon)$, $c_\beta = \frac{1}{3}\tan\frac{\beta}{2}$ are positive constants.

We prove this theorem in the remainder of this section. The upper bound follows from Proposition 2.7 which shows a stronger result that lnfs_β is bounded from above by the distance to a subset of the medial axis characterized by an angle condition. This set also contains all critical points of the distance function d_{X}. First, we establish this result.

Definition 2.6 *The* θ-medial axis *$M_\theta \subseteq M$ of X is defined as the set of points $m \in M$ where there exist two points $x, y \in \Pi(m)$ such that $\angle xmy \geq 2\theta$.*

[1] We note that $\tilde{\mathbf{N}}_p$ and $\mathbf{N}_p$ here are subspaces of $\mathbb{R}^k$. The angle between them refers to the smallest non-zero *principle angle* between these two subspaces as used in the literature.

We will see later that the concept of θ-medial axis is also used as a bridge between geometry and topology for our inference result. Our algorithm *does not* approximate M_θ, but rather, approximates the distances to it by the the *lean set*.

Proposition 2.7 *Let θ, β be two positive constants so that $\frac{\pi}{2} \geq \theta \geq \beta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon$ for a sufficiently small $\varepsilon \leq \frac{1}{8}\sin^2\theta$. Let x be any point in X. Then, $\mathrm{lnfs}_\beta(x) = d(x, L_\beta) \leq c \cdot d(x, M_\theta)$ where $c = 1 + \cos\theta + \varepsilon$ is a positive constant.*

Proof: Let $m = \operatorname{argmin} d(x, M_\theta)$. By definition, we have a pair of points $s, t \in \Pi(m)$ in the manifold X so that the line segments sm and tm subtends an angle larger than or equal to 2θ and both sm and tm are normal to X at s and t respectively. Let $p \in P$ and $q \in P$ be the nearest sample points to s and t respectively. By the ε-sampling condition of P, we have that $d(p, s) \leq \varepsilon\mathrm{lfs}(s)$ and thus $\angle(\mathbf{N}_s, \mathbf{N}_p) \leq \varepsilon$.

In Appendix A, we show that the pair (p, q) is β-good, hence its midpoint $\frac{p+q}{2}$ belongs to L_β. Notice that $\max\{\mathrm{lfs}(s), \mathrm{lfs}(t)\} \leq d(s, m) = d(t, m)$, and due to the ε-sampling condition, $d(\frac{p+q}{2}, \frac{s+t}{2}) \leq \varepsilon d(s, m)$. We then have:

$$\begin{aligned} & d(\frac{p+q}{2}, m) \leq d(\frac{s+t}{2}, m) + d(\frac{p+q}{2}, \frac{s+t}{2}) \leq (\cos\theta + \varepsilon) d(s, m); \\ \Rightarrow \quad & d(x, L_\beta) \leq d(x, \frac{p+q}{2}) \leq d(x, m) + d(m, \frac{p+q}{2}) \leq d(x, m) + d(s, m)(\cos\theta + \varepsilon). \end{aligned} \tag{1}$$

Since s is a closest point of m in X, we have $d(s, m) = d(m, \mathsf{X}) \leq d(x, m)$. Combining this with (1), it follows that

$$d(x, L_\beta) \leq (1 + \cos\theta + \varepsilon) \cdot d(x, m).$$

■

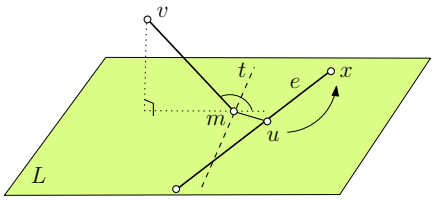
We bound the distance $d(x, M_\theta)$ with $\mathrm{wlfs}(x)$ by observing the following. The critical points of a distance function $d : \mathbb{R}^k \to \mathbb{R}$ can be characterized by points $x \in \mathbb{R}^k$ that have the zero gradient ∇d along every unit vector originating at x; see Grove [18]. It is also known that the critical points of the distance function d_X lie in the medial axis M. They are points $m \in M$ so that the convex hull $\mathrm{Conv}\,(\Pi(m))$ of all nearest neighbors of m in X contains m. Intiutively, this means that there exists a pair of points x, y in $\Pi(m)$ so that the angle $\angle xmy$ is large. We use this angle condition to avoid the critical points. Specifically, we show the following result for manifolds of arbitrary codimension which helps to make the angle condition precise.

Proposition 2.8 *Let the ambient dimension $k \geq 1$ and $m \in M$ be a critical point of the distance function d_X. There exists a pair of points $x, y \in \Pi(m)$ so that $\angle xmy \geq \frac{\pi}{2}$.*

Proof: It is known that any critical point m of the distance function d_X is in the convex hull $C = \mathrm{Conv}\,\Pi(m)$ of the points in $\Pi(m)$. This convex hull C is a j-polytope for some $j \leq k$. We can assume that j is at least 2, because otherwise, C is an edge with endpoints say $x, y \in \Pi(m)$, and $\angle xmy = \pi > \frac{\pi}{2}$.

Now consider the subspace $\mathbb{R}^j \subseteq \mathbb{R}^k$ that contains the j-polytope C. Choose an arbitrary 2-flat H passing through m in this $\mathbb{R}^j$. The intersection of H and C is a polygon that contains m. There is at least a pair of vertices u, v of this polygon so that $\pi \geq \angle umv \geq \frac{\pi}{2}$. The vertices u and v are the intersection of the 2-flat with the two codimension-2 faces U and V of C respectively which are $(j-2)$-faces.

Let e be the maximal line segment contained in U that connects u and a vertex of U. We can show that, one can choose an endpoint, say x, of e so that the angle $\angle umv$ remains at least $\frac{\pi}{2}$ when u assumes the position of x. To see this consider the plane L spanned by the line of e and the point m (see figure on the right). Let t be the line perpendicular to the orthogonal projection of mv. Observe that all points $z \in e$ makes an angle $\angle zmv$ of at least $\frac{\pi}{2}$ if z lies in the halfplane of L delimited by t which does not contain the projection of mv. Then, one of the endpoints of e must satisfy this condition because $u \in e$ does so to ensure $\angle umv \geq \frac{\pi}{2}$.



The chosen endpoint x of e is either a vertex of C or a point in a lower dimensional face of U. Keeping u at x, we can let v coincide with a similar endpoint of a line segment in V while keeping the angle $\angle umv$ at least $\frac{\pi}{2}$. Therefore, continuing this process, u and v either reach a vertex of C or a lower dimensional face. It follows that both will reach a vertex of C eventually while keeping the angle $\angle umv \geq \frac{\pi}{2}$. These two vertices qualify for x and y in the proposition. ∎

Remark 2.9 *We remark that the above bound of $\frac{\pi}{2}$ can be further tightened with a term depending on the dimension k. However, the bound of $\frac{\pi}{2}$ suffices for our results.*

The following assertion is now immediate.

Proposition 2.10 *For $\theta \leq \frac{\pi}{4}$, every point $x \in \mathsf{X}$ satisfies $d(x, M_\theta) \leq \mathrm{wlfs}(x)$.*

Propositions 2.7 and 2.10 together proves the upper bound of the lnfs_β claimed in Theorem 2.5. Next, we show the lower bound.

Proposition 2.11 *For every sample point $p \in P$, we have $\mathrm{lnfs}_\beta(p) > c_2 \cdot \mathrm{lfs}(p)$ where $c_2 = \frac{2c_0c_\beta}{1+c_0+2c_0c_\beta}$ and $c_0 = \sin(\beta - \nu_\varepsilon)$.*

Proof: Let z be the nearest point to p in L_β, and (p', q') the β-good pair that gives rise to z (thus z is the midpoint of $p'q'$). By definition of a β-good pair, $\angle(\tilde{\mathbf{N}}_{p'}, p'q')) \leq \frac{\pi}{2} - \beta$ and hence $\angle(\mathbf{N}_{p'}, p'q') \leq \frac{\pi}{2} - \beta + \nu_\varepsilon$. There is a medial ball B tangent to the manifold X at p' so that the half line $p'o$ going through the center o of this ball B realizes the angle $\angle(\mathbf{N}_{p'}, p'q')$. Hence, $\angle op'q' \leq \frac{\pi}{2} - \beta + \nu_\varepsilon$. It follows that

$$d(p', z) = \frac{1}{2} d(p', q') \geq d(p', o) \cos(\frac{\pi}{2} - \beta + \nu_\varepsilon) \geq c_0 \cdot \mathrm{lfs}(p'), \text{ where } c_0 = \sin(\beta - \nu_\varepsilon). \tag{2}$$

The empty ball condition of the β-good pair means that $2c_\beta d(p', z) \leq d(p, z)$, that is, $d(p', z) \leq \frac{d(p,z)}{2c_\beta}$. It then follows that

$$d(p, p') \leq d(p, z) + d(p', z) \leq (1 + \frac{1}{2c_\beta}) d(p, z).$$

By the 1-Lipschitz property of the lfs function and (2), we have:

$$\begin{aligned} \mathrm{lfs}(p) &\leq \mathrm{lfs}(p') + d(p, p') \leq \mathrm{lfs}(p') + (1 + \frac{1}{2c_\beta}) d(p, z) \leq \frac{1}{c_0} d(p', z) + (1 + \frac{1}{2c_\beta}) d(p, z) \\ &\leq \frac{1}{2c_0c_\beta} d(p, z) + (1 + \frac{1}{2c_\beta}) d(p, z) = (1 + \frac{1}{2c_\beta} + \frac{1}{2c_0c_\beta}) \cdot d(p, z). \end{aligned}$$

Setting $c_2 = \frac{1}{1+\frac{1}{2c_\beta}+\frac{1}{2c_0c_\beta}} = \frac{2c_0c_\beta}{1+c_0+2c_0c_\beta}$, we have that $d(p, z) = \mathrm{lnfs}_\beta(p) \geq c_2 \cdot \mathrm{lfs}(p)$, which proves the proposition. ∎

We will see later that, β is fixed at a constant value of $\frac{\pi}{5}$. For this choice of β, c_2 is not unusually small.

2.2 Computations for sparsification

In this section we describe the algorithm LEAN that takes a standard ε-dense sample P w.r.t. lfs of a hidden manifold $\mathsf{X} \subset \mathbb{R}^k$ of known intrinsic dimension, and outputs a *sparsified set* $Q \subseteq P$. The set Q is both adaptive and locally uniform as stated afterward in Theorem 2.12. The parameter ρ is chosen later to be a fixed constant less than 1.

Algorithm 1 LEAN(P, β, ρ)

1: $L_\beta := \emptyset$;
2: **for** every pair $(p, q) \in P \times P$ **do**
3: if (p, q) is a β-good pair then $L_\beta := L_\beta \cup \{\frac{p+q}{2}\}$
4: **end for**
5: Put P in a max priority queue $\mathbb{Q}$ with priority $\mathrm{lnfs}_\beta(p)$ for $p \in P$;
6: **while** $\mathbb{Q}$ not empty **do**
7: $q :=$extract-max($\mathbb{Q}$); $Q := Q \cup \{q\}$;
8: delete any p from $\mathbb{Q}$ if $d(q, p) \leq \rho \mathrm{lnfs}_\beta(q)$
9: **end while**

The sparsification is based on the lean set L_β, which is computed in lines 2–4 of the algorithm. We note that checking whether a pair (p, q) is β-good or not requires no parameter other than β, which is set to a fixed constant $\frac{\pi}{5}$ later in the homology inference algorithm. Clearly, $|L_\beta| = O(|P|^2)$ (see Section 2.3 for improving $|L_\beta|$ to $O(|P|)$). There is one implementation detail which involves the estimation of the normal space $\tilde{\mathbf{N}}_p$ for every point $p \in P$. This estimation step is oblivious to any parameter but requires the intrinsic dimension s of X to be known.

We estimate the tangent space T_p (thus the normal space) of X at a point $p \in P$ as follows. Let s be the intrinsic dimension of the manifold X. Let $p_1 \in P$ be the nearest neighbor of p in $P \setminus \{p\}$. Suppose we have already obtained points $\sigma_i = \{p, p_1, \ldots, p_i\}$ with $i < s$. Let $\mathrm{aff}(\sigma_i)$ denote the affine hull of the points in σ_i. Next, we choose $p_{i+1} \in P$ that is closest to p among all points forming an angle within the range $[\frac{\pi}{2} - \frac{\pi}{5}, \frac{\pi}{2}]$ with $\mathrm{aff}(\sigma_i)$. We add p_{i+1} to the set and obtain $\sigma_{i+1} = \{p, p_1, \ldots, p_i, p_{i+1}\}$. This process is repeated until $i + 1 = s$, the dimension of X, at which point we have obtained $s + 1$ points $\sigma_s = \{p, p_1, \ldots, p_s\}$. We use $\mathrm{aff}(\sigma_s)$ to approximate the tangent space T_p. The simplex σ_s obtained this way has good thickness property, which by Corollary 2.6 in [5] implies that the angle between the tangent space and the estimated tangent space at p (thus also the angle between the normal space and the estimated normal space at p) is bounded by $O(\varepsilon)$. The big-O hides terms depending only on the intrinsic property of the manifold. See Appendix B for details. In other words, we have that the error ν_ε in the estimated normal spaces (as required in Assumption 2.2) is $O(\varepsilon)$.

Next, we put the points of P in a priority queue and process them in non-decreasing order of their distances to L_β. We iteratively remove the point q with maximum value of $d(q, L_\beta)$ from the queue and proceed as follows. We put q into the sparse set Q and delete any point from the queue that lies at a distance of at most $\rho \mathrm{lnfs}_\beta(q)$ from q. Since we consider points in non-decreasing order of their distances to L_β, no earlier point that is already in the sparse set Q can be deleted by this process.

Determining if a pair (p, q) is β-good takes $O(|P|)$ time. This linear complexity is mainly due to the range queries for balls required for testing the 'empty ball' condition 2 for β-goodness. Therefore, for $L_\beta = O(|P|^2)$, the algorithm spends $O(|P|^3)$ time in total. This can be slightly improved to $O(|P|^{2-\frac{1}{k}} 2^{O(\log^* |P|)})$ using general spherical range query data structure in the ambient space $\mathbb{R}^k$ [1]. Once the lean set is computed, the computation of lnfs for all points involves computing the nearest neighbor in L_β for each point $p \in P$. Using the method described in section 2.3, we can bring down the lean set size to $O(|P|)$. Then, computing lnfs_β takes at most $O(|P|^2)$ time in total. The actual sparsification in steps 6-9 takes only

$O(|Q|^2) = O(|P|^2)$ time.

We show that the sparsification by LEAN leaves the point set Q locally uniform w.r.t. lnfs_β. The proof appears in Appendix A.

Theorem 2.12 *Let P be a sample of a manifold $\mathsf{X} \subseteq \mathbb{R}^k$, which is ε-dense w.r.t.* lfs*. For $\rho \leq \frac{1}{12}$, the output of $Lean(P, \beta, \rho)$ is a $(\frac{4}{3}\rho, \rho)$-uniform sample of X w.r.t. lnfs_β when $\varepsilon > 0$ is sufficiently small.*

2.3 Linear-size Lean Set

Observe that, the size $|L_\beta|$ is $O(n^2)$ if the input sample P has size n. This is far less than $O(n^{\lceil \frac{k}{2} \rceil})$, k being the ambient dimension, which one incurs if the medial axis is approximated with the Voronoi diagrams [10, 16]. We can further thin down the lean set to a linear size $O(n)$ for any fixed k by the following simple strategy:

For every $p \in P$, among all β-good pairs (p, q) it forms, we choose the pair (p, q^*) such that the distance $d(p, q^*)$ is the smallest. We call this pair (p, q^*) the *minimal β-good pair* for p. We now take a reduced lean set, denoted by $\widehat{L}_\beta$, as the collection of midpoints of these minimal β-good pairs. Obviously, $|\widehat{L}_\beta| = O(n)$.

Below we show that this reduced lean set can replace the original lean set L_β: it only worsens the distance from a sample point to the lean set by an additional constant factor. Note that this is the only distance required by the algorithm (and the homology inference in Theorem 3.10). In particular, we have the following result.

Theorem 2.13 *For any point $p \in P$, we have that* $\mathrm{lnfs}_\beta(p) \leq d(p, \widehat{L}_\beta) \leq (1 + \frac{1}{c_\beta})\mathrm{lnfs}_\beta(p)$.

Proof: The left inequality is trivial since $\widehat{L}_\beta \subseteq L_\beta$. We will show the right inequality. Fix any sample point $p \in P$, and let $m \in L_\beta$, the midpoint of a β-good pair (s, t), be p's nearest neighbor in the original lean set L_β.

Let (s, t^*) be the minimal β-good pair for s, and m^* its midpoint. We now show that $d(p, \widehat{L}_\beta) \leq d(p, m^*) \leq (1 + \frac{1}{c_\beta})d(p, L_\beta)$. Indeed, since (s, t^*) is the minimal β-good pair for s, we have that $d(s, t) \geq d(s, t^*)$. Hence

$$d(m, m^*) \leq d(m, s) + d(s, m^*) \leq \frac{1}{2}(d(s, t) + d(s, t^*)) \leq d(s, t).$$

At the same time, by the empty-ball property of a β-good pair, we have that $d(p, m) \geq c_\beta d(s, t)$; that is, $d(s, t) \leq \frac{1}{c_\beta} d(p, m)$. Putting everything together, we obtain:

$$d(p, \widehat{L}_\beta) \leq d(p, m^*) \leq d(p, m) + d(m, m^*) \leq d(p, m) + d(s, t) \leq (1 + \frac{1}{c_\beta})d(p, m) = (1 + \frac{1}{c_\beta})d(p, L_\beta).$$

The claim then follows. ∎

3 Homology inference

In this section, we aim to infer homology groups of a hidden manifold X from its point samples. Let $\mathsf{H}_i(\cdot)$ denote the i-dimensional homology group. It refers to the singular homology when the argument is a manifold or a compact set, and to the simplicial homology when it is a simplicial complex. All homology groups in this paper are assumed to be defined over the finite field $\mathbb{Z}_2$. For details on homology groups, see e.g. [21].

The homology inference from a point sample of a hidden manifold X has been researched extensively in the literature [9, 12, 15, 22]. However, most of these works assume that the given sample $P \subset \mathsf{X}$ is

globally dense, that is, ε-dense w.r.t. to the *infimum* of lfs or wlfs. This strong assumption lets one to infer the homology from an appropriate offset of P w.r.t. the distance $d(x, P)$, which is represented with the union of balls of equal radii around the sample points. As we indicated in the introduction, unfortunately, when the sample is *adaptive* (ε-dense w.r.t. a non-constant function ϕ), there may not be such choice of a global radius so that the offset captures the topology of X.

To circumvent this problem, one needs to scale the distance with the function ϕ that provides the adaptivity. This idea was used in [9] where ϕ is taken as lfs. Approximating lfs is difficult, so we use lnfs_β instead for scaling. Observe that the offset may intersect the medial axis, but we argue that we can compute relevant offsets that never contains the critical points of the scaled distance, thereby ensuring topological fidelity.

3.1 Scaled distance and its offsets

In what follows, we develop the results in more generality by scaling the distance d_{X} with the distance to a finite set $L \subset \mathbb{R}^k$. Later, in computations, we replace L by the lean set $L_{\frac{\pi}{5}}$ and the distance $d(x, L)$ with $\mathrm{lnfs}_{\frac{\pi}{5}}$ for $x \in \mathsf{X}$. Recall that $\Pi(x)$ denotes the set of closest neighbors of x in X.

Definition 3.1 *Given a finite set $L \subset \mathbb{R}^k$ such that $L \cap \mathsf{X} = \emptyset$, Let $h_L : \mathbb{R}^k \to \mathbb{R}$ be a scaled distance to the manifold where*

$$h_L(x) = \frac{d(x, \mathsf{X})}{d(x, \mathsf{X}) + d(x, L)} = \frac{d(x, \Pi(x))}{d(x, \Pi(x)) + d(x, L)} \text{ and let } \mathsf{X}_\alpha = h_L^{-1}[0, \alpha).$$

We avoid the obvious choice of $h_L(x) = \frac{d(x,\mathsf{X})}{d(x,L)}$ because that makes $h_L(x)$ unbounded at L. We are interested in analyzing the topology of the α-offsets (which are open) X_α of h_L when $\mathsf{X}_\alpha \setminus \mathsf{X}$ does not include any critical points of h_L. Clearly, $\mathsf{X} \subset \mathsf{X}_\alpha$ for any $\alpha > 0$, as $L \cap \mathsf{X} = \emptyset$. This brings us to the concept of flow induced by the distance function which was studied in [18] and later used in the context of sampling theory [10, 19, 20]. The vector field $\nabla_{d_{\mathsf{X}}}$ as we defined earlier is not continuous. However, as it is shown in [20], there exists a continuous flow $\mathsf{F} : \mathbb{R}^k \setminus \mathsf{X} \times \mathbb{R}^+ \to \mathbb{R}^k \setminus \mathsf{X}$ such that $\mathsf{F}(x, t) = x + \int_0^t \nabla_{d_{\mathsf{X}}}(\mathsf{F}(x, \tau))d\tau$. For a point $x \in \mathbb{R}^k \setminus \mathsf{X}$, the image $\mathsf{F}(x, [0, t))$ of an half-open interval $[0, t)$ is called its *flow line*. For a point $x \notin \mathsf{X} \cup M$, where M is the medial axis of X, the flowline $\mathsf{F}(x, [0, \infty))$ first coincides with the line segment $x\Pi(x)$ which is normal to the manifold X. Once it reaches the medial axis M, it stays in M. We show that h_L increases along the flow line of d_{X} in the α-offset that we are interested in. This, in turn, implies that the α-offset of our interest avoids the critical points of h_L.

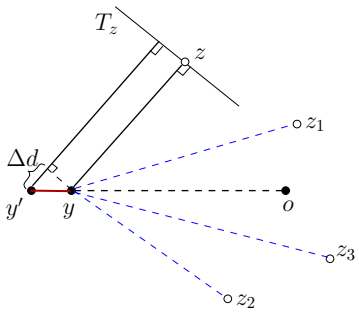
Proposition 3.2 *For $\theta \leq \frac{\pi}{4}$, $\alpha < \frac{\cos 2\theta}{1+\cos 2\theta}$ and $M_\theta \cap \mathsf{X}_\alpha = \emptyset$, the function h_L increases along the flow line on the piece $\mathsf{X}_\alpha \cap \mathsf{F}(x, [0, \infty))$ where x is any point in $\mathsf{X}_\alpha \setminus \mathsf{X}$.*

Proof: First, observe that, due to Proposition 2.8, we can assert that $\mathsf{X}_\alpha \setminus \mathsf{X}$ contains no critical point of d_{X} since $\mathsf{X}_\alpha \cap M_\theta = \emptyset$ and $\theta \leq \frac{\pi}{4}$. Therefore, flow lines for every point $x \in \mathsf{X}_\alpha \setminus \mathsf{X}$ are (topological) segments. Consider an arbitrary point $y = \mathsf{F}(x, t)$ such that $y \in \mathsf{X}_\alpha$. Set $d = d(y, \mathsf{X})$ and $\tilde{d} = d(y, L)$. Since $y \in \mathsf{X}_\alpha$, we have

$$h_L(y) < \alpha \implies \frac{\tilde{d}}{d} > \frac{1 - \alpha}{\alpha}. \tag{3}$$

For arbitrary small $\Delta t > 0$, let Δd and $\Delta\tilde{d}$ denote the changes in the distances d and $\tilde{d}$ respectively when we move on the flow line from $y = \mathsf{F}(x, t)$ to $y' = \mathsf{F}(x, t + \Delta t)$. Observe that by the triangle inequality, $\Delta\tilde{d} = |d(y, L) - d(y', L)| \leq d(y, y')$. We claim that $\Delta d \geq d(y, y') \cdot \cos 2\phi$ where ϕ is the supremum angle so that any point of $\mathsf{X}_\alpha \cap M$ belongs to M_ϕ.

The flow line $\mathsf{F}(x, [0, \infty))$ follows a direction that is normal to the manifold X when it does not lie in the medial axis M of X. If y lies on a portion of the flow line which is normal to the manifold X, then it is easy to see that $\Delta d = |d(y, \mathsf{X}) - d(y', \mathsf{X})| = d(y, y') \geq d(y, y') \cdot \cos 2\phi$. If y lies on a portion of the flow line which is contained in the medial axis M, then the definition of ϕ implies that, for any two points $z_1, z_2 \in \Pi(y)$, the angle $\angle z_1 y z_2 \leq 2\phi$. At the same time, it is known that if $y \in M$, then the flow direction $\nabla_{d_\mathsf{X}}(y)$ at $y = \mathsf{F}(x, t)$ points in the direction of $\overrightarrow{oy}$ where o is the center of the minimum enclosing ball for $\Pi(y)$ (see e.g, [20]). In fact, o must be contained in the convex hull of points in $\Pi(y)$. This further leads to that there exists a pair of points $z_1, z_2 \in \Pi(y)$ so that the angle between $\overrightarrow{oy}$ and $\overrightarrow{zy}$ for any $z \in \Pi(y)$ is at most the angle $\angle z_1 y z_2$, which is at most 2ϕ. See the figure for an illustration where $\Pi(y) = \{z, z_1, z_2, z_3\}$, and T_z is the intersection of the tangent space of X at z with the plane spanned by o, y, z. Hence, in the limit as $y' \to y$, $\Delta d \to d(y, y') \cdot \cos \angle oyz$ for some $z \in \Pi(y)$, implying $\Delta d \geq d(y, y') \cdot \cos(2\phi)$.



Finally, note that in the claim, we require that $M_\theta \cap \mathsf{X}_\alpha = \emptyset$. By definition of ϕ, this means that $\theta > \phi$. Hence, for $\theta \leq \frac{\pi}{4}$, $\frac{\cos 2\theta}{1+\cos 2\theta} = 1 - \frac{1}{1+\cos 2\theta} \leq 1 - \frac{1}{1+\cos 2\phi}$. The condition $\alpha < \frac{\cos 2\theta}{1+\cos 2\theta}$ now provides that $\frac{1}{\cos 2\phi} < \frac{1-\alpha}{\alpha}$. It follows that:

$$\frac{\Delta\tilde{d}}{\Delta d} \leq \frac{1}{\cos 2\phi} < \frac{1-\alpha}{\alpha} \leq \frac{\tilde{d}}{d} \Longrightarrow \frac{\tilde{d} + \Delta\tilde{d}}{d + \Delta d} < \frac{\tilde{d}}{d} \Longrightarrow h_L(\mathsf{F}(x,t)) < h_L(\mathsf{F}(x, t + \Delta t)).$$

■

Now, we will show that the α-offset X_α remains homotopy equivalent to X if α is chosen appropriately. For the standard distance function d_X, such a result is well known [9, 11] which uses the critical point theory of Grove [18]. Here, we need the result for the scaled distance h_L for which we cannot use this theory. We establish it using Proposition 3.2 and a result from topology presented in Spanier [24].

Proposition 3.3 *Let $\theta \leq \frac{\pi}{4}$ and $\alpha < \frac{\cos 2\theta}{1+\cos 2\theta}$. Let X_α be as defined in proposition 3.2 where $\mathsf{X}_\alpha \cap M_\theta = \emptyset$. For any $\alpha' > 0$ where $\alpha' \leq \alpha$, the inclusion $X_{\alpha'} \xrightarrow{i} X_\alpha$ induces an isomoprhism $\mathsf{H}_q(\mathsf{X}_{\alpha'}) \overset{i_*}{\cong} \mathsf{H}_q(\mathsf{X}_\alpha)$ in each dimension $q \geq 0$. Furthermore, $\mathsf{H}_q(\mathsf{X}_{\alpha'}) \cong \mathsf{H}_q(\mathsf{X})$.*

Proof: For any real $a > 0$, set $D_a := \mathbb{R}^d - \mathsf{X}_a$. Let $B = Closure\,(D_{\alpha'} \setminus D_\alpha)$. Any point $x \in B$ has a flow line $\mathsf{F}(x, [0, t])$ along which h_L strictly increases (Proposition 3.2). In particular, there is a unit vector originating at x along which ∇_{h_L} does not vanish. Therefore, B does not contain any critical point of h_L. It follows that $D_{\alpha'}$ deformation retracts to D_α along the gradient flow of d_X which means that the inclusion $D_\alpha \subseteq D_{\alpha'}$ is a homotopy equivalence and hence, for each integer $q \geq 0$, $\mathsf{H}_q(D_\alpha) \to \mathsf{H}_q(D_{\alpha'})$ is an isomorphism induced by inclusion.

Now we use a result from page 289, Spanier [24] for pairs of subsets (A, B) in $\mathbb{R}^d$. The result implies that, with homology and cohmology defined over $\mathbb{Z}_2$, the following diagram commutes for every dimension $q \geq 0$ where $i : (A', B') \subseteq (A, B)$, $j : (\mathbb{R}^d - B, \mathbb{R}^d - A) \subseteq (\mathbb{R}^d - B', \mathbb{R}^d - A')$, and i^*, j_* are inclusion induced homomorphisms:

$$\begin{array}{ccc} \mathsf{H}_{(d-q)}(\mathbb{R}^d - B, \mathbb{R}^d - A) & \xrightarrow{\cong} & \mathsf{H}^q(A, B) \\ \downarrow{\scriptstyle j_*} & & \downarrow{\scriptstyle i^*} \\ \mathsf{H}_{(d-q)}(\mathbb{R}^d - B', \mathbb{R}^d - A') & \xrightarrow{\cong} & \mathsf{H}^q(A', B') \end{array}$$

Taking both B, B' to be empty sets, and $A = X_\alpha$, $A' = X_{\alpha'}$, we get the following diagram that commutes:

$$\begin{array}{ccc} \mathsf{H}_{(d-q)}(\mathbb{R}^d, D_\alpha) & \xrightarrow{\cong} & \mathsf{H}^q(\mathsf{X}_\alpha, \emptyset) \\ \downarrow j_* & & \downarrow i^* \\ \mathsf{H}_{(d-q)}(\mathbb{R}^d, D_{\alpha'}) & \xrightarrow{\cong} & \mathsf{H}^q(\mathsf{X}_{\alpha'}, \emptyset) \end{array}$$

Since $\mathsf{H}_{(d-q)}(D_\alpha) \to \mathsf{H}_{(d-q)}(D_{\alpha'})$ is an isomorphism induced by inclusion, we have that j_*, the left vertical map, is an isomorphism. Then, from the commutativity of the diagram, we get that, i^*, the right vertical map, is also an isomorphism. Switching the cohomology with homology using the universal coefficient theorem under $\mathbb{Z}_2$, we get that

$$\mathsf{H}_q(\mathsf{X}_{\alpha'}) = \mathsf{H}_q(\mathsf{X}_{\alpha'}, \emptyset) \xrightarrow{i_*} \mathsf{H}_q(\mathsf{X}_\alpha, \emptyset) = \mathsf{H}_q(\mathsf{X}_\alpha).$$

This completes the proof of the first claim.

Now we show the second cliam that $\mathsf{H}_q(\mathsf{X}) \cong \mathsf{H}_q(\mathsf{X}_{\alpha'})$. For any point $x \in \mathsf{X}_{\alpha''} \setminus \mathsf{X}$, a flow line $\mathsf{F}(x, [0, t])$ cannot re-enter $\mathsf{X}_{\alpha''}$ once it exits because of the monotonicity of h_L. This means $\mathsf{F}(x, [0, t])$ intersects $\mathsf{X}_{\alpha''}$ in one connected segment. Let x' be the unique point where $\mathsf{F}(x, [0, t])$ intersects the hypersurface $h_L^{-1}(\alpha'')$. Since X is compact and smooth, by choosing $\alpha'' > 0$ sufficiently small, one can ensure that $\mathsf{F}(x, [0, t]) \cap \mathsf{X}_{\alpha''}$ lies on the normal line segment $x\Pi(x)$, for all $x \in \mathsf{X}_{\alpha''} \setminus \mathsf{X}$. It implies that $\mathsf{X}_{\alpha''}$ intersects the normal lines to X in a connected segment along which $\mathsf{X}_{\alpha''}$ can be retracted to X completing the proof that $\mathsf{H}_q(\mathsf{X}_{\alpha''})$ is isomorphic to $\mathsf{H}_q(\mathsf{X})$ for a sufficiently small $\alpha'' > 0$. We can extend the claim to all $\alpha' \leq \alpha$ by using the previous claim that inclusion induces an isomorphism $\mathsf{H}_q(\mathsf{X}_{\alpha''}) \cong \mathsf{H}_q(\mathsf{X}_{\alpha'})$ for all $\alpha'' \leq \alpha' \leq \alpha$. ■

3.2 Interleaving and inference

Our goal is to interleave the α-offsets of h_L with the union of a set of open balls $\bigcup B$ centered at the sample points because then, following the approach in [12], we can relate the topology of the nerve complex of $\bigcup B$ with that of X. For the standard distance function d_X, the offsets restricted to the sample P provide the required set of balls because $d_\mathsf{X}|_P$ approximates d_X. Unfortunately, offsets of h_L restricted to P are not necessarily union of geometric balls centering points in P. Nevertheless, we show that a set of balls whose radii are proportional to the distances to L have the necessary property.

First, we consider the union of balls, one for every point in X. Let $\bigcup B_\alpha$ denote the union of open balls $B(x, r)$ for every $x \in \mathsf{X}$ where $r = \alpha d(x, L)$. One has the following interleaving result.

Proposition 3.4 $\mathsf{X}_{\frac{\alpha}{1+2\alpha}} \subseteq \bigcup B_\alpha \subseteq \mathsf{X}_\alpha$.

Proof: First we show the left inclusion. Let x be any point in $\mathsf{X}_{\frac{\alpha}{1+2\alpha}}$, and y an arbitrary point from $\Pi(x)$ (i.e, $d(x, y) = d(x, \Pi(x))$). Then we have,

$$\begin{aligned} \frac{d(x,y)}{2d(x,y) + d(y,L)} &= \frac{d(x,y)}{d(x,y) + (d(x,y) + d(y,L))} \\ &\leq \frac{d(x,y)}{d(x,y) + d(x,L)} < \frac{\alpha}{1+2\alpha} \text{ since } x \in \mathsf{X}_{\frac{\alpha}{1+2\alpha}} \end{aligned}$$

It then follows that

$$(1+2\alpha)d(x,y) < 2\alpha d(x,y) + \alpha d(y,L) \Longrightarrow d(x,y) < \alpha d(y,L) \Longrightarrow x \in \bigcup B_\alpha.$$

We now prove the second inclusion. Let x be any point in $\bigcup B_\alpha$. Let $z \in \mathsf{X}$ be a point so that $x \in B(z, \alpha d(z, L))$; that is, $d(x, z) < \alpha d(z, L)$. Such a point exists by the definition of $\bigcup B_\alpha$. Using triangle inequality, we have:

$$h_L(x) = \frac{d(x, \mathsf{X})}{d(x, \mathsf{X}) + d(x, L)} \leq \frac{d(x, z)}{d(x, z) + d(x, L)} \leq \frac{d(x, z)}{d(z, L)} < \frac{\alpha d(z, L)}{d(z, L)} = \alpha.$$

■

We extend the above interleaving result to the union of open balls whose centers are restricted only to a sample $P \subset \mathsf{X}$. For convenience we define the following sampling condition closely related the ε-dense sampling condition.

Definition 3.5 *A finite set $P \subset \mathsf{X}$ is a (δ, L)-sample of X if every point $x \in \mathsf{X}$ has a point $p \in P$ so that $d(x, p) \leq \delta d(p, L)$. Furthermore, let $\bigcup P_\alpha = \bigcup_{p \in P} B(p, \alpha d(p, L))$ denote the union of scaled open balls around sample points in P.*

Remark 3.6 *A δ-dense sample w.r.t. lnfs_β is also a $(\frac{\delta}{1-\delta}, L_\beta)$-sample of X. Conversely, a (δ, L_β)-sample of X is also a $\frac{\delta}{1-\delta}$-dense sample w.r.t. lnfs_β. These follow from the fact that lnfs_β is 1-Lipschitz.*

Proposition 3.7 *For a (δ, L)-sample P of X and any $\alpha > 0$, we have $\mathsf{X}_{\frac{\alpha}{1+2\alpha}} \subseteq \bigcup P_{\alpha+\delta+\alpha\delta} \subseteq \mathsf{X}_{\alpha+\delta+\alpha\delta}$.*

Proof: Recall that by definition $\bigcup B_\alpha = \bigcup_{x \in \mathsf{X}} B(x, \alpha d(x, L))$. By the (δ, L)-sampling condition of P, as well as triangle inequality, we have $\bigcup B_\alpha \subseteq \bigcup P_{\alpha+\delta+\alpha\delta}$. Combining this with the left inclusion in Proposition 3.4, we have

$$\mathsf{X}_{\frac{\alpha}{1+2\alpha}} \subseteq \bigcup B_\alpha \subseteq \bigcup P_{\alpha+\delta+\alpha\delta}.$$

The second inclusion follows because $\bigcup P_{\alpha+\delta+\alpha\delta} \subseteq \bigcup B_{\alpha+\delta+\alpha\delta}$ and $\bigcup B_{\alpha+\delta+\alpha\delta} \subseteq \mathsf{X}_{\alpha+\delta+\alpha\delta}$ (Proposition 3.4). ■

With the isomorphisms in the homology groups of the offset of our scaled distance function (Proposition 3.3) and the interleaving result (Proposition 3.7), we can infer the homology of the hidden manifold X from the union of balls $\bigcup P_\alpha$.

Suppose that P is a (δ, L)-sample of the manifold X. Recall that $\bigcup P_\alpha$ denotes the union of open balls $\bigcup_{p \in P} B(p, \alpha d(p, L))$ centered at each point $p \in P$, with radius $\alpha d(p, L)$. Note that the parameter α does not stand for *distance threshold*, but a *scale parameter* for the distance $d(p, L)$. This parameter is universal for all points, while the distance $d(p, L)$ makes the union of balls adaptive.

By manipulating the result in Proposition 3.7, one obtains that, for $\alpha + \delta \leq \frac{1}{4}$ and $\alpha' = \frac{\alpha}{2(1-\alpha)}$,

$$\mathsf{X}_{\frac{\alpha}{2}} \subseteq \bigcup P_{\alpha'+\delta+\alpha'\delta} \subseteq \bigcup P_{\frac{5}{4}\alpha'+\delta} \subseteq \bigcup P_{\alpha+\delta}.$$

When $\alpha + \delta \leq \frac{1}{6}$ and $\alpha' = \frac{\alpha+\delta}{1-2(\alpha+\delta)}$, similar manipulation gives

$$\mathsf{X}_{\alpha+\delta} \subseteq \bigcup P_{\alpha'+\delta+\alpha'\delta} \subseteq \bigcup P_{\frac{7}{6}\alpha'+\delta} \subseteq \bigcup P_{\frac{11}{4}(\alpha+\delta)} \subseteq \bigcup P_{3(\alpha+\delta)}.$$

So, for $\alpha + \delta \leq \frac{1}{6}$, we obtain

$$\mathsf{X}_{\frac{\alpha}{2}} \subseteq \bigcup P_{\alpha+\delta} \subseteq \mathsf{X}_{\alpha+\delta} \subseteq \bigcup P_{3(\alpha+\delta)} \subseteq \mathsf{X}_{3(\alpha+\delta)}, \tag{4}$$

which leads to inclusion-induced homomorphisms at the homology level that interleave:

$$\mathsf{H}_i(\mathsf{X}_{\frac{\alpha}{2}}) \to \mathsf{H}_i\left(\bigcup P_{\alpha+\delta}\right) \to \mathsf{H}_i(\mathsf{X}_{\alpha+\delta}) \to \mathsf{H}_i\left(\bigcup P_{3(\alpha+\delta)}\right) \to \mathsf{H}_i(\mathsf{X}_{3(\alpha+\delta)}).$$

On the other hand, if $3(\alpha+\delta) < \frac{\cos 2\theta}{1+\cos 2\theta}$ and $\mathsf{X}_{3(\alpha+\delta)} \cap M_\theta = \emptyset$, we can use Proposition 3.3 and Lemma 3.2 in [12] to claim that

$$\text{image}\left(\mathsf{H}_i\left(\bigcup P_{\alpha+\delta}\right) \to \mathsf{H}_i\left(\bigcup P_{3(\alpha+\delta)}\right)\right) \cong \mathsf{H}_i(\mathsf{X}).$$

Let $C^\alpha(P)$ denote the nerve of $\bigcup P_\alpha$. One can recognize the resemblance between $C^\alpha(P)$ and the well-known Čech complex. Both are nerves of unions of balls, but unlike Čech complexes, $C^\alpha(P)$ is the nerve of a union of balls that may have different radii; recall that α denotes a fraction relative to a distance rather than an absolute distance. The Nerve Lemma [6] provides that $C^\alpha(P)$ is homotopy equivalent to $\bigcup P_\alpha$. Also, the argument of Chazal and Oudot [12] to prove Theorem 3.5 can be extended to claim that for any $i \geq 0$,

$$\text{rank}\left(\mathsf{H}_i(C^{\alpha+\delta}(P)) \to \mathsf{H}_i(C^{3(\alpha+\delta)}(P)) = \text{rank}\left(\mathsf{H}_i\left(\bigcup P_{\alpha+\delta}\right) \to \mathsf{H}_i\left(\bigcup P_{3(\alpha+\delta)}\right)\right) = \text{rank}\,\mathsf{H}_i(\mathsf{X}).\right.$$

The complex $C^\alpha(P)$ interleaves with another complex $R^\alpha(P)$ that is reminiscent of the interleaving of the Čech with the Vietoris-Rips complexes. Specifically, let

$$R^\alpha(P) := \{\sigma \mid d(p,q) < \alpha(d(p,L) + d(q,L)) \text{ for every edge } pq \text{ of } \sigma\}.$$

It is easy to observe that $R^\alpha(P)$ is the completion of the 1-skeleton of $C^\alpha(P)$ and the following inclusions hold as in the case of the original Čech and Vietoris-Rips complexes.

$$C^\alpha(P) \subseteq R^\alpha(P) \subseteq C^{2\alpha}(P) \text{ for any } \alpha \geq 0.$$

Now, by choosing $\alpha+\delta \leq \frac{1}{6}\frac{\cos 2\theta}{1+\cos 2\theta}$ (which also implies $\alpha+\delta \leq \frac{1}{12}$ since $\frac{\cos 2\theta}{1+\cos 2\theta} \leq \frac{1}{2}$), we have a sequence similar to (4) that eventually induces the following sequence:

$$\mathsf{H}_i(C^{\alpha+\delta}(P)) \to \mathsf{H}_i(R^{\alpha+\delta}(P)) \to \mathsf{H}_i(C^{2(\alpha+\delta)}(P)) \to \mathsf{H}_i(C^{6(\alpha+\delta)}(P)) \to \mathsf{H}_i(R^{6(\alpha+\delta)}(P)) \to \mathsf{H}_i(C^{12(\alpha+\delta)}(P)).$$

In particular, following a similar argument as before, we have that

$$\text{rank}\left(\mathsf{H}_i(C^{\alpha+\delta}(P)) \to \mathsf{H}_i(C^{12(\alpha+\delta)}(P)) = \text{rank}\left(\mathsf{H}_i(C^{2(\alpha+\delta)}(P)) \to \mathsf{H}_i(C^{6(\alpha+\delta)}(P)) = \text{rank}\,\mathsf{H}_i(\mathsf{X})\right.\right.$$

as long as $12(\alpha+\delta) \leq \frac{\cos 2\theta}{1+\cos 2\theta}$ and $\mathsf{X}_{12(\alpha+\delta)} \cap M_\theta = \emptyset$. By using the standard results of interleaving [12] on this sequence, we obtain that

$$\text{rank}(\mathsf{H}_i(R^{\alpha+\delta}(P)) \to \mathsf{H}_i(R^{6(\alpha+\delta)}(P))) = \text{rank}\,\mathsf{H}_i(\mathsf{X}).$$

Theorem 3.8 *Let $\mathsf{X} \subset \mathbb{R}^k$ be a manifold and $L \subset \mathbb{R}^k$ be a finite set where $L \cap \mathsf{X} = \emptyset$. For some $\delta > 0$, let P be a (δ, L)-sample of X. Let $\theta \leq \frac{\pi}{4}$, and $\alpha+\delta \leq \frac{1}{12}\frac{\cos 2\theta}{1+\cos 2\theta}$. If $\mathsf{X}_{12(\alpha+\delta)} \cap M_\theta = \emptyset$, then* $\text{rank}(\mathsf{H}_i(\mathsf{X})) = \text{rank}(\mathsf{H}_i(R^{\alpha+\delta}(P)) \to \mathsf{H}_i(R^{6(\alpha+\delta)}(P)))$*, for any $i \geq 0$.*

3.3 Computations for topology inference

In step 3 of LEANTOPO, we compute the persistence homology induced by the inclusion $R^{2\rho}(Q) \to R^{12\rho}(Q)$ where $\rho = \frac{1}{26}\frac{\cos 2\beta}{1+\cos 2\beta}$. When the parameter ε is sufficiently small and $\beta = \frac{\pi}{5}$, we can find a value θ such that $\frac{\pi}{4} \geq \theta \geq \beta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon$ and $2\rho = \frac{1}{13}\frac{\cos 2\beta}{1+\cos 2\beta} \leq \frac{1}{12}\frac{\cos 2\theta}{1+\cos 2\theta}$. This is precisely what is needed for the homology inference in Theorem 3.8. More specifically, recall by 7 in the proof of Theorem 2.12, the output sparsified set of points Q is a $(\delta, L_{\frac{\pi}{5}})$-sample for $\delta = \frac{6}{5}\rho$. The algorithm implicitly sets $\alpha = 2\rho - \delta = \frac{4}{5}\rho$ such that $\alpha+\delta = 2\rho \leq \frac{1}{12}\frac{\cos 2\theta}{1+\cos 2\theta}$ when ε is sufficiently small. Theorem 3.8 requires further that the offset $\mathsf{X}_{\alpha'} := h^{-1}_{L_{\frac{\pi}{5}}}[0, \alpha')$ is disjoint from M_θ for $\alpha' = 12(\alpha+\delta)$ which we establish using the following proposition.

Algorithm 2 LEANTOPO(P)

1: $\beta := \frac{\pi}{5}$; $\rho := \frac{1}{26}\frac{\cos 2\beta}{1+\cos 2\beta}$; $Q := Lean(P, \beta, \rho)$;
2: Compute the complexes $R^{2\rho}(Q)$ and $R^{12\rho}(Q)$;
3: Compute the persistence induced by the inclusion $R^{2\rho}(Q) \to R^{12\rho}(Q)$.

Proposition 3.9 *Let $\alpha' \leq \frac{1}{1+\cos\theta+\varepsilon}$ and θ be such that $\frac{\pi}{2} \geq \theta \geq \frac{\pi}{5} + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon$ for a sufficiently small $\varepsilon \leq \frac{1}{8}\sin^2\theta$. Then, $M_\theta \cap \mathsf{X}_{\alpha'} = \emptyset$.*

Proof: We prove the result by contradiction. Assume that there exists a point $x \in M_\theta \cap \mathsf{X}_{\alpha'}$. Define m and s as in the proof of Proposition 2.7. With $\beta = \frac{\pi}{5}$, the assumed conditions for $\theta, \beta, \varepsilon$ are same as in Proposition 2.7, and thus we can arrive at the inequality 1 in its proof. Since s is a closest point of m in X, we have $d(s,m) = d(m,\mathsf{X}) \leq d(x,m) + d(x,\mathsf{X})$. Combining this with (1), it follows that, for any $x \in \mathsf{X}_{\alpha'}$,

$$d(x, L_{\frac{\pi}{5}}) \leq (1 + \cos\theta + \varepsilon) \cdot d(x,m) + (\cos\theta + \varepsilon) \cdot d(x,\mathsf{X}).$$

Since $x \in \mathsf{X}_{\alpha'}$, $h_{L_{\frac{\pi}{5}}}(x) < \alpha'$ implies that $d(x,\mathsf{X}) < \frac{\alpha'}{1-\alpha'} d(x, L_{\frac{\pi}{5}})$. Hence $d(x, L_{\frac{\pi}{5}}) < c \cdot d(x,m) = c \cdot d(x, M_\theta)$ for the positive constant $c = \frac{1+\cos\theta+\varepsilon}{1-\frac{\alpha'}{1-\alpha'}(\cos\theta+\varepsilon)} = 1 + \frac{\cos\theta+\varepsilon}{1-\alpha'(1+\cos\theta+\varepsilon)}$.

On the other hand, since $x \in M_\theta \cap \mathsf{X}_{\alpha'}$ and since $\mathsf{X} \cap M_\theta$ is empty, $x \notin \mathsf{X}$. Thus, $d(x,\mathsf{X}) > 0$. Since $x \in \mathsf{X}_{\alpha'}$ and $h_{L_{\frac{\pi}{5}}}(x) < \alpha'$, we have that $d(x, L_{\frac{\pi}{5}}) > \frac{1-\alpha'}{\alpha'} d(x,\mathsf{X})$. Hence $d(x, L_{\frac{\pi}{5}}) > 0$ as well since $\alpha' < 1$. This further implies that $d(x, M_\theta) > 0$ because according to the above derivation, $d(x, M_\theta) \geq \frac{1}{c} d(x, L_{\frac{\pi}{5}})$ for $c > 0$. This however contradicts the fact that $x \in M_\theta \cap \mathsf{X}_{\alpha'} \in M_\theta$. Hence our assumption is wrong and there is no such point $x \in M_\theta \cap \mathsf{X}_{\alpha'}$. ∎

Theorem 3.10 *Let $\mathsf{X} \subset \mathbb{R}^k$ be a smooth compact manifold without boundary of known intrinsic dimension. Let P be an ε-dense sample of X w.r.t.* lfs. LEANTOPO(P) *computes the rank of $\mathsf{H}_i(\mathsf{X})$ for any $i \geq 0$ when ε is sufficiently small.*

Proof: Since $\frac{\cos 2\theta}{1+\cos 2\theta} \leq \frac{1}{1+\cos\theta+\varepsilon}$ for $\theta \leq \frac{\pi}{2}$ and small enough ε, one has the fact that $\alpha' = 12(\alpha+\delta) \leq \frac{\cos 2\theta}{1+\cos 2\theta}$ implies that $\alpha' \leq \frac{1}{1+\cos\theta+\varepsilon}$. This means that the parameters α, and θ set by the algorithm LEANTOPO satisfy the conditions required by Proposition 3.9. Hence, $\mathsf{X}_{12(\alpha+\delta)} \cap M_\theta = \emptyset$. Therefore, all conditions for Theorem 3.8 hold for the sparsified set Q output by LEAN, and it then follows that $\mathrm{rank}(\mathsf{H}_i(\mathsf{X})) = \mathrm{rank}(\mathsf{H}_i(R^{2\rho}(Q)) \to \mathsf{H}_i(R^{12\rho}(Q)))$, for any $i \geq 0$. ∎

We remark that a particular interesting feature of Algorithm LEANTOPO is that, we only need to set the parameter β to a universal constant $\frac{\pi}{5}$. All other parameters such as the angle and radius conditions for choosing β-good pairs and the saprsification radius are determined by this choice of the angle β. This makes LEANTOPO parameter-free; see also our experimental results in Section 4. At the same time, the above Theorem states that its output is guaranteed to be correct as the input set of samples P becomes sufficiently dense.

4 Experiments and discussion

We experimented with LEANTOPO primarily on curve and surface samples. We used thresholds for sparsification that are more aggressive than predicted by our analysis. For example, our analysis predicts that for $\beta = \frac{\pi}{5}$, the constant $c_\beta = \frac{1}{3}\tan\frac{\beta}{2} \approx 0.11$, but we kept it at 0.5. We kept the same thresholds for all models to ensure that we don't fine tune it for different input. The sparsification ratio $\frac{r_q}{d(q,L_\beta)}$ is kept at 0.5, and the r for computing the complex R^r is kept at 0.65 in all cases. Table 1 below shows the details. The rank of

H_1 homology is computed correctly by our algorithm for all these data. The sparsified points are shown in Figure 1.

Name	input #points	output #points	c_β	sparsification ratio	r for R^r	$\mathrm{r}ank\, H_1$
MOTHERCHILD	126500	5267	0.5	0.5	0.7	8
BOTIJO	101529	7600	0.5	0.5	0.7	10
KITTEN	134448	1914	0.5	0.5	0.7	2
CURVEHELIX	1000	235	0.5	0.5	0.7	1

Table 1: Experiments on a curve and three surface samples.

Extensions. One obvious question that remains open is how to extend the scope of our sparsification strategy to larger class of input, such as noisy data samples and/or samples from compact spaces rather than manifolds.

Noise: We observe that, for Hausdorff noise, where samples are assumed to lie within a small offset of the manifold, our method can be applied. However, a parameter giving the extent of this Hausdorff noise needs to be supplied. With this parameter, one can estimate the normals reliably from the noisy but dense sample. The step where we compute the lean set, requires an empty ball test which also needs this parameter because otherwise noise can collaborate to provide a false impression that some spurious manifolds have been sampled. Given the ambiguity that a noisy sample can be dense for two topologically different spaces, it may be impossible to avoid a parameter that eliminates different such possibilities. Nevertheless, our method would free the user from specifying a threshold for building the complexes.

In an experiment, we added artificial noise on the three surface samples as shown in Figure 2 to test robustness of our algorithm. We added a uniform displacement to each sample point along the normal direction. The displacement ranged from -0.5% to 0.5% times the diameter of the model. We modified our algorithm to ignore all leanset points formed by two points closer than a threshold which is picked as a multiple of the diameter of the model. Other thresholds were kept the same as in the previous experiment. Results in Table 2 show that the algorithm can tolerate noise in case there is a known upper limit on the noise level.

The more general noise model which allows outliers would also be worthwhile to investigate. One may explore the 'distance to measure' technique proposed in [8] for this case. But, it is not clear how to adapt the entire development in this paper to this setting. One possibility is to eliminate all outliers first to make the noise only Hausdorff, and then apply the technique for Hausdorff noise as alluded in the previous paragraph. This will certainly require more parameters to be supplied by the user.

Compact sets: The case for compact sets is perhaps more challenging. The normal spaces are not well

Threshold (multiple of noise scale)	0	1	2	3	4	5	6	7	8	9	10	11	12	13
MOTHERCHILD	18196	1636	37	8	8	8	8	8	8	8	8	8	7	7
BOTIJO	14565	14580	1462	10	10	10	10	10	10	10	10	8	8	8
KITTEN	20506	20572	1314	2	2	2	2	2	2	2	2	2	2	2

Table 2: Experiments on 3 surfaces with artificial noise. The table shows resulting H_1 of each model under different threshold. Experiments show that the influence of noise is removed when we pick threshold greater than or equal to 3 times of the noise scale. The threshold might introduce problem when it is too large.

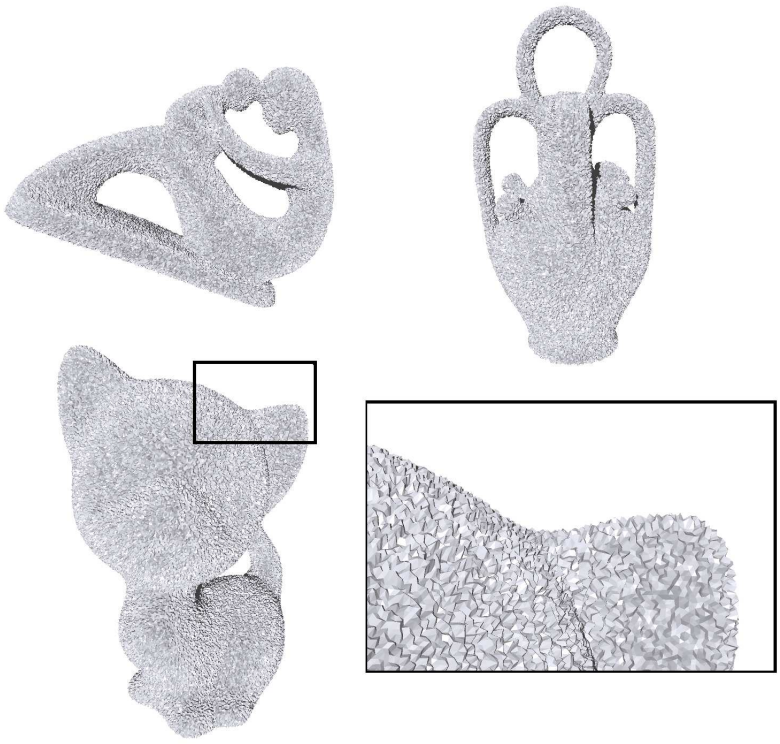


Figure 2: Noisy samples. Meshes are created only for rendering.

defined everywhere for such spaces. Thus, we need to devise a different strategy to compute the lean set. The theory of compacts developed in the context of topology inference in [7] may be useful here. Computing the lean sets efficiently in high dimensions for compact spaces remains a formidable open problem.

Acknowledgment

This work was partially supported by the NSF grants CCF-1064416, CCF-1116258, CCF 1318595, and CCF 1319406.

References

[1] P. K. Agarwal. Range searching. Chapter 36, *Handbook Discr. Comput. Geom.*, J. E. Goodman, J. O'Rourke (eds.), Chapman & Hall/CRC, Boca Raton, Florida, 2004.

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

[3] N. Amenta, M. Bern, and D. Eppstein. The crust and β-skele ton: combinatorial curve reconstruction. *Graphical Models and Image Process.* **60** (1998), 125–135.

[4] N. Amenta, S. Choi, and R. K. Kolluri. The power crust, union of balls, and the medial axis transform. *Comput. Geom. Theory Appl.* **19** (2001) 127–153.

[5] J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. Tech Report Inria-00440337, version 2, (2011), http://hal.inria.fr/inria-00440337.

[6] K. Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fund. Math* **35**, (1948) 217-234.

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

[8] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for measures based on distance functions. *Found. Comput. Math.*, Springer Verlag (Germany), **11** (2011), pp.733-751.

[9] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Comput. Geom: Theory and Applications* **40** (2008), 156–170.

[10] F. Chazal and A. Lieutier. The λ-medial axis. *Graph. Models* **67** (2005), 304–331.

[11] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discrete Comput. Geom.***37** (2007), 601–607.

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

[13] S.-W. Cheng, J. Jin and M-K. Lau. A fast and simple surface reconstruction algorithm. *Proc. 28th Ann. Sympos. Comput. Geom.* (2012), 69–78.

[14] K. Clarkson. Building triangulations using ε-nets. *Proc. 38th Annu. Sympos. Theory Comput.* (20016), 316–335.

[15] T. K. Dey, F. Fan, and Y. Wang. Graph induced complex on point data. *Comput. Geom.: Theory and Applications* **48** (2015), 575–588.

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

[17] S. Funke and E. A. Ramos. Smooth-surface reconstruction in near-linear time. *Proc. 13th Annu. ACM-SIAM Sympos. Discrete Algorithms* (2002), 781–790.

[18] K. Grove. Critical point theory for distance functions. *Proc. Symposia in Pure Mathematics* **54**, American Mathematical Society, Providence, RI, 1993.

[19] J. Giesen and M. John. The flow complex: A data structure for geometric modeling. *Comput. Geom.: Theory and Applications* **39** (2008), 178–190.

[20] A. Lieutier. Any open bounded subset of $\mathbb{R}^n$ has the same homotopy type as its medial axis. *J. Comput. Aided Design* **36** (2004), 1029–1046.

[21] James R. Munkres. Elements of Algebraic Topology. Addison–Wesley Publishing Company, Menlo Park, 1984.

[22] D. Sheehy. Linear-Size Approximations to the Vietoris-Rips Filtration. *Discrete Comput. Geom.* **49** (2013), 778–796.

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

[24] E. H. Spanier. Algebraic topology. *Springer-Verlag*.

A Missing Proofs

Proving that (p,q) is β-good for Proposition 2.7. We know that $\angle(\mathbf{N}_s, st) \le \pi/2 - \theta$ which implies that $d(s,t) \ge 2d(s,m)\sin\theta \ge 2\mathrm{lfs}(s)\sin\theta$. Consider the triangle pst. By triangle inequality, $d(p,t) \ge d(s,t) - d(s,p) \ge (2\sin\theta - \varepsilon)\mathrm{lfs}(s)$. The angle $\angle pts$ is at most

$$\arcsin\frac{d(p,s)}{d(p,t)} \le \arcsin\frac{\varepsilon\mathrm{lfs}(s)}{(2\sin\theta-\varepsilon)\mathrm{lfs}(s)} \le \frac{4}{3}\cdot\frac{\varepsilon}{2\sin\theta-\varepsilon}. \tag{5}$$

The last inequality follows from that $\arcsin(x) \le cx$ for $x \le \frac{\sqrt{c^2-1}}{c}$. In our case, choose $c = \frac{4}{3}$. Since $\sqrt{\varepsilon} \le \frac{\sin\theta}{2\sqrt{2}} \le \frac{1}{2}$, we have that

$$\frac{\varepsilon}{2\sin\theta-\varepsilon} \le \frac{\varepsilon}{4\sqrt{\varepsilon}-\varepsilon} = \frac{\sqrt{\varepsilon}}{4-\sqrt{\varepsilon}} \le \frac{1}{7} \le \frac{\sqrt{c^2-1}}{c}.$$

Now assume without loss of generality that $\mathrm{lfs}(s) \ge \mathrm{lfs}(t)$. Then,

$$d(p,q) \ge d(s,t) - d(p,s) - d(q,t) \ge d(s,t) - 2\varepsilon\mathrm{lfs}(s) \ge 2(\sin\theta-\varepsilon)\mathrm{lfs}(s).$$

Recall that $d(t,p) \ge (2\sin\theta-\varepsilon)\mathrm{lfs}(s)$. Considering the triangle tpq, we have

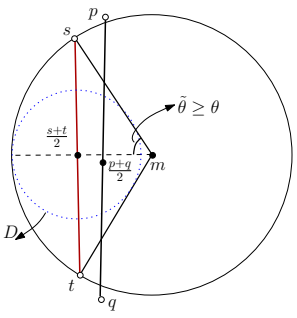
$$\angle tpq \le \arcsin\frac{d(q,t)}{d(p,t)} \le \arcsin\frac{\varepsilon\mathrm{lfs}(t)}{2(\sin\theta-\varepsilon)\mathrm{lfs}(s)} \le \arcsin\frac{\varepsilon\mathrm{lfs}(s)}{2(\sin\theta-\varepsilon)\mathrm{lfs}(s)} \le \frac{4}{3}\cdot\frac{\varepsilon}{2(\sin\theta-\varepsilon)}, \tag{6}$$

where the last inequality follows from a similar argument used for (5).

We know that, $\angle(\mathbf{N}_p, \mathbf{N}_s) \le \varepsilon$, $\angle(\tilde{\mathbf{N}}_p, \mathbf{N}_p) \le \nu_\varepsilon$, and $\angle(pq, st) \le \angle pts + \angle tpq$. Combining these with (5), (6) and the assumption that $\sqrt{\varepsilon} \le \frac{1}{2\sqrt{2}}\sin\theta (\le \frac{1}{2})$, we have that

$$\begin{aligned}\angle(pq, \tilde{\mathbf{N}}_p) &\le \angle(pq,st) + \angle(st,\mathbf{N}_s) + \angle(\mathbf{N}_s,\mathbf{N}_p) + \angle(\mathbf{N}_p,\tilde{\mathbf{N}}_p) \le \frac{8}{3}\cdot\frac{\varepsilon}{2\sin\theta-2\varepsilon} + \frac{\pi}{2} - \theta + \varepsilon + \nu_\varepsilon \\ &\le \frac{8}{3}\cdot\frac{\sqrt{\varepsilon}}{4\sqrt{2}-2\sqrt{\varepsilon}} + \frac{\sqrt{\varepsilon}}{2} + \frac{\pi}{2} - \theta + \nu_\varepsilon \le \frac{\pi}{2} - \theta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon\end{aligned}$$

Similar bound holds for $\angle(pq, \tilde{\mathbf{N}}_q)$. It follows that the pair (p,q) satisfies the first condition of being β-good, as long as $\frac{\pi}{2} - \theta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon \le \frac{\pi}{2} - \beta$. This is guaranteed by requiring $\theta \ge \beta + \frac{3}{2}\sqrt{\varepsilon} + \nu_\varepsilon$ (as specified in the proposition).



Next, we argue that (p,q) also satisfies the second condition of being β-good. To do so, let $\tilde{\theta} = \frac{1}{2}\angle smt$ be half of the angle spanned by sm and tm. Note that by the definition of θ-medial axis M_θ, we have that $\tilde{\theta} \ge \theta$. See the right figure for an illustration. First, observe that the ball $D = B(\frac{s+t}{2}, r)$ with $r = d(s,m)\cdot(1-\cos\tilde{\theta})$ does not intersect X, since this ball is contained inside the medial ball $B(m, d(s,m))$. The midpoint $\frac{p+q}{2}$ of pq is at most $\varepsilon\mathrm{lfs}(s) \le \varepsilon d(s,m)$ distance away from $\frac{s+t}{2}$ because both p and q are at most $\varepsilon\mathrm{lfs}(s)$ away from s and t (assuming w.o.l.g $\mathrm{lfs}(s) \ge \mathrm{lfs}(t)$). This means that the ball $D' = B(\frac{p+q}{2}, r')$ centering at the midpoint of pq and with radius $r' = d(s,m)\cdot(1-\cos\tilde{\theta}-\varepsilon)$ is contained in the ball D and thus does not have any point of X and hence P inside.

On the other hand, note that

$$d(p,q) \le d(s,t) + 2\varepsilon\mathrm{lfs}(s) \le 2d(s,m)\sin\tilde{\theta} + 2\varepsilon d(s,m) = 2d(s,m)(\sin\tilde{\theta}+\varepsilon).$$

Thus, the second condition for p, q being a good pair is satisfied as long as

$$c_\beta \leq \frac{1-\cos\tilde{\theta}-\varepsilon}{2(\sin\tilde{\theta}+\varepsilon)} \leq \frac{r'}{d(p,q)}.$$

Consider the function $f(x) = \frac{1-\cos x-\varepsilon}{\sin x+\varepsilon}$, its derivative $f'(x)$ is greater than 0 for $x \in [0, \pi/2]$. Indeed,

$$f'(x) = \frac{\sin x \cdot (\sin x+\varepsilon) - (1-\cos x-\varepsilon)\cdot\cos x}{(\sin x+\varepsilon)^2} = \frac{1-\cos x+\varepsilon\sin x+\varepsilon\cos x}{(\sin x+\varepsilon)^2} \geq 0.$$

Hence $f(x)$ is an increasing function, and $f(\tilde{\theta}) \geq f(\theta)$ since $\tilde{\theta} \geq \theta$. In other words, the second condition for (p, q) being a good pair is satisfied as long as $c_\beta \leq \frac{1-\cos\theta-\varepsilon}{2(\sin\theta+\varepsilon)}$. To further simplify it, note that using $\varepsilon \leq \frac{1}{8}\sin^2\theta$, one can show that $\frac{4\varepsilon}{\sin\theta} \leq \tan\frac{\theta}{2}$. Combining this with $\frac{1-\cos\theta}{\sin\theta} = \tan\frac{\theta}{2}$, we then have

$$\frac{1-\cos\theta-\varepsilon}{2(\sin\theta+\varepsilon)} \geq \frac{1-\cos\theta-\varepsilon}{\frac{9}{4}\sin\theta} = \frac{4}{9}\tan\frac{\theta}{2} - \frac{4\varepsilon}{9\sin\theta} \geq \frac{4}{9}\tan\frac{\theta}{2} - \frac{1}{9}\tan\frac{\theta}{2} = \frac{1}{3}\tan\frac{\theta}{2} \geq \frac{1}{3}\tan\frac{\beta}{2}.$$

Hence as $c_\beta \leq \frac{1}{3}\tan\frac{\beta}{2}$, the ball $B(\frac{p+q}{2}, c_\beta d(p,q))$ is contained in D' and thus contains no point in P. Therefore, the pair (p, q) is β-good and its midpoint is in L_β.

Proof of Theorem 2.12. Let x be any point in X to which p is the nearest sample point in P. Then, $d(x,p) \leq \varepsilon\mathrm{lfs}(x) \leq \varepsilon'\mathrm{lfs}(p)$ where $\varepsilon' = \frac{\varepsilon}{1-\varepsilon}$. If p is retained in Q, $d(x,Q) \leq \varepsilon\mathrm{lfs}(x) \leq \varepsilon'\mathrm{lfs}(p) \leq \frac{\varepsilon'}{c_2}d(p,L_\beta) \leq \frac{6\rho}{5}d(p,L_\beta)$ for sufficiently small $\varepsilon > 0$, where c_2 is the constant from Proposition 2.11. Now consider the case when p is deleted while processing another point, say $q \in P$. By the sparsification procedure in lines 5–9, $d(q,L_\beta) \geq d(p,L_\beta)$ and q will remain in Q since we process points in non-decreasing order of their distances to L_β. Using Proposition 2.11, we then have:

$$\begin{aligned} d(x,q) &\leq d(x,p) + d(p,q) \leq \varepsilon\mathrm{lfs}(x) + \rho d(q,L_\beta) \leq \varepsilon'\mathrm{lfs}(p) + \rho d(q,L_\beta) \\ &\leq \frac{\varepsilon'}{c_2}d(p,L_\beta) + \rho d(q,L_\beta) \leq (\frac{\varepsilon'}{c_2}+\rho)d(q,L_\beta) \leq \frac{6\rho}{5}d(q,L_\beta). \end{aligned}$$

The last inequality holds when ε is sufficiently small (in which case the estimation error ν_ε in the normal space is also small). Therefore,

$$d(x,q) \leq \frac{6\rho}{5}\mathrm{lnfs}_\beta(q) \tag{7}$$

Now applying Remark 3.6, Q is also $(\frac{4}{3}\rho)$-dense because $\frac{\frac{6\rho}{5}}{1-\frac{6\rho}{5}} \leq \frac{4}{3}\rho$ for $\rho \leq \frac{1}{12}$.

The fact that Q is ρ-sparse w.r.t. lnfs_β follows easily from the sparsification procedure.

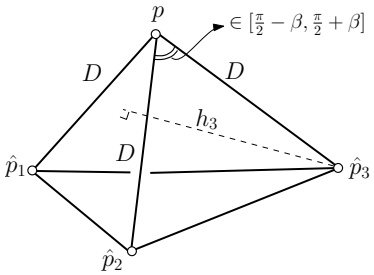
B Estimation of Normal/Tangent Space

Here, we provide the justification for the claimed bound of $O(\varepsilon)$ on the tangent space estimation(and thus the normal space) of the hidden manifold X at a sample point $p \in P$. For completion, we restate the procedure described in section 2.2 for estimating the tangent space T_p. Set $\beta = \frac{\pi}{5}$ for the calculations to follow. Let s denote the intrinsic dimension of the manifold X, which we assume is known a-priori. Let $p_1 \in P$ be the nearest neighbor of p in $P \setminus \{p\}$. Suppose we have already obtained points $\sigma_i = \{p, p_1, \ldots, p_i\}$ with $i < s$. Let $\mathrm{aff}(\sigma_i)$ denote the affine hull of the points in σ_i. Next, we choose $p_{i+1} \in P$ that is closest to p among

all points forming an angle within the range $[\frac{\pi}{2}-\beta, \frac{\pi}{2}]$ with $\mathrm{aff}(\sigma_i)$. We add p_{i+1} to the set and obtain $\sigma_{i+1} = \{p, p_1, \ldots, p_i, p_{i+1}\}$. This process is repeated until $i = s$, at which point we have obtained $s+1$ points $\sigma_s = \{p, p_1, \ldots, p_s\}$. We use $\mathrm{aff}(\sigma_s)$ to approximate the tangent space T_p. We now show that the simplex σ_i is "fat". In particular, we will leverage a result (Corollary 2.6) of [5] to bound the angle between the true tangent space T_p and approximate tangent space $\mathrm{aff}(\sigma_i)$.

More specifically, we first modify the simplex σ_i to another one $\widehat{\sigma}_i$ as follows. Let D denote the longest length of any edge incident to p in σ_i. Later we will prove that $D = O(\frac{\varepsilon \mathrm{lfs}(p)}{\sin\beta})$. Now, we extend each edge pp_j along the same line segment but to $p\hat{p}_j$ such that $\|p\hat{p}_j\| = D$. The resulting simplex spanned by $\{p, \hat{p}_1, \ldots, \hat{p}_i\}$ is denoted by $\widehat{\sigma}_i$. By construction, $\mathrm{aff}(\sigma_i) = \mathrm{aff}(\widehat{\sigma}_i)$. Hence, we only need to bound the angle $\angle(T_p, \mathrm{aff}(\widehat{\sigma}_i))$. Corollary 2.6 of [5] states that $\sin\angle(\mathrm{aff}(\widehat{\sigma}_i), T_p) \leq \frac{L^{i+2}}{\mathrm{Vol}(\widehat{\sigma}_i)\cdot S\cdot \mathrm{lfs}(p)}$, where L and S are the longest and shortest edge length of $\widehat{\sigma}_i$ respectively; while $\mathrm{Vol}(\widehat{\sigma}_i)$ stands for the volume of the simplex $\widehat{\sigma}_i$. To use this result, we bound the terms L, S, and $\mathrm{Vol}(\widehat{\sigma}_i)$.

See the figure on right for an illustration. First, we bound the angle between any two $p\hat{p}_\ell$ and $p\hat{p}_j$, for $\ell, j \in [1, i]$. Assume w.o.l.g. that $j > \ell$. By construction, $p\hat{p}_j$ forms an angle α such that $\alpha \in [\frac{\pi}{2}-\beta, \frac{\pi}{2}]$ with $\mathrm{aff}(\sigma_{j-1})$. It follows that $\alpha \leq \angle(p\hat{p}_\ell, p\hat{p}_j) \leq \pi - \alpha$, that is, $\angle(p\hat{p}_\ell, p\hat{p}_j) \in [\frac{\pi}{2}-\beta, \frac{\pi}{2}+\beta]$. Therefore, the edge length $d(\hat{p}_\ell, \hat{p}_j)$ satisfies



$$d(\hat{p}_\ell, \hat{p}_j) = 2D\cdot\sin\frac{1}{2}\angle(p\hat{p}_\ell, p\hat{p}_j) \in [2D\cdot\sin(\frac{\pi}{4}-\frac{\beta}{2}), 2D\cdot\sin(\frac{\pi}{4}+\frac{\beta}{2})].$$

Therefore the longest edge length L in simplex $\widehat{\sigma}_i$ is at most $L \leq 2D\cdot\sin(\frac{\pi}{4}+\frac{\beta}{2})$, while the smallest edge length S in simplex $\widehat{\sigma}_i$ is at least $S \geq \min\{D, 2D\cdot\sin(\frac{\pi}{4}-\frac{\beta}{2})\}$.

Next, we bound the volume $\mathrm{Vol}(\widehat{\sigma}_i)$ of $\widehat{\sigma}_i$, which we do inductively. We claim that $\mathrm{Vol}(\widehat{\sigma}_i) \geq \frac{D\cdot(D\cdot\cos\beta)^{i-1}}{i!}$. This claim holds when $i = 1$ in which case $\mathrm{Vol}(\widehat{\sigma}_1) = d(p, \hat{p}_1) = D$. Assume it holds for $i-1$. Then, we have that $\mathrm{Vol}(\widehat{\sigma}_i) = \frac{1}{i}d(\hat{p}_i, \mathrm{aff}(\widehat{\sigma}_{i-1}))\cdot\mathrm{Vol}(\widehat{\sigma}_{i-1})$, where $h_i = d(\hat{p}_i, \mathrm{aff}(\widehat{\sigma}_{i-1}))$ is the height of the simplex $\widehat{\sigma}_i$ using $\widehat{\sigma}_{i-1}$ as the base facet. On the other hand, by construction $\angle(p\hat{p}_i, \mathrm{aff}(\widehat{\sigma}_{i-1}) \geq \frac{\pi}{2}-\beta$, which gives

$$h_i = d(p, \hat{p}_i)\cdot\sin\angle(p\hat{p}_i, \mathrm{aff}(\widehat{\sigma}_{i-1})) \geq D\cdot\cos\beta.$$

It follows that $\mathrm{Vol}(\widehat{\sigma}_i) \geq \frac{D}{i}\cdot\cos\beta\cdot\mathrm{Vol}(\widehat{\sigma}_{i-1}) \geq \frac{D\cdot(D\cdot\cos\beta)^{i-1}}{i!}$, which then proves the claim inductively.

Now we derive an upper bound on D. Inductively, assume that for $1 \leq i \leq s$,

$$D \leq 13\varepsilon\mathrm{lfs}(p) \text{ and } \theta_i = \angle(\mathrm{aff}(\sigma_i), T_p) \leq \arcsin\left(\frac{i!2^{i+2}D\sin^{i+2}(\frac{\pi}{4}+\frac{\beta}{2})}{\cos^{i-1}\beta\sin(\frac{\pi}{4}-\frac{\beta}{2})\mathrm{lfs}(p)}\right).$$

For $i = 1$ and sufficiently small ε, it is true because the nearest point p_1 to p satisfies $d(p, p_1) \leq 3\varepsilon\mathrm{lfs}(p)$ and also $\sin\angle(pp_1, T_p) \leq \frac{3}{2}\varepsilon$ (this follows easily from the ε-dense sampling condition, see e.g. Corollary 3.1 and Lemma 3.4 [16]) For induction consider the time when we choose p_i. Consider the projection $\tilde{\sigma}_{i-1}$ of σ_{i-1} onto T_p and the $(i-1)$-dimensional affine subspace $\mathrm{aff}(\tilde{\sigma}_{i-1})$ of T_p containing this projection. By our inductive hypothesis, $\angle(\mathrm{aff}\sigma_{i-1}, \mathrm{aff}\tilde{\sigma}_{i-1}) \leq \theta_{i-1}$. Let F be the subspace of T_p orthogonal to $\mathrm{aff}\tilde{\sigma}_{i-1}$ and let $x \in F$ be such that $d(x, p) = 10\varepsilon\mathrm{lfs}(p)$. The closest point $\tilde{x} \in \mathsf{X}$ of x to X has $d(x, \tilde{x}) = O(\varepsilon^2\mathrm{lfs}(p))$. Therefore, we can assume that

$$9\varepsilon\mathrm{lfs}(p) \leq d(p, \tilde{x}) \leq 11\varepsilon\mathrm{lfs}(p)$$

when $\varepsilon > 0$ is sufficiently small. There is a sample point $p' \in P$ with $d(\tilde{x}, p') \leq \varepsilon\mathrm{lfs}(\tilde{x})$. This means that the angle $\angle p\tilde{x}, pp'$ is at most $\arcsin(\frac{\varepsilon\mathrm{lfs}(x)}{9\varepsilon\mathrm{lfs}(p)}) = \arcsin\frac{1}{8}$ when ε is sufficiently small. It follows that $\angle(pp', \mathrm{aff}(\sigma_{i-1})) \geq \frac{\pi}{2} - \arcsin\frac{1}{8} - \theta_{i-1}$. One can make θ_{i-1} arbitrarily small by choosing ε sufficiently

small. Therefore, if $\beta = \frac{\pi}{5}$ and ε is small enough, we have $\angle pp', \mathrm{aff}(\sigma_{i-1}) \in [\frac{\pi}{2} - \beta, \frac{\pi}{2}]$. Since p_i is chosen with the smallest distance from p satisfying the above angle condition, we have, for small enough ε,

$$d(p, p_i) \leq d(p, p') \leq d(p, \tilde{x}) + d(\tilde{x}, p') \leq 11\varepsilon \mathrm{lfs}(p) + \varepsilon \mathrm{lfs}(x) \leq 13\varepsilon \mathrm{lfs}(p).$$

Since D cannot be larger than the maximum between older D from stage $i-1$ and $d(p, p_i)$, one has $D \leq 13\varepsilon \mathrm{lfs}(p)$. Combining all these with Corollary 2.6 of [5], we obtain that $\sin \angle(\mathrm{aff}(\widehat{\sigma}_i), T_p) = \sin \theta_i$ as claimed.

Evaluating $\sin \theta_i$ we obtain $\sin \theta_i = O(\frac{D}{\mathrm{lfs}(p)}) = O(\varepsilon)$ for all $i \in [1, s]$ where the big-O notation hides constants depending exponentially on the intrinsic dimension s and $\cos \beta$. In other words, the angle ν_ε between the approximate tangent space and the true tangent space (thus between the approximate normal space and the true normal space) at any sample point is bounded by $O(\varepsilon)$, where the big-O notations hides constant depending on the angle β and intrinsic dimension s of the manifold X.