Local Homology Transfer and Stratification Learning

Paul Bendich
bendich@math.duke.edu
Department of Mathematics, Duke University and IST Austria

Bei Wang
beiwang@sci.utah.edu
SCI Institute, University of Utah

Sayan Mukherjee
sayan@stat.duke.edu
Departments of Statistical Science, Mathematics, and Computer Science, Duke University

Abstract

A stratified space is a collection of manifolds of different dimensions which fit together uniformly inside some larger space. The objective of this paper is to show that data sampled from such a space can be clustered by strata. We first define a multi-scale notion of stratified spaces, providing a stratification at different scales which are indexed by a radius parameter. We then use methods derived from kernel and cokernel persistent homology to cluster the data points into different strata. We prove a correctness guarantee for this clustering method under certain topological conditions. We then provide a probabilistic guarantee for the clustering for the point sample setting – we provide bounds on the minimum number of sample points required to state with high probability which points belong to the same strata. Finally, we give an explicit algorithm for the clustering.

1 Introduction

Manifold learning is a basic problem in geometry, topology, and statistical inference that has received a great deal of recent attention. One formulation of the problem is: given a point cloud of data sampled from a manifold in an ambient space $\mathbb{R}^N$, infer the dimension and structure of the underlying manifold. A limitation of this problem statement is that it does not apply to sets that are not manifolds. For example, we may consider the more general class of stratified spaces that can be decomposed into strata – manifolds of varying dimension each of which fit together in some uniform way inside the higher dimensional space.

In this paper, we study the following problem in stratification learning: given a point cloud sampled from a stratified space, how do we cluster points that belong to the same stratum together while keeping points in different stratum apart. Intuitively, a reasonable strategy would be to place two points in the same piece of stratum if they "look the same locally" – they have identical neighborhoods within the larger space at some very small scale. However, the notion of "local" becomes unclear in the context of the uncertainty induced from sampling, since everything becomes noisy at small scales. In response, we introduce a radius or scale parameter r and define a notion of local equivalence at each scale r.

We will use tools derived from algebraic topology. In particular, we define local equivalence between points via maps that transfer information carried by local homology groups, and we then use persistent homology [10] methods to infer the properties of these maps.

Prior Work. Consistency in manifold learning has often been recast as homology inference – as the number of points in a point cloud goes to infinity, the homology inferred from the point cloud converges to the true homology of the underlying space. Results of this nature have been given for manifolds [17, 18] and a large class of compact subsets of Euclidean space [5]. Stronger results in homology inference for closed subsets of a metric space are given in [7].

Geometric approaches to stratification learning have also been developed. These include inference of a mixture of linear subspaces [15], mixture models for general stratified spaces [12], and generalized Principal Component Analysis (GPCA) [20] which was developed for dimension reduction for mixtures of manifolds.

The study of stratified spaces has long been a focus of pure mathematics; see, for example, [11, 21]. The problem of inference for the local homology groups of a sampled stratified space in a deterministic setting has been addressed in [1].

Contributions. In this paper we propose an approach to stratification learning based on local homology inference. The results in this paper are:

(1) A topological definition of two points belonging to the same strata by assessing the multi-scale local structure of the points through a local homology transfer map. (Definition 3.1);

(2) Topological conditions on the point sample under which this characterization holds (Theorem 3.2);

(3) Finite sample bounds for the minimum number of points required in the sample to state with high probability which points belong to the same strata (Theorem 4.1);

(4) An algorithm that computes which points belong to the same strata (Section 5).

2 Background

We first describe general persistence modules [4], focusing mainly on those that arise from maps between absolute or relative homology groups induced by inclusions of topological spaces or pairs of such spaces. We then discuss stratifications and their connection to the local homology groups of a topological space. Basics on homology itself are assumed; for a readable background, see [16] or [13], or [10] for a more computationally oriented treatment.

Persistence modules. For simplicity, our treatment of persistence modules adapted from [4] is restricted to $\mathbb{Z}/2\mathbb{Z}$-vector spaces. Let A be some subset of $\mathbb{R}$. A *persistence module* $\mathcal{F}_A$ is a collection $\{F_\alpha\}_{\alpha\in A}$ of $\mathbb{Z}/2\mathbb{Z}$-vector spaces, together with a family $\{f_\alpha^\beta : F_\alpha \to F_\beta\}_{\alpha\leq\beta\in A}$ of linear maps such that $\alpha \leq \beta \leq \gamma$ implies $f_\alpha^\gamma = f_\beta^\gamma \circ f_\alpha^\beta$. We will assume that the index set A is either $\mathbb{R}$ or $\mathbb{R}_{\geq 0}$ and not explicitly state indices unless necessary.

A real number α is said to be a *regular value* of the persistence module $\mathcal{F}$ if there exists some $\varepsilon > 0$ such that

the map $f_{\alpha-\delta}^{\alpha+\delta}$ is an isomorphism for each $\delta < \varepsilon$. Otherwise we say that α is a *critical value* of the persistence module; if $A = \mathbb{R}_{\geq 0}$, then $\alpha = 0$ will always be considered to be a critical value. We say that $\mathcal{F}$ is *tame* if it has a finite number of critical values and if all the vector spaces F_α are of finite rank. Any tame $\mathbb{R}_{\geq 0}$-module $\mathcal{F}$ must have a smallest non-zero critical value $\rho(\mathcal{F})$; we call this number the *feature size* of the persistence module. Assume $\mathcal{F}$ is tame and so we have a finite ordered list of critical values $0 = c_0 < c_1 < \ldots < c_m$. We choose regular values $\{a_i\}_{i=0}^m$ such that $c_{i-1} < a_{i-1} < c_i < a_i$ for all $1 \leq i \leq m$, and we adopt the shorthand notation $F_i \equiv F_{a_i}$ and $f_i^j : F_i \to F_j$, for $0 \leq i \leq j \leq m$. A vector $v \in \mathcal{F}_i$ is said to be *born* at level i if $v \notin \operatorname{im} f_{i-1}^i$, and such a vector *dies* at level j if $f_i^j(v) \in \operatorname{im} f_{i-1}^j$ but $f_i^{j-1}(v) \notin \operatorname{im} f_{i-1}^{j-1}$. This is illustrated in Figure 1 (a). We then define $P^{i,j}$ to be the vector space of vectors that are born at level i and then subsequently die at level j, and let $\beta^{i,j}$ denote its rank.

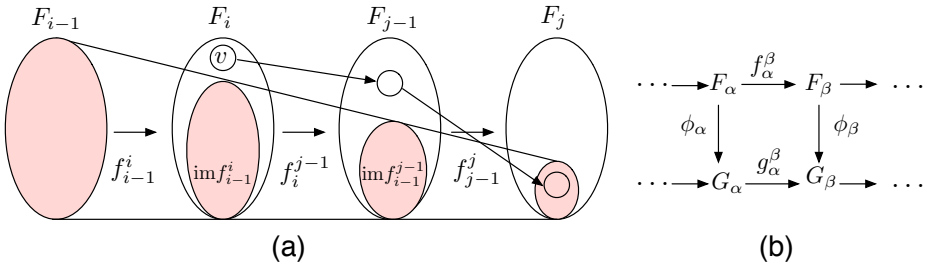

Figure 1: (a) The vector v is born at level i and then it dies at level j. (b) Commuting diagrams for (co)kernel modules.

Persistence diagrams. The information contained within a tame module $\mathcal{F}$ can be compactly represented by a *persistence diagram*, $\operatorname{Dgm}(\mathcal{F})$, which is a multi-set of points in the extended plane. It contains $\beta^{i,j}$ copies of the points (c_i, c_j), as well as infinitely many copies of each point along the major diagonal $y = x$. In Figure 2 (a) the persistence diagrams for a curve and a point cloud sampled from it are displayed, see below for a full explanation of this figure.

For any two points $u = (x, y)$ and $u' = (x', y')$ in the extended plane, we define $||u - u'||_\infty = \max\{|x - x'|, |y - y'|\}$. We define the *bottleneck distance* between any two persistence diagrams D and D' to be:

$$d_B(D, D') = \inf_{\Gamma: D \to D'} \sup_{u \in D} ||u - \Gamma(u)||_\infty,$$

where Γ ranges over all bijections from D to D'. Under certain conditions described in the full version, persistence diagrams are stable under this distance.

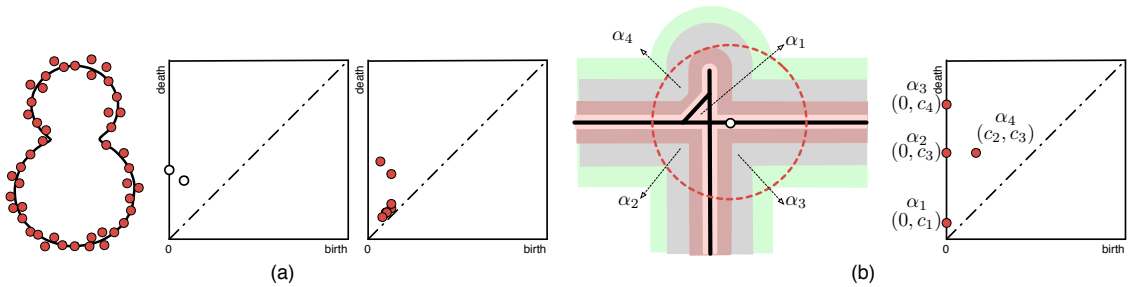

Figure 2: (a) Illustration of a point cloud and its persistence diagram: left, $\mathbb{X}$ is the curve embedded as shown in the plane and $\mathbb{U}$ is the point cloud; middle, the persistence diagram $\operatorname{Dgm}_1(d_{\mathbb{X}})$; right, the persistence diagram $\operatorname{Dgm}_1(d_{\mathbb{U}})$. The diagrams are generated by thickening $\mathbb{X}$ (or $\mathbb{U}$) while tracking the birth and death of homology classes. (b) Illustration of relative homology and its persistence diagram: left, the space $\mathbb{X}$ is in solid line and the closed ball B has dotted boundary; right, the persistence diagram for the module $\{\mathsf{H}_1(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}$. Here, α goes through four non-zero critical values $c_1 < c_2 < c_3 < c_4$ that correspond to the four colored level sets, where the points in the persistence diagram correspond to the birth and death of the four relative homology classes respectively. In particular, α_4 is created when the level set at value c_2 touches B.

(Co)Kernel modules. Suppose now that we have two persistence modules $\mathcal{F}$ and $\mathcal{G}$ along with a family of maps $\{\phi_\alpha : F_\alpha \to G_\alpha\}$ which commute with the module maps – for every pair $\alpha \leq \beta$, we have $g_\alpha^\beta \circ \phi_\alpha = \phi_\beta \circ f_\alpha^\beta$. In other words, every square commutes in the diagram shown in Figure 1 (b). Then, for each pair of real numbers

$\alpha \leq \beta$, the restriction of f_α^β to $\ker \phi_\alpha$ maps into $\ker \phi_\beta$, giving rise to a new kernel persistence module, with persistence diagram denoted by $\mathrm{Dgm}(\ker \phi)$. Similarly, we obtain a cokernel persistence module, with diagram $\mathrm{Dgm}(\mathrm{cok}\, \phi)$.

Homology and distance functions. Consider a family of topological spaces $\{\mathbb{X}_\alpha\}$, along with inclusions $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ for all $\alpha \leq \beta$. The inclusions induce maps $\mathsf{H}_j(\mathbb{X}_\alpha) \to \mathsf{H}_j(\mathbb{X}_\beta)$, for each homological dimension $j \geq 0$, and hence we have persistence modules for each j. Defining $\mathsf{H}(\mathbb{X}_\alpha) = \bigoplus_j \mathsf{H}_j(\mathbb{X}_\alpha)$ and taking direct sums of maps in the obvious way, will also give one large direct-sum persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}$.

Given a compact topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$, we define $d_\mathbb{X}$ as the distance function which maps each point in the ambient space to the distance from its closest point in $\mathbb{X}$. We let $\mathbb{X}_\alpha$ denote the sublevel set $d_\mathbb{X}^{-1}[0, \alpha]$; each sublevel set should be thought of as a thickening of $\mathbb{X}$ within the ambient space. Increasing the thickening parameter produces a growing family of sublevel sets, giving rise to the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}_{\alpha \in \mathbb{R}_{\geq 0}}$; we denote the persistence diagram of this module by $\mathrm{Dgm}(d_\mathbb{X})$ and use $\mathrm{Dgm}_j(d_\mathbb{X})$ for the diagrams of the individual modules for each homological dimension j. In Figure 2 (a), we see an example of such an $\mathbb{X}$ embedded in the plane, along with the persistence diagram $\mathrm{Dgm}_1(d_\mathbb{X})$. We also have the persistence diagram $\mathrm{Dgm}_1(d_\mathbb{U})$, where $\mathbb{U}$ is a dense point sample of $\mathbb{X}$. Note that the two diagrams are quite close in bottleneck distance. Indeed, the difference between the two diagrams will always be upper-bounded by the Hausdorff distance between the space and its sample.

We can also have persistence modules of relative homology groups. For example, referring to the left of Figure 2 (b), we let $\mathbb{X}$ be the space drawn in solid lines and B the closed ball whose boundary is drawn as a dotted circle. By restricting $d_\mathbb{X}$ to B and also to ∂B, we produce pairs of sublevel sets $(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)$. Using the maps induced by the inclusions of pairs, we obtain the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}_{\alpha \in \mathbb{R}_{\geq 0}}$ of relative homology groups. The persistence diagram, for homological dimension 1, appears on Figure 2 (b) right. Here, α goes through four non-zero critical values $c_1 < c_2 < c_3 < c_4$ that correspond to the four level sets, where the points in the persistence diagrams (Figure 2 (b) right) correspond to the birth and death of the four relative homology classes respectively.

Stratified spaces. We assume that we have a topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$. A (purely) *d-dimensional stratification of* $\mathbb{X}$ is a decreasing sequence of closed subspaces $\mathbb{X} = \mathbb{X}_d \supseteq \mathbb{X}_{d-1} \supseteq \ldots \mathbb{X}_0 \supseteq \mathbb{X}_{-1} = \emptyset$, such that for each i, the i-dimensional *stratum* $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ is a (possibly empty) i-manifold. The connected components of $\mathbb{S}_i$ are called i-dimensional *pieces*. See Figure 3 (a) for an illustration.

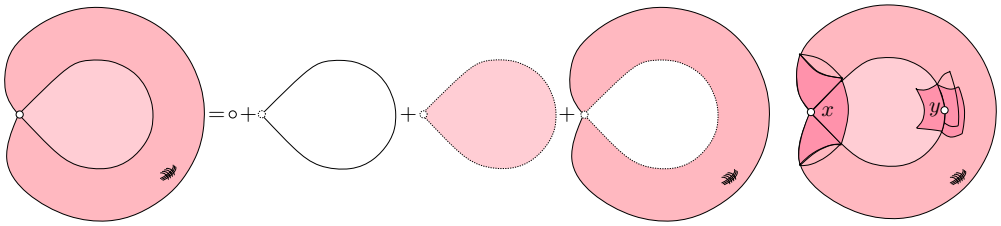

Figure 3: (a) The coarsest stratification of a pinched torus with a spanning disc stretched across the hole. (b) The space in (a) is a cs-space, where the x and y are respectively in the 0-stratum and the 1-stratum, their neighborhoods are highlighted.

One usually also imposes a requirement to ensure that the various pieces fit together uniformly. There are a number of different ways this can be done (see [14] for an extensive survey). For example, one might assume that for each $x \in \mathbb{S}_i$, there exists a small enough neighborhood $N(x) \subseteq \mathbb{X}$ and a $(d-i-1)$-dimensional stratified space L_x such that $N(x)$ is stratum-preserving homeomorphic to the product of an i-ball and the cone on L_x; one can then show that the space L_x depends only on the particular piece containing x. This definition, formally known as a *cs-space*, is illustrated in Figure 3 (b). Since the topology on $\mathbb{X}$ is that inherited from the ambient space, this neighborhood $N(x)$ will take the form $\mathbb{X} \cap B_r(x)$, where $B_r(x)$ is a small enough ball around x in the ambient space.

We note that the above definition requires all strata to be contained within the closure of the top-dimensional stratum. It is also possible, of course, to have spaces where this is not the case: for example, a two-dimensional

plane that has been punctured by a line. In this case, a slight adjustment to the above definitions can be made in order to impose similar notions of uniformity.

Local homology and homology stratifications. Recall ([16]) that the local homology groups of a space $\mathbb{X}$ at a point $x \in \mathbb{X}$ are the groups $\mathsf{H}_i(\mathbb{X}, \mathbb{X} - x)$ in each homological dimension i. If $\mathbb{X}$ happens to be a d-manifold, or if x is simply a point in the top-dimensional stratum of a d-dimensional stratification, then these groups are rank one in dimension d and trivial in all other dimensions. On the other hand, the local homology groups for lower-stratum points can be more interesting; for example if x is the crossing point in Figure 2 (b), then $\mathsf{H}_1(\mathbb{X}, \mathbb{X} - x)$ has rank three.

If x and y are close enough points in a particular piece of the same stratum, then there is a natural isomorphism between their local homology groups $\mathsf{H}(\mathbb{X}, \mathbb{X} - x) \cong \mathsf{H}(\mathbb{X}, \mathbb{X} - y)$, which can be understood in the following manner. Taking a small enough radius r and using excision, we see that the two local homology groups in question are in fact just $\mathsf{H}(\mathbb{X} \cap B_r(x), \mathbb{X} \cap \partial B_r(x))$ and $\mathsf{H}(\mathbb{X} \cap B_r(y), \mathbb{X} \cap \partial B_r(y))$. Both of these groups will then map, via intersection of chains, isomorphically into the group $\mathsf{H}(\mathbb{X} \cap B_r(x) \cap B_r(y), \partial(B_r(x) \cap B_r(y))$, and the isomorphism above is then derived from these two maps. In [19], the authors define the concept of a homology stratification of a space $\mathbb{X}$. Briefly, they require a decomposition of $\mathbb{X}$ into pieces such that the locally homology groups are locally constant across each piece; more precisely, that the maps discussed above be isomorphisms for each pair of close enough points in each piece.

3 Topological Inference Theorem

From the discussion above, it is easy to see that any stratification of a topological space will also be a homology stratification. The converse is unfortunately false. However, we can build a useful analytical tool based on the contrapositive: given two points in a point cloud, we can hope to state, based on their local homology groups and the maps between them, that the two points should not be placed in the same piece of any stratification. To do this, we first adapt the definition of these local homology maps into a more multi-scale and robust framework. More specifically, we introduce a radius parameter r and a notion of local equivalence, $\sim_r$, which allows us to group the points of $\mathbb{X}$, as well as of the ambient space, into strata at this radius scale. We then give the main result of this section: topological conditions under which the point cloud $\mathbb{U}$ can be used to infer the strata at different radius scales.

3.1 Local Equivalence

Local homology intersection map. We assume that we are given some topological space $\mathbb{X}$ embedded in some Euclidean space in $\mathbb{R}^N$. For each radius $r \geq 0$, and for each pair of points $p, q \in \mathbb{R}^N$, we define the following homology map $\phi^{\mathbb{X}}(p, q, r)$:

$$\mathsf{H}(\mathbb{X} \cap B_r(p), \mathbb{X} \cap \partial B_r(p)) \to \mathsf{H}(\mathbb{X} \cap B_r(p) \cap B_r(q), \mathbb{X} \cap \partial(B_r(p) \cap B_r(q))). \tag{1}$$

Intuitively, this map can be understood as taking a chain, throwing away the parts that lie outside the smaller range, and then modding out the new boundary. Alternatively, one may think of it as being induced by a combination of inclusion and excision; for a formal and technical definition, see [2].

For example, consider the space $\mathbb{X}$ drawn in the plane as shown in Figures 4 (a), (b), and (c). For each pair (p, q) of points shown in the three parts of the figure, we let $f = \phi^{\mathbb{X}}(p, q, r)$ and $g = \phi^{\mathbb{X}}(q, p, r)$. Then the points p and q are considered to have the same local structure if f and g are both isomorphisms; equivalently, if $\ker f = \operatorname{cok} f = 0$ and if $\ker g = \operatorname{cok} g = 0$. In part (a), $\ker(f) \neq 0$, since the classes α_2 and α_3 go to zero when passing to the intersection. In part (b), there is a class $\gamma_2 \in \operatorname{cok} f$. The maps f and g in part (c) are both isomorphisms.

Returning to the general case, we use these maps to impose an equivalence relation on $\mathbb{R}^N$.

Definition 3.1 (Local equivalence)**.** *Two points x and y are said to have equivalent local structure at radius r, denoted $x \sim_r y$, iff there exists a chain of points $x = x_0, x_1, \ldots, x_m = y$ from $\mathbb{X}$ such that, for each $1 \leq i \leq m$, the maps $\phi^{\mathbb{X}}(x_{i-1}, x_i, r)$ and $\phi^{\mathbb{X}}(x_i, x_{i-1}, r)$ are both isomorphisms.*

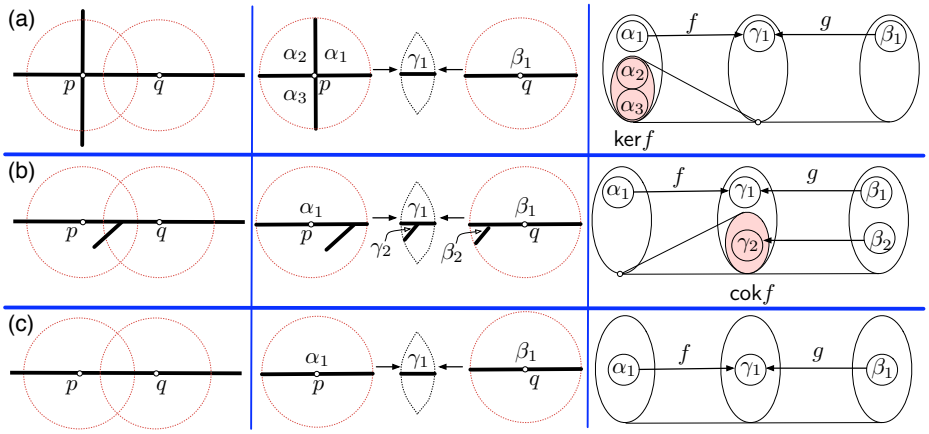

Figure 4: Let $f = \phi^{\mathbb{X}}(p, q, r)$ and $g = \phi^{\mathbb{X}}(q, p, r)$. The local homology classes are labeled in their corresponding locations. (a) p and q do not have the same local structure at radius r since $\ker f \neq 0$. (b) p and q do not have the same local structure at radius r since $\operatorname{cok} f \neq 0$. (c) p and q have the same local structure at radius r since $\ker f = \operatorname{cok} f = 0$ and $\ker g = \operatorname{cok} g = 0$.

In other words, x and y have the same local structure at this radius iff they can be connected by a chain of points which are pairwise close enough and whose local homology groups at radius r transfer isomorphically into each other via the intersection maps.

Different choices of r will of course lead to different equivalence classes. For example, consider the space $\mathbb{X}$ drawn in the plane as shown in the left half of Figure 5 (a). At the radius drawn, point z is equivalent to the cross point and is not equivalent to either the point x or y. Note that some points from the ambient space will now be considered equivalent to x and y, and some others will be equivalent to z. On the other hand, a smaller choice of radius would result in all three of x, y, and z belonging to the same equivalence class.

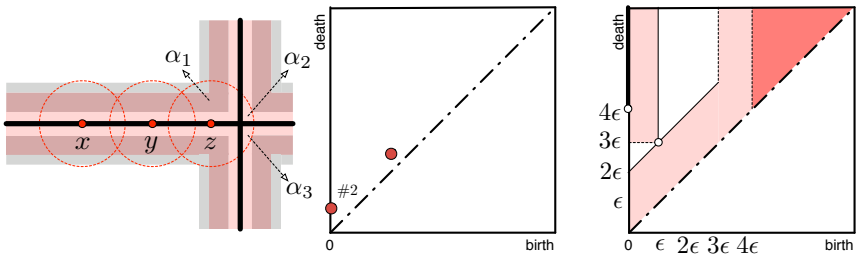

Figure 5: (a) Illustration of equivalence relation: left, $x \sim_r y$, $y \nsim_r z$; right, the 1-dim persistence diagram, for the kernel of the map going from the z ball into its intersection with the y ball. A number, i.e., $\#2$, labeling a point in the persistence diagram indicates its multiplicity. (b) Regions in $\mathbb{X}$-diagrams and $\mathbb{U}$-diagrams. The point in the $\mathbb{X}$-diagrams lie either along the solid black line or in the darkly shaded region. Adding the lightly shaded regions, we get the region of possible points in the $\mathbb{U}$-diagrams.

(Co)Kernel persistence. In order to relate the point cloud $\mathbb{U}$ to the equivalence relation $\sim_r$, we must first define a multi-scale version of the maps $\phi^{\mathbb{X}}(p, q, r)$; we do so by gradually thickening the space $\mathbb{X}$ using the sublevel sets of its distance function. For each $p, q \in \mathbb{R}^N$ and $r, \alpha \geq 0$, we will consider the intersection map $\phi^{\mathbb{X}}_{\alpha}(p, q, r)$, which is defined by substituting $\mathbb{X}_\alpha$ for $\mathbb{X}$ in (1). Note of course that $\phi^{\mathbb{X}}(p, q, r) = \phi^{\mathbb{X}}_0(p, q, r)$.

For the moment, we fix a choice of p, q, and r, and we use the following shorthand, $B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap B_r(p)$, $\partial B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap \partial B_r(p)$, $B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap B_r(p) \cap B_r(q)$, $\partial B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap \partial(B_r(p) \cap B_r(q))$, and we also often write $B^{\mathbb{X}}_p = B^{\mathbb{X}}_p(0)$ and $B^{\mathbb{X}}_{pq} = B^{\mathbb{X}}_{pq}(0)$. By replacing $\mathbb{X}$ with $\mathbb{U}$ in this shorthand, we also write $B^{\mathbb{U}}_p(\alpha) = \mathbb{U}_\alpha \cap B_r(p)$, and so forth.

For any pair of non-negative real values $\alpha \leq \beta$ the inclusion $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ gives rise to the following commutative diagram:

$$\begin{array}{ccc} \mathsf{H}(B_p^{\mathbb{X}}(\alpha), \partial B_p^{\mathbb{X}}(\alpha)) & \xrightarrow{\phi_\alpha^{\mathbb{X}}} & \mathsf{H}(B_{pq}^{\mathbb{X}}(\alpha), \partial B_{pq}^{\mathbb{X}}(\alpha)) \\ \downarrow & & \downarrow \\ \mathsf{H}(B_p^{\mathbb{X}}(\beta), \partial B_p^{\mathbb{X}}(\beta)) & \xrightarrow{\phi_\beta^{\mathbb{X}}} & \mathsf{H}(B_{pq}^{\mathbb{X}}(\beta), \partial B_{pq}^{\mathbb{X}}(\beta)) \end{array} \tag{2}$$

Hence there are maps $\ker \phi_\alpha^{\mathbb{X}} \to \ker \phi_\beta^{\mathbb{X}}$ and $\operatorname{cok} \phi_\alpha^{\mathbb{X}} \to \operatorname{cok} \phi_\beta^{\mathbb{X}}$. Allowing α to increase from 0 to ∞ gives rise to two persistence modules, $\{\ker \phi_\alpha^{\mathbb{X}}\}$ and $\{\operatorname{cok} \phi_\alpha^{\mathbb{X}}\}$, with diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ and $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$. Recall that a homomorphism is an isomorphism iff its kernel and cokernel are both zero. In our context then, the map $\phi^{\mathbb{X}}$ is an isomorphism iff neither $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ nor $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$ contain any points on the y-axis above 0.

Example. As shown in the left part of Figure 5 (a), x, y, and z are points sampled from a cross embedded in the plane. Taking r as drawn, the right part of Figure 5 (a) displays $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$, where $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(z, y, r)$; we now explain this diagram in some detail. The group $\mathsf{H}_1(B_z^{\mathbb{X}}, \partial B_z^{\mathbb{X}})$ has rank three; as a possible basis we might take the three local homology classes represented by α_1, α_2, and α_3, which are pairs of segments defining the northwest-facing right angle, the northeast-facing right angle, and the southeast-facing right angle. Under the intersection map $\phi^{\mathbb{X}} = \phi_0^{\mathbb{X}}$, the first of these classes α_1 maps to the generator of $\mathsf{H}_1(B_{zy}^{\mathbb{X}}, \partial B_{zy}^{\mathbb{X}})$, while the other two map to zero. Hence $\ker \phi_0^{\mathbb{X}}$ has rank two. As $\mathbb{X}$ starts to thicken into the ambient space, both classes in this kernel eventually die, one at the α value which fills in the northeast corner of the larger ball, and the other at the α value which fills in the southeast corner; these two values are the same here due to symmetry in the picture. At this value, the map $\phi_\alpha^{\mathbb{X}}$ becomes an isomorphism and it remains so until the intersection of the two balls fills in completely. This gives birth to a new kernel class which subsequently dies when the larger ball finally fills in. The diagram $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$ thus contains three points; the leftmost two show that the map $\phi^{\mathbb{X}}$ is not an isomorphism, and thus that z and y do not have the same local structure at the chosen radius level.

3.2 Topological Inference Theorem

Given a point cloud $\mathbb{U}$ sampled from $\mathbb{X}$, we consider the following question: for a radius r, how can we infer whether or not any given pair of points in $\mathbb{U}$ has the same local structure at this radius? In this subsection, we prove a theorem which describes the circumstances under which we can make the above inference. Naturally, any inference will require that we use $\mathbb{U}$ to judge whether or not the maps $\phi^{\mathbb{X}}(p, q, r)$ are isomorphisms. The basic idea is that if $\mathbb{U}$ is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by $\mathbb{U}$ will be good enough approximations of the diagrams defined by $\mathbb{X}$.

(Co)Kernel stability. Again we fix p, q, and r, and write $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$. For each $\alpha \geq 0$, we let $\mathbb{U}_\alpha = d_{\mathbb{U}}^{-1}[0, \alpha]$. We consider $\phi_\alpha^{\mathbb{U}} = \phi_\alpha^{\mathbb{U}}(p, q, r)$, defined by replacing $\mathbb{X}$ with $\mathbb{U}_\alpha$ in (1), as

$$\mathsf{H}(\mathbb{U}_\alpha \cap B_r(p), \mathbb{U}_\alpha \cap \partial B_r(p)) \to \mathsf{H}(\mathbb{U}_\alpha \cap B_r(p) \cap B_r(q), \mathbb{U}_\alpha \cap \partial(B_r(p) \cap B_r(q))). \tag{3}$$

Running α from 0 to ∞, we obtain two more persistence modules, $\{\ker \phi_\alpha^{\mathbb{U}}\}$ and $\{\operatorname{cok} \phi_\alpha^{\mathbb{U}}\}$, with diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{U}})$.

If $\mathbb{U}$ is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by $\mathbb{U}$ will be good approximations of the diagrams defined by $\mathbb{X}$. More precisely, we have the following theorem, see the full version for the proof.

Theorem 3.1 ((Co)Kernel Diagram Stability)**.** *The bottleneck distances between the (co)kernel diagrams of $\phi^{\mathbb{U}}$ and $\phi^{\mathbb{X}}$ are upper-bounded by the Hausdorff distance between $\mathbb{U}$ and $\mathbb{X}$:*

$$d_B(\mathrm{Dgm}(\ker \phi^{\mathbb{U}}), \mathrm{Dgm}(\ker \phi^{\mathbb{X}})) \leq d_H(\mathbb{U}, \mathbb{X}), \quad d_B(\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{U}}), \mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})) \leq d_H(\mathbb{U}, \mathbb{X}).$$

Main inference result. We now suppose that we have a point sample $\mathbb{U}$ of a space $\mathbb{X}$, where the Hausdorff distance between the two is no more than some ε. In this case, we call $\mathbb{U}$ an ε-approximation of $\mathbb{X}$. Given two points $p, q \in U$ and a fixed radius r, we set $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$, and we wish to determine whether or not $\phi^{\mathbb{X}}$ is an isomorphism. Since we only have access to the point sample $\mathbb{U}$, we instead compute the diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{U}})$; we provide an algorithm for doing this in Section 5.

Given any persistence diagram $\mathcal{D}$, which we recall is a multi-set of points in the extended plane, and two positive real numbers $a < b$, we let $\mathcal{D}(a, b)$ denote the intersection of $\mathcal{D}$ with the portion of the extended plane which lies

above $y = b$ and to the left of $x = a$; note that these points correspond to classes which are born no later than a and die no earlier than b.

For a fixed choice of p, q, r, we consider the following two persistence modules: $\{\mathsf{H}(B_p^{\mathbb{X}}(\alpha), \partial B_p^{\mathbb{X}}(\alpha))\}$ and $\{\mathsf{H}(B_{pq}^{\mathbb{X}}(\alpha), \partial B_{pq}^{\mathbb{X}}(\alpha))\}$. We let $\sigma(p, r)$ and $\sigma(p, q, r)$ denote their respective feature sizes and then set $\rho(p, q, r)$ to their minimum. Geometrically, $\rho(p, q, r)$ is related to a local *reach* and the gradient of $d_{\mathbb{X}}$ (as detailed in [2]).

We now give the main theorem of this section, which states that we can use $\mathbb{U}$ to decide whether or not $\phi^{\mathbb{X}}(p, q, r)$ is an isomorphism as long as $\rho(p, q, r)$ is large enough relative to the sampling density, see the full version for its proof.

Theorem 3.2 (Topological Inference Theorem)**.** *Suppose that we have an ε-approximation $\mathbb{U}$ from $\mathbb{X}$. Then for each pair of points $p, q \in \mathbb{R}^N$ such that $\rho = \rho(p, q, r) \geq 4\varepsilon$, the map $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$ is an isomorphism iff* $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})(\varepsilon, 3\varepsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})(\varepsilon, 3\varepsilon) = \emptyset$.

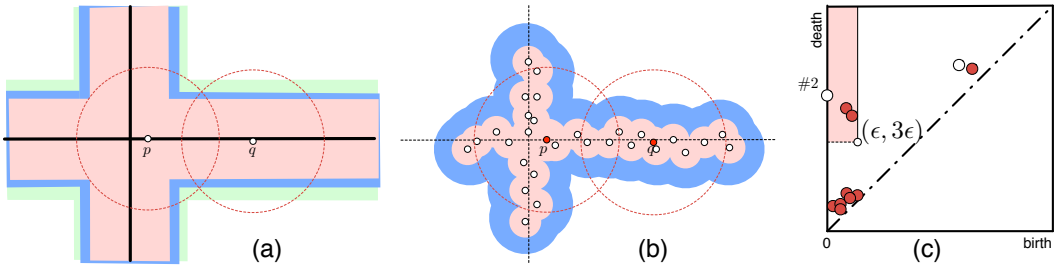

Figure 6: (a) Space $\mathbb{X}$ is the black cross, with p, q, r as drawn. (b) Space $\mathbb{U}$ is an ε-approximation of $\mathbb{X}$. (c) The kernel persistence diagram of $\phi^{\mathbb{X}}(p, q, r)$ ($\mathbb{X}$-diagram) is shown to contain black empty points; the kernel diagram of $\phi^{\mathbb{U}}(p, q, r)$ ($\mathbb{U}$-diagram) is shown to contain red filled points. Suppose only $\mathbb{U}$ is given, we can use $\mathbb{U}$-diagram to infer that points p and q are not locally equivalent.

Figure 5 (b) illustrates Theorem 3.2, that is, under certain topological conditions, $\phi^{\mathbb{X}}$ is an isomorphism if and only if certain regions in the $\mathbb{U}$-diagrams are empty. For example, suppose $\mathbb{X}$ is the cross shown in Figure 6 (a), with p, q, r as drawn. p and q are locally different at this radius level, as shown by the presence of two black empty points on the y-axis of the kernel persistence $\mathbb{X}$-diagram (Figure 6 (c)). Suppose $\mathbb{X}$ is unknown and we are only given $\mathbb{U}$, an ε-approximation of $\mathbb{X}$ (Figure 6 (b)). From the kernel $\mathbb{U}$-diagram, which has two points in the relevant rectangle, we can infer that p and q do not have the same local structure at radius level r by applying Theorem 3.2.

4 Probabilistic Inference Theorem

The topological inference of Section 3 states conditions under which the point sample $\mathbb{U}$ can be used to infer stratification properties of the space $\mathbb{X}$. The basic condition is that the Hausdorff distance between the two must be small. In this section we describe a probabilistic model for generating the point sample $\mathbb{U}$, and we provide an estimate of how large this point sample should be to infer stratification properties of the space $\mathbb{X}$ with a quantified measure of confidence. More specifically, we provide a local estimate, based on $\rho(p, q, r)$ and $\rho(q, p, r)$, of how many sample points are needed to infer the local relationship at radius level r between two fixed points p and q; this same theorem can be used to give a global estimate of the number of points needed for inference between any pair of points whose ρ-values are above some fixed low threshold.

Sampling strategy. We assume $\mathbb{X}$ to be compact. Since the stratified space $\mathbb{X}$ can contain singularities and maximal strata of varying dimensions, some care is required in the sampling design. Consider for example a sheet of area one, punctured by a line of length one. In this case, sampling from a naively constructed uniform measure on this space would result in no points being drawn from the line. This same issue arose and was dealt with in [18], although in a slightly different approach than we will develop.

A sampling strategy that will deal with the problem of varying dimensions is to use a mixture model. In the example of the sheet and line, a uniform measure would be placed on the sheet, while another uniform measure would be placed on the line, and a mixture probability is placed on the two measures; for example, each measure could be drawn with probability $1/2$. We now formalize this approach. Consider each (non-empty) i-dimensional

stratum $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ of $\mathbb{X}$. All strata that are included in the closure of some higher-dimensional strata, in other words all non-maximal strata, are not considered in the model. A uniform measure is assigned to the closure of each maximal stratum, $\mu_i(\mathbb{S}_i)$, this is possible since each such closure is compact. We assume a finite number of maximal strata K and assign to the closure of each such stratum a probability $p_i = 1/K$. This implies the following density $f(x) = \frac{1}{K}\sum_{j=1}^{K} \nu_i(X = x)$, where ν_i is the density corresponding to measure μ_i. The point sample is generated from the following model: $\mathbb{U} = \{x_1, ..., .x_n\} \overset{iid}{\sim} f(x)$. We call this model M.

Lower bounds on the sample size of the point cloud. Our main theorem is the probabilistic analogue of Theorem 3.2. An immediate consequence of this theorem is that, for two points $p, q \in \mathbb{U}$, we can infer with probability at least $1-\xi$ whether p and q are locally equivalent, $p \sim_r q$. The confidence level $1-\xi$ will be a monotonic function of the size of the point sample. The theorem involves a parameter $v(\rho)$, for each positive ρ, which is based on the volume of the intersection of ρ-balls with $\mathbb{X}$. First we note that each maximal stratum of $\mathbb{X}$ comes with its own notion of volume: in the plane punctured by a line example, we measure volume in the plane and in the line as area and length, respectively. The volume $\mathrm{vol}\,(\mathbb{Y})$ of any subspace $\mathbb{Y}$ of $\mathbb{X}$ is the sum of the volumes of the intersections of $\mathbb{Y}$ with each maximal stratum. For $\rho > 0$, we define $v(\rho) = \inf_{x\in\mathbb{X}} \frac{\mathrm{vol}\,(B_{\rho/32}(x)\cap\mathbb{X})}{\mathrm{vol}\,(\mathbb{X})}$. We then have our main theorem, whose proof appears in the full version:

Theorem 4.1 (Local Probabilistic Sampling Theorem)**.** *Let $\{x_1, x_2, ..., x_n\}$ be drawn from model M. Fix a pair of points $p, q \in \mathbb{R}^N$ and a positive radius r, and put $\rho = \min\{\rho(p,q,r), \rho(q,p,r)\}$. If $n \geq \frac{1}{v(\rho)}\left(\log\frac{1}{v(\rho)} + \log\frac{1}{\xi}\right)$, then, with probability greater than $1-\xi$ we can correctly infer whether or not $\phi^{\mathbb{X}}(p,q,r)$ and $\phi^{\mathbb{X}}(q,p,r)$ are both isomorphisms.*

To extend the above theorem to a more global result, one can pick a positive ρ and radius r, and consider the set of all pairs of points (p, q) such that $\rho \leq \min\{\rho(p,q,r), \rho(q,p,r\}$. Applying Theorem 4.1 uniformly to all pairs of points will give the minimum number of sample points needed to settle the isomorphism question for all of the intersection maps between all pairs.

5 Algorithm

The theorems in the previous sections give conditions under which a point cloud $\mathbb{U}$, sampled from a stratified space $\mathbb{X}$, can be used to infer the local equivalences between points on $\mathbb{X}$. We now switch gears slightly, and imagine clustering the $\mathbb{U}$-points into strata. The basic strategy is to build a graph on the point set, with edges corresponding to positive isomorphism judgments. The connected components of this graph will then be our proposed strata. A crucial subroutine in the clustering algorithm is the computation of the diagrams $\mathrm{Dgm}(\ker\phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\,\phi^{\mathbb{U}})$, for $\phi^{\mathbb{U}} = \phi^{\mathbb{U}}(p,q,r)$ between all pairs $(p,q) \in \mathbb{U}\times\mathbb{U}$. We will focus our attention in this section on the computation of the (co)kernel diagrams, for details on the entire algorithm and its robustness see the full version.

To compute the diagrams $\mathrm{Dgm}(\ker\phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\,\phi^{\mathbb{U}})$ we require for each $\alpha \geq 0$ a simplicial analogue of the map $\phi^{\mathbb{U}}_{\alpha} : \mathsf{H}(B^{\mathbb{U}}_{p}(\alpha), \partial B^{\mathbb{U}}_{p}(\alpha)) \to \mathsf{H}(B^{\mathbb{U}}_{pq}(\alpha), \partial B^{\mathbb{U}}_{pq}(\alpha))$. We define, for each $\alpha \geq 0$ (a) two pairs of simplicial complexes $L_0(\alpha) \subseteq L(\alpha)$ and $K_0(\alpha) \subseteq K(\alpha)$, and (b) a relative homology map between them $\psi_\alpha : \mathsf{H}(L(\alpha), L_0(\alpha)) \to \mathsf{H}(K(\alpha), K_0(\alpha))$. In the full version, we give a correctness proof that $\mathrm{Dgm}(\ker\phi^{\mathbb{U}}) = \mathrm{Dgm}(\ker\psi)$ and $\mathrm{Dgm}(\mathrm{cok}\,\phi^{\mathbb{U}}) = \mathrm{Dgm}(\mathrm{cok}\,\psi)$.

5.1 Preliminaries

To construct the simplicial complexes in our algorithm, we will compute Voronoi diagrams and nerves of sets of collections derived from these Voronoi diagrams.

Voronoi diagram. Given a finite collection $\mathbb{U}$ of points in $\mathbb{R}^N$ and $u_i \in \mathbb{U}$, then the *Voronoi cell* of u_i is defined to be:

$$V_i = V(u_i) = \{x \in \mathbb{R}^N \mid ||x - u_i|| \leq ||x - u_j||, \forall u_j \in \mathbb{U}\}.$$

The set of cells V_i covers the entire space and forms the *Voronoi diagram* of $\mathbb{R}^N$, denoted as $\mathrm{Vor}\,(\mathbb{U}|\mathbb{R}^N)$. If we restrict each V_i to some subset $\mathbb{X} \subseteq \mathbb{R}^N$, then the set of cells $V_i \cap \mathbb{X}$ forms a *restricted Voronoi diagram*, denoted as $\mathrm{Vor}\,(\mathbb{U}|\mathbb{X})$. For a simplex σ with vertices in $\mathbb{U}$, we set $V_\sigma = \cap_{u_i \in \sigma} V_i$.

Nerves. The *nerve* $N(\mathcal{C})$ of a finite collection of sets $\mathcal{C}$ is defined to be the abstract simplicial complex with vertices corresponding to the sets in $\mathcal{C}$ and with simplices corresponding to all non-empty intersections among these sets, $N(\mathcal{C}) = \{S \subseteq \mathcal{C} \mid \bigcap S \neq \emptyset\}$. Every abstract simplicial complex can be geometrically realized, and therefore the concept of homotopy type makes sense. Under certain conditions, for example whenever the sets in $\mathcal{C}$ are all closed and convex subsets of Euclidean space ([10], p.59), the nerve of $\mathcal{C}$ has the same homotopy type, and thus the same homology groups, as the union of sets in $\mathcal{C}$. This implies we can compute $\mathsf{H}(U_\alpha)$, the absolute homology of the thickened point cloud, by computing the nerve of the collection of sets $V_i \cap \mathbb{U}_\alpha$.

The nerve of the restricted Voronoi diagram $\mathrm{Vor}\,(\mathbb{U}|\mathbb{X})$ is called the *restricted Delaunay triangulation*, denoted as $\mathrm{Del}\,(\mathbb{U}|\mathbb{X})$. It contains the set of simplices σ for which $V_\sigma \cap \mathbb{X} \neq \emptyset$.

Power cells, lunes, and moons. We need to compute the relative homology groups $\mathsf{H}(B_p^{\mathbb{U}}(\alpha), \partial B_p^{\mathbb{U}}(\alpha))$ and $\mathsf{H}(B_{pq}^{\mathbb{U}}(\alpha), \partial B_{pq}^{\mathbb{U}}(\alpha))$. The direct argument used to compute absolute homology based on the nerve does not apply to computing relative homology groups since the collection of the sets $V_i \cap \partial B_p^{\mathbb{U}}(\alpha)$ and $V_i \cap \partial B_{pq}^{\mathbb{U}}(\alpha)$ need not be convex.

To get around this problem, we first define the *power cell* with respect to $B_r(p)$, $P(\alpha)$, as $P(\alpha) = \{x \in \mathbb{R}^N \mid ||x-p||^2 - r^2 \leq ||x-u||^2 - \alpha^2, \forall u \in \mathbb{U}\}$, and we set $P_0(\alpha) = B_r(p) - \operatorname{int} P(\alpha)$. Replacing p with q in this formula gives $Q(\alpha)$, the power cell with respect to $B_r(q)$. Finally, we set $Z(\alpha) = P(\alpha) \cap Q(\alpha)$, and $Z_0(\alpha) = (B_r(p) \cap B_r(q)) - \operatorname{int} Z(\alpha)$. These definitions are illustrated in Figure 7 (a). Note that $P_0(\alpha)$ and $Z_0(\alpha)$ are both contained in $\mathbb{U}_\alpha$.

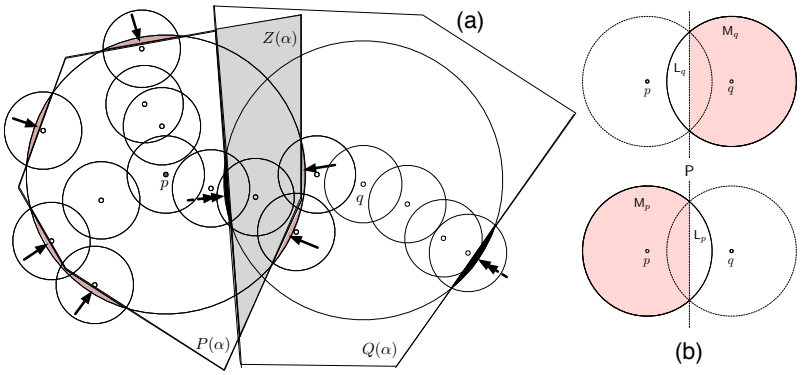

Figure 7: (a) Illustration of intersection power cell $Z(\alpha)$, as the grey shaded region. The unshaded convex regions are $P(\alpha)$ and $Q(\alpha)$ respectively. The dark pink and black shaded regions (pointed by single and double arrows) correspond to $P_0(\alpha)$ and $Q_0(\alpha)$ respectively. (b) Illustration of the lune and the moon. The shaded regions are the respective moons. The white regions within solid circles are the respective lunes.

It turns out that replacing $\partial B_p^{\mathbb{U}}(\alpha)$ with $P_0(\alpha)$ and $\partial B_{pq}^{\mathbb{U}}(\alpha)$ with $Z_0(\alpha)$ has no effect on the relative homology groups in question. That is, the spaces $(B_p^U(\alpha), \partial B_p^U(\alpha))$ and $(B_p^U(\alpha), P_0(\alpha))$ are homotopy equivalent, so are the spaces $(B_{pq}^U(\alpha), \partial B_{pq}^U(\alpha))$ and $(B_{pq}^U(\alpha), Z_0(\alpha))$. Consequently, their homology groups are isomorphic. The first part of this statement was proven in [1], and a proof of the second appears in [2]. The sets $V_i \cap P_0(\alpha)$ are convex [1]. Unfortunately, it is still possible for $V_i \cap Z_0(\alpha)$ to be non-convex, which requires a further subdivision of the Voronoi cells by bisection. Consider the hyperplane P of points in $\mathbb{R}^N$ which are equidistant from p and q. This will divide $\mathbb{R}^N$ into two half-spaces with P_p and P_q denoting the half-spaces containing p and q. Given P_p we define the *p-lune*, L_p, and *p-moon*, M_p, as follows (see Figure 7 (b)): $\mathsf{L}_p = \mathsf{P}_q \cap B_r(p), \mathsf{M}_p = \mathsf{P}_p \cap B_r(p)$.

The lune and the moon divide each Voronoi cell into two parts, $V_i^{\mathsf{L}} = V_i \cap \mathsf{L}_p$ and $V_i^{\mathsf{M}} = V_i \cap \mathsf{M}_p$. These sets are obviously convex, assuming they are non-empty, since they are each the intersection of two convex sets. It also turns out that the non-empty sets among $V_i^{\mathsf{L}} \cap Z_0(\alpha)$ and $V_i^{\mathsf{M}} \cap Z_0(\alpha)$ are convex; see [2] for a proof.

5.2 Algorithm to compute simplicial analogues

Our algorithm to compute simplicial analogues contains two steps: (a) defining the simplicial complexes and (b) defining the corresponding relative homology simplicial maps. We first define the pairs of simplicial complexes $L_0(\alpha) \subseteq L(\alpha)$ and $K_0(\alpha) \subseteq K(\alpha)$. Set $\mathcal{A}$ to be the collection of the non-empty sets among $V_i^{\mathsf{L}} \cap B_p^{\mathbb{U}}(\alpha)$ and $V_i^{\mathsf{M}} \cap B_p^{\mathbb{U}}(\alpha)$. Define $\mathcal{A}_0$ as the collection of the nonempty sets among $V_i^{\mathsf{L}} \cap P_0(\alpha)$ and $V_i^{\mathsf{M}} \cap P_0(\alpha)$. Note that $\cup\mathcal{A} = B_p^{\mathbb{U}}(\alpha)$ and $\cup\mathcal{A}_0 = P_0(\alpha)$. Taking the nerve of both collections, we define the simplicial complexes $L(\alpha) = N(\mathcal{A})$ and $L_0(\alpha) = N(\mathcal{A}_0)$. Similarly, we define $\mathcal{C}$ and $\mathcal{C}_0$ to be the collections of the non-empty sets among, respectively, $V_i^{\mathsf{L}} \cap B_{pq}^{\mathbb{U}}(\alpha)$ and $V_i^{\mathsf{M}} \cap B_{pq}^{\mathbb{U}}(\alpha)$, and $V_i^{\mathsf{L}} \cap Z_0(\alpha)$ and $V_i^{\mathsf{M}} \cap Z_0(\alpha)$. We define $K(\alpha) = N(\mathcal{C})$ and $K_0(\alpha) = N(\mathcal{C}_0)$. See Figure 8 for an example of the simplicial complexes constructed in $\mathbb{R}^2$ for a given $\mathbb{U}$. To demonstrate how our algorithm works, we test it on synthetic data shown in the full version.

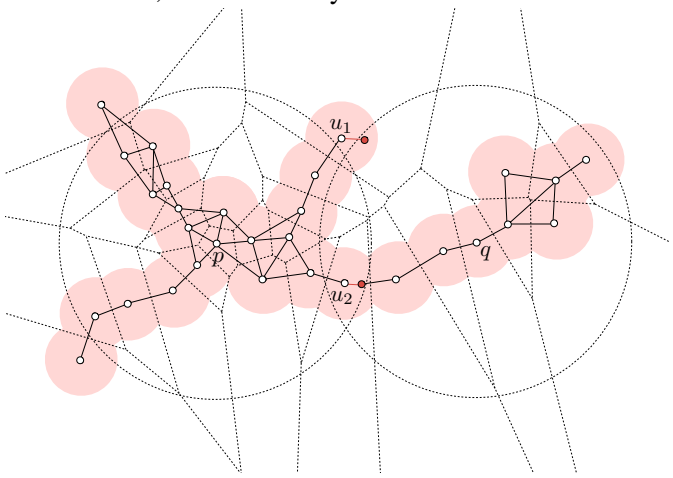

Figure 8: Illustration of the simplicial complexes constructed around two points p and q. The underlying Voronoi decomposition of the space is shown in thin dotted lines. u_1 and u_2 in $\mathbb{U}$ are the points whose restricted Voronoi regions intersect with the lune at non-convex regions.

To define the maps $\psi_\alpha : \mathsf{H}(L(\alpha), L_0(\alpha)) \to \mathsf{H}(K(\alpha), K_0(\alpha))$ we need the following technical lemma:

Lemma 5.1 (Containment Lemma)**.** *Assume that a simplex σ is in $L_0(\alpha)$. If σ is also in $K(\alpha)$, then σ is in $K_0(\alpha)$, as well, see the full version for the proof.*

To define ψ_α, we first construct a chain map $g = g_\alpha : \mathsf{C}(L(\alpha)) \to \mathsf{C}(K(\alpha))$ as follows. Given a simplex $\sigma \in L(\alpha)$, we define $g(\sigma) = \sigma$ if $\sigma \in K(\alpha)$, and $g(\sigma) = 0$ otherwise; we then extend g to a chain map by linearity. Using the Containment Lemma, we see that $g(\mathsf{C}(L_0(\alpha))) \subseteq \mathsf{C}(K_0(\alpha))$, and thus g descends to a relative chain map $f = f_\alpha : \mathsf{C}(L(\alpha), L_0(\alpha)) \to \mathsf{C}(K(\alpha), K_0(\alpha))$. Since f clearly commutes with all boundary operators, it induces a map on relative homology, this is our $\psi = \psi_\alpha$. To compute the diagrams involving ψ, we reduce various boundary matrices via (co)kernel persistence algorithm described in [8], in time at most cubic in the size of the simplicial complexes representing the data.

6 Discussion

We have presented a first step towards learning stratified spaces. There are several open issues of interest including: algorithmic efficiency and scaling with dimension using Rips or Witness complexes [9] instead of Delaunay triangulation, robustness of the algorithm and weighting local equivalence, and extensions to the noisy setting [17] when the mixture is concentrated around the stratified space.

Specifically, the algorithm to compute the (co)kernel diagrams from the thickened point cloud is based on an adaption of Delaunay triangulation and the power-cell construction. This algorithm should be quite slow when the dimensionality of the ambient space is high due to the runtime complexity of Delaunay triangulation. One idea to address this bottleneck is to use Rips or Witness complexes [9]. Another approach is to use dimension reduction techniques such as principal components analysis (PCA) or random projection that approximately preserve distance [6] as a preprocessing step. Another idea that may work if the ambient dimension is not too high is using faster algorithms to construct Delaunay triangulations [3].

References

[1] Paul Bendich, David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy Morozov. Inferring local homology from sampled stratified spaces. In *Proceedings 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 536–546, 2007.

[2] Paul Bendich, Bei Wang, and Sayan Mukherjee. Towards stratification learning through homology inference. http://arxiv.org/abs/1008.3572, August 2010.

[3] Jean-Daniel Boissonnat, Olivier Devillers, and Samuel Hornus. Incremental construction of the delaunay triangulation and the delaunay graph in medium dimension. *Proceedings 25th Annual Symposium on Computational Geometry*, pages 208–216, 2009.

[4] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings 25th Annual Symposium on Computational Geometry*, pages 237–246, 2009.

[5] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in euclidean space. *Discrete and Computational Geometry*, 41:461–479, 2009.

[6] Kenneth L. Clarkson. Tighter bounds for random projections of manifolds. *Proceedings 24th Annual Symposium on Computational Geometry*, pages 39–48, 2008.

[7] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, **37**:103–120, 2007.

[8] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy Morozov. Persistence homology for kernels, images and cokernels. *Proceedings 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1011–1020, 2009.

[9] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Symposium on Point-Based Graphics*, pages 157–166, 2004.

[10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[11] Mark Goresky and Robert MacPherson. *Stratified Morse Theory*. Springer-Verlag, 1988.

[12] Gloria Haro, Gregory Randall, and Guillermo Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Advances in NIPS*, 17, 2005.

[13] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[14] Bruce Hughes and Shmuel Weinberger. Surgery and stratified spaces. *Surveys on Surgery Theory*, pages 311–342, 2000.

[15] Gilad Lerman and Teng Zhang. Probabilistic recovery of multiple subspaces in point clouds by geometric lp minimization, 2010.

[16] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley, Redwood City, California, 1984.

[17] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Computational Geometry*, 39:419–441, 2008.

[18] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning from noisy data. Manuscript, 2008.

[19] Colin Rourke and Brian Sanderson. Homology stratifications and intersection homology. *Geometry and Topology Monographs*, 2:455–472, 1999.

[20] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1945 – 1959, 2005.

[21] Shmuel Weinberger. *Chicago Lectures in Mathematics*, chapter The topological classification of stratified spaces. University of Chicago Press, Chicago, IL, 1994.

Acknowledgments

All the authors would like to thank Herbert Edelsbrunner and John Harer for useful discussions and suggestions. PB would like to thank David Cohen-Steiner and Dmitrity Morozov for helpful discussion, and SM would like to thank Shmuel Weinberger for useful comments. SM and BW would like to acknowledge the support of NIH Grants R01 CA123175-01A1 and P50 GM 081883, and SM would like to acknowledge the support of NSF Grant DMS-07-32260.

Local Homology Transfer and Stratification Learning

Paul Bendich
bendich@math.duke.edu
Department of Mathematics, Duke University and IST Austria

Bei Wang
beiwang@sci.utah.edu
SCI Institute, University of Utah

Sayan Mukherjee
sayan@stat.duke.edu
Departments of Statistical Science, Mathematics, and Computer Science, Duke University

Abstract

A stratified space is a collection of manifolds of different dimensions which fit together uniformly inside some larger space. The objective of this paper is to show that data sampled from such a space can be clustered by strata. We first define a multi-scale notion of stratified spaces, providing a stratification at different scales which are indexed by a radius parameter. We then use methods derived from kernel and cokernel persistent homology to cluster the data points into different strata. We prove a correctness guarantee for this clustering method under certain topological conditions. We then provide a probabilistic guarantee for the clustering for the point sample setting – we provide bounds on the minimum number of sample points required to state with high probability which points belong to the same strata. Finally, we give an explicit algorithm for the clustering.

1 Introduction

Manifold learning is a basic problem in geometry, topology, and statistical inference that has received a great deal of recent attention. One formulation of the problem is: given a point cloud of data sampled from a manifold in an ambient space $\mathbb{R}^N$, infer the dimension and structure of the underlying manifold. A limitation of this problem statement is that it does not apply to sets that are not manifolds. For example, we may consider the more general class of stratified spaces that can be decomposed into strata – manifolds of varying dimension each of which fit together in some uniform way inside the higher dimensional space.

In this paper, we study the following problem in stratification learning: given a point cloud sampled from a stratified space, how do we cluster points that belong to the same stratum together while keeping points in different stratum apart. Intuitively, a reasonable strategy would be to place two points in the same piece of stratum if they "look the same locally" – they have identical neighborhoods within the larger space at some very small scale. However, the notion of "local" becomes unclear in the context of the uncertainty induced from sampling, since everything becomes noisy at small scales. In response, we introduce a radius or scale parameter r and define a notion of local equivalence at each scale r.

We will use tools derived from algebraic topology. In particular, we define local equivalence between points via maps that transfer information carried by local homology groups, and we then use persistent homology [14] methods to infer the properties of these maps.

Prior Work. Consistency in manifold learning has often been recast as homology inference – as the number of points in a point cloud goes to infinity, the homology inferred from the point cloud converges to the true homology of the underlying space. Results of this nature have been given for manifolds [25, 26] and a large class of compact subsets of Euclidean space [7]. Stronger results in homology inference for closed subsets of a metric space are given in [11].

Geometric approaches to stratification learning have also been developed. These include inference of a mixture of linear subspaces [22], mixture models for general stratified spaces [18], and generalized Principal Component Analysis (GPCA) [28] which was developed for dimension reduction for mixtures of manifolds.

The study of stratified spaces has long been a focus of pure mathematics; see, for example, [17, 29]. The problem of inference for the local homology groups of a sampled stratified space in a deterministic setting has been addressed in [3].

Contributions. In this paper we propose an approach to stratification learning based on local homology inference. The results in this paper are:

(1) A topological definition of two points belonging to the same strata by assessing the multi-scale local structure of the points through a local homology transfer map. (Definition 3.1);

(2) Topological conditions on the point sample under which this characterization holds (Theorem 3.2);

(3) Finite sample bounds for the minimum number of points required in the sample to state with high probability which points belong to the same strata (Theorem 4.1);

(4) An algorithm that computes which points belong to the same strata (Section 5).

2 Background

We first describe general persistence modules [6], focusing mainly on those that arise from maps between absolute or relative homology groups induced by inclusions of topological spaces or pairs of such spaces. We then discuss stratifications and their connection to the local homology groups of a topological space. Basics on homology itself are assumed; for a readable background, see [24] or [19], or [14] for a more computationally oriented treatment.

Persistence modules. For simplicity, our treatment of persistence modules adapted from [6] is restricted to $\mathbb{Z}/2\mathbb{Z}$-vector spaces. Let A be some subset of $\mathbb{R}$. A *persistence module* $\mathcal{F}_A$ is a collection $\{F_\alpha\}_{\alpha\in A}$ of $\mathbb{Z}/2\mathbb{Z}$-vector spaces, together with a family $\{f_\alpha^\beta : F_\alpha \to F_\beta\}_{\alpha\leq\beta\in A}$ of linear maps such that $\alpha \leq \beta \leq \gamma$ implies $f_\alpha^\gamma = f_\beta^\gamma \circ f_\alpha^\beta$. We will assume that the index set A is either $\mathbb{R}$ or $\mathbb{R}_{\geq 0}$ and not explicitly state indices unless necessary.

A real number α is said to be a *regular value* of the persistence module $\mathcal{F}$ if there exists some $\varepsilon > 0$ such that

the map $f_{\alpha-\delta}^{\alpha+\delta}$ is an isomorphism for each $\delta < \varepsilon$. Otherwise we say that α is a *critical value* of the persistence module; if $A = \mathbb{R}_{\geq 0}$, then $\alpha = 0$ will always be considered to be a critical value. We say that $\mathcal{F}$ is *tame* if it has a finite number of critical values and if all the vector spaces F_α are of finite rank. Any tame $\mathbb{R}_{\geq 0}$-module $\mathcal{F}$ must have a smallest non-zero critical value $\rho(\mathcal{F})$; we call this number the *feature size* of the persistence module. Assume $\mathcal{F}$ is tame and so we have a finite ordered list of critical values $0 = c_0 < c_1 < \ldots < c_m$. We choose regular values $\{a_i\}_{i=0}^m$ such that $c_{i-1} < a_{i-1} < c_i < a_i$ for all $1 \leq i \leq m$, and we adopt the shorthand notation $F_i \equiv F_{a_i}$ and $f_i^j : F_i \to F_j$, for $0 \leq i \leq j \leq m$. A vector $v \in \mathcal{F}_i$ is said to be *born* at level i if $v \notin \operatorname{im} f_{i-1}^i$, and such a vector *dies* at level j if $f_i^j(v) \in \operatorname{im} f_{i-1}^j$ but $f_i^{j-1}(v) \notin \operatorname{im} f_{i-1}^{j-1}$. This is illustrated in Figure 1 (a). We then define $P^{i,j}$ to be the vector space of vectors that are born at level i and then subsequently die at level j, and let $\beta^{i,j}$ denote its rank.

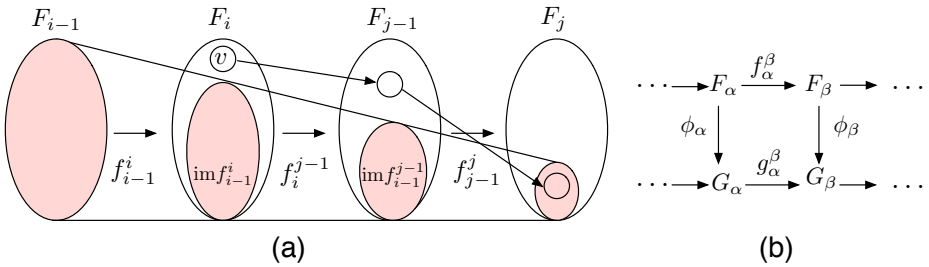

Figure 1: (a) The vector v is born at level i and then it dies at level j. (b) Commuting diagrams for (co)kernel modules.

Persistence diagrams. The information contained within a tame module $\mathcal{F}$ can be compactly represented by a *persistence diagram*, $\operatorname{Dgm}(\mathcal{F})$, which is a multi-set of points in the extended plane. It contains $\beta^{i,j}$ copies of the points (c_i, c_j), as well as infinitely many copies of each point along the major diagonal $y = x$. In Figure 3 (a) the persistence diagrams for a curve and a point cloud sampled from it are displayed, see below for a full explanation of this figure.

For any two points $u = (x, y)$ and $u' = (x', y')$ in the extended plane, we define $||u - u'||_\infty = \max\{|x - x'|, |y - y'|\}$. We define the *bottleneck distance* between any two persistence diagrams D and D' to be:

$$d_B(D, D') = \inf_{\Gamma : D \to D'} \sup_{u \in D} ||u - \Gamma(u)||_\infty,$$

where Γ ranges over all bijections from D to D'. Under certain conditions which we now describe, persistence diagrams will be stable under the bottleneck distance.

Two persistence modules $\mathcal{F}$ and $\mathcal{G}$ are said to be *strongly ϵ-interleaved* if, for some positive ϵ, there exist two families $\{\xi_\alpha : F_\alpha \to G_{\alpha+\epsilon}\}_\alpha$ and $\{\psi_\alpha : G_\alpha \to F_{\alpha+\epsilon}\}$ of linear maps which commute with the module maps $\{f_\alpha^\beta\}$ and $\{g_\alpha^\beta\}$ in the appropriate manner. More precisely, we require that, for each $\alpha \leq \beta$, the four diagrams in Figure 2 all commute.

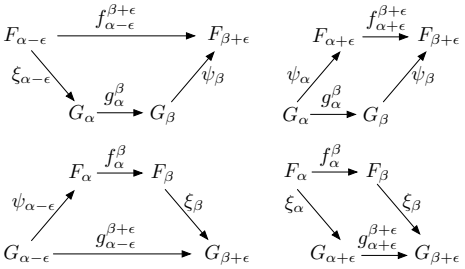

Figure 2: Commuting diagrams for strongly interleaving persistence modules.

We can now state the diagram stability result ([6], Theorem 4.4), that we will need below.

Theorem 2.1 (Diagram Stability Theorem)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two tame persistence modules and $\epsilon > 0$. If $\mathcal{F}$ and $\mathcal{G}$ are strongly ϵ-interleaved, then $d_B(\operatorname{Dgm}(\mathcal{F}), \operatorname{Dgm}(\mathcal{G})) \leq \epsilon$.*

When we wish to compute the persistence diagram associated to a module $\mathcal{F}$, it is often convenient to substitute another module $\mathcal{G}$, usually one defined in terms of simplicial complexes or other computable objects. The following theorem ([14], p.159) gives a condition under which this is possible.

Theorem 2.2 (Persistence Equivalence Theorem)**.** *Given two persistence modules $\mathcal{F}$ and $\mathcal{G}$, suppose there exist for each α isomorphisms $F_\alpha \cong G_\alpha$ which commute with the module maps, then* $\mathrm{Dgm}(\mathcal{F}) = \mathrm{Dgm}(\mathcal{G})$.

That is, if all the vertical maps are isomorphisms and all squares commute in the following diagram, then $\mathrm{Dgm}(\mathcal{F}) = \mathrm{Dgm}(\mathcal{G})$.

$$\begin{array}{ccccc} \ldots & \to F_\alpha & \to F_\beta & \to \ldots \\ & \uparrow\cong & \uparrow\cong & \\ \ldots & \to G_\alpha & \to G_\beta & \to \ldots \end{array}$$

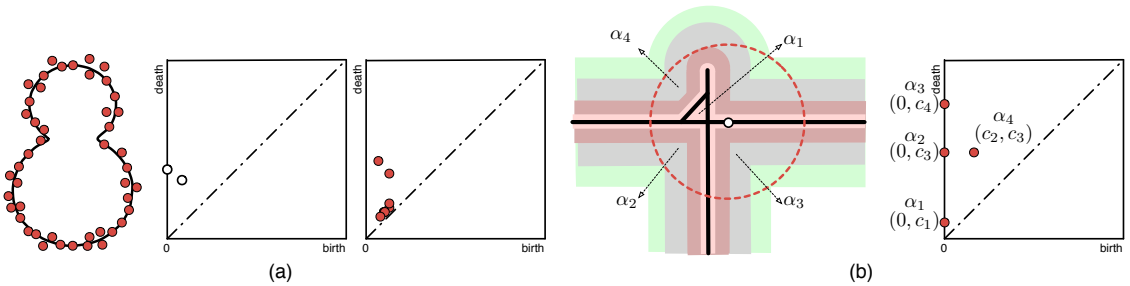

Figure 3: (a) Illustration of a point cloud and its persistence diagram: left, $\mathbb{X}$ is the curve embedded as shown in the plane and $\mathbb{U}$ is the point cloud; middle, the persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{X}})$; right, the persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{U}})$. The diagrams are generated by thickening $\mathbb{X}$ (or $\mathbb{U}$) while tracking the birth and death of homology classes. (b) Illustration of relative homology and its persistence diagram: left, the space $\mathbb{X}$ is in solid line and the closed ball B has dotted boundary; right, the persistence diagram for the module $\{\mathsf{H}_1(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}$. Here, α goes through four non-zero critical values $c_1 < c_2 < c_3 < c_4$ that correspond to the four colored level sets, where the points in the persistence diagram correspond to the birth and death of the four relative homology classes respectively. In particular, α_4 is created when the level set at value c_2 touches B.

(Co)Kernel modules. Suppose now that we have two persistence modules $\mathcal{F}$ and $\mathcal{G}$ along with a family of maps $\{\phi_\alpha : F_\alpha \to G_\alpha\}$ which commute with the module maps – for every pair $\alpha \leq \beta$, we have $g_\alpha^\beta \circ \phi_\alpha = \phi_\beta \circ f_\alpha^\beta$. In other words, every square commutes in the diagram shown in Figure 1 (b). Then, for each pair of real numbers $\alpha \leq \beta$, the restriction of f_α^β to $\ker \phi_\alpha$ maps into $\ker \phi_\beta$, giving rise to a new kernel persistence module, with persistence diagram denoted by $\mathrm{Dgm}(\ker \phi)$. Similarly, we obtain a cokernel persistence module, with diagram $\mathrm{Dgm}(\mathrm{cok}\, \phi)$.

Homology and distance functions. Consider a family of topological spaces $\{\mathbb{X}_\alpha\}$, along with inclusions $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ for all $\alpha \leq \beta$. The inclusions induce maps $\mathsf{H}_j(\mathbb{X}_\alpha) \to \mathsf{H}_j(\mathbb{X}_\beta)$, for each homological dimension $j \geq 0$, and hence we have persistence modules for each j. Defining $\mathsf{H}(\mathbb{X}_\alpha) = \bigoplus_j \mathsf{H}_j(\mathbb{X}_\alpha)$ and taking direct sums of maps in the obvious way, will also give one large direct-sum persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}$.

Given a compact topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$, we define $d_{\mathbb{X}}$ as the distance function which maps each point in the ambient space to the distance from its closest point in $\mathbb{X}$. We let $\mathbb{X}_\alpha$ denote the sublevel set $d_{\mathbb{X}}^{-1}[0, \alpha]$; each sublevel set should be thought of as a thickening of $\mathbb{X}$ within the ambient space. Increasing the thickening parameter produces a growing family of sublevel sets, giving rise to the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}_{\alpha \in \mathbb{R}_{\geq 0}}$; we denote the persistence diagram of this module by $\mathrm{Dgm}(d_{\mathbb{X}})$ and use $\mathrm{Dgm}_j(d_{\mathbb{X}})$ for the diagrams of the individual modules for each homological dimension j. In Figure 3 (a), we see an example of such an $\mathbb{X}$ embedded in the plane, along with the persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{X}})$. We also have the persistence diagram $\mathrm{Dgm}_1(d_{\mathbb{U}})$, where $\mathbb{U}$ is a dense point sample of $\mathbb{X}$. Note that the two diagrams are quite close in bottleneck distance. Indeed, the difference between the two diagrams will always be upper-bounded by the Hausdorff distance between the space and its sample.

We can also have persistence modules of relative homology groups. For example, referring to the left of Figure 3 (b), we let $\mathbb{X}$ be the space drawn in solid lines and B the closed ball whose boundary is drawn as a dotted circle. By restricting $d_{\mathbb{X}}$ to B and also to ∂B, we produce pairs of sublevel sets $(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)$. Using the maps induced by the inclusions of pairs, we obtain the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}_{\alpha \in \mathbb{R}_{\geq 0}}$ of relative homology groups. The persistence diagram, for homological dimension 1, appears on Figure 3 (b) right. Here, α goes through four non-zero critical values $c_1 < c_2 < c_3 < c_4$ that correspond to the four level sets, where the points in the persistence diagrams (Figure 3 (b) right) correspond to the birth and death of the four relative homology classes respectively.

Stratified spaces. We assume that we have a topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$. A (purely) *d-dimensional stratification of* $\mathbb{X}$ is a decreasing sequence of closed subspaces $\mathbb{X} = \mathbb{X}_d \supseteq \mathbb{X}_{d-1} \supseteq \ldots \mathbb{X}_0 \supseteq \mathbb{X}_{-1} = \emptyset$, such that for each i, the i-dimensional *stratum* $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ is a (possibly empty) i-manifold. The connected components of $\mathbb{S}_i$ are called i-dimensional *pieces*. See Figure 4 (a) for an illustration.

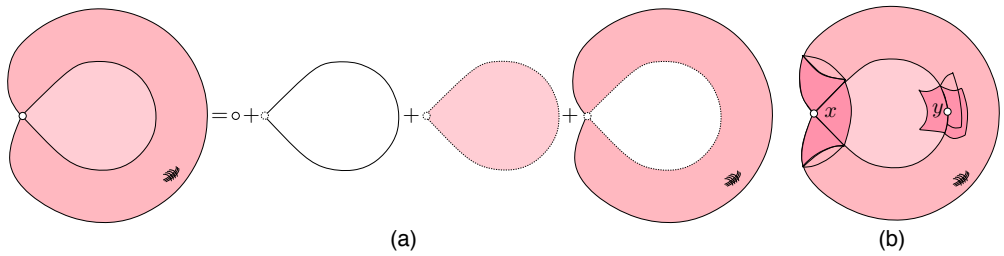

Figure 4: (a) The coarsest stratification of a pinched torus with a spanning disc stretched across the hole. (b) The space in (a) is a cs-space, where the x and y are respectively in the 0-stratum and the 1-stratum, their neighborhoods are highlighted.

One usually also imposes a requirement to ensure that the various pieces fit together uniformly. There are a number of different ways this can be done (see [20] for an extensive survey). For example, one might assume that for each $x \in \mathbb{S}_i$, there exists a small enough neighborhood $N(x) \subseteq \mathbb{X}$ and a $(d-i-1)$-dimensional stratified space L_x such that $N(x)$ is stratum-preserving homeomorphic to the product of an i-ball and the cone on L_x; one can then show that the space L_x depends only on the particular piece containing x. This definition, formally known as a *cs-space*, is illustrated in Figure 4 (b). Since the topology on $\mathbb{X}$ is that inherited from the ambient space, this neighborhood $N(x)$ will take the form $\mathbb{X} \cap B_r(x)$, where $B_r(x)$ is a small enough ball around x in the ambient space.

We note that the above definition requires all strata to be contained within the closure of the top-dimensional stratum. It is also possible, of course, to have spaces where this is not the case: for example, a two-dimensional plane that has been punctured by a line. In this case, a slight adjustment to the above definitions can be made in order to impose similar notions of uniformity.

Local homology and homology stratifications. Recall ([24]) that the local homology groups of a space $\mathbb{X}$ at a point $x \in \mathbb{X}$ are the groups $\mathsf{H}_i(\mathbb{X}, \mathbb{X} - x)$ in each homological dimension i. If $\mathbb{X}$ happens to be a d-manifold, or if x is simply a point in the top-dimensional stratum of a d-dimensional stratification, then these groups are rank one in dimension d and trivial in all other dimensions. On the other hand, the local homology groups for lower-stratum points can be more interesting; for example if x is the crossing point in Figure 3 (b), then $\mathsf{H}_1(\mathbb{X}, \mathbb{X} - x)$ has rank three.

If x and y are close enough points in a particular piece of the same stratum, then there is a natural isomorphism between their local homology groups $\mathsf{H}(\mathbb{X}, \mathbb{X} - x) \cong \mathsf{H}(\mathbb{X}, \mathbb{X} - y)$, which can be understood in the following manner. Taking a small enough radius r and using excision, we see that the two local homology groups in question are in fact just $\mathsf{H}(\mathbb{X} \cap B_r(x), \mathbb{X} \cap \partial B_r(x))$ and $\mathsf{H}(\mathbb{X} \cap B_r(y), \mathbb{X} \cap \partial B_r(y))$. Both of these groups will then map, via intersection of chains, isomorphically into the group $\mathsf{H}(\mathbb{X} \cap B_r(x) \cap B_r(y), \partial(B_r(x) \cap B_r(y))$, and the isomorphism above is then derived from these two maps. In [27], the authors define the concept of a homology stratification of a space $\mathbb{X}$. Briefly, they require a decomposition of $\mathbb{X}$ into pieces such that the locally homology groups are locally constant across each piece; more precisely, that the maps discussed above be isomorphisms for each pair of close enough points in each piece.

3 Topological Inference Theorem

From the discussion above, it is easy to see that any stratification of a topological space will also be a homology stratification. The converse is unfortunately false. However, we can build a useful analytical tool based on the contrapositive: given two points in a point cloud, we can hope to state, based on their local homology groups and the maps between them, that the two points should not be placed in the same piece of any stratification. To do this, we first adapt the definition of these local homology maps into a more multi-scale and robust framework. More specifically, we introduce a radius parameter r and a notion of local equivalence, $\sim_r$, which allows us to group the points of $\mathbb{X}$, as well as of the ambient space, into strata at this radius scale. We then give the main result of this section: topological conditions under which the point cloud $\mathbb{U}$ can be used to infer the strata at different radius scales.

3.1 Local Equivalence

Local homology intersection map. We assume that we are given some topological space $\mathbb{X}$ embedded in some Euclidean space in $\mathbb{R}^N$. For each radius $r \geq 0$, and for each pair of points $p, q \in \mathbb{R}^N$, we define the following homology map $\phi^{\mathbb{X}}(p, q, r)$:

$$\mathsf{H}(\mathbb{X} \cap B_r(p), \mathbb{X} \cap \partial B_r(p)) \to \mathsf{H}(\mathbb{X} \cap B_r(p) \cap B_r(q), \mathbb{X} \cap \partial(B_r(p) \cap B_r(q))). \tag{1}$$

Intuitively, this map can be understood as taking a chain, throwing away the parts that lie outside the smaller range, and then modding out the new boundary. Alternatively, one may think of it as being induced by a combination of inclusion and excision; for a formal and technical definition, see [4].

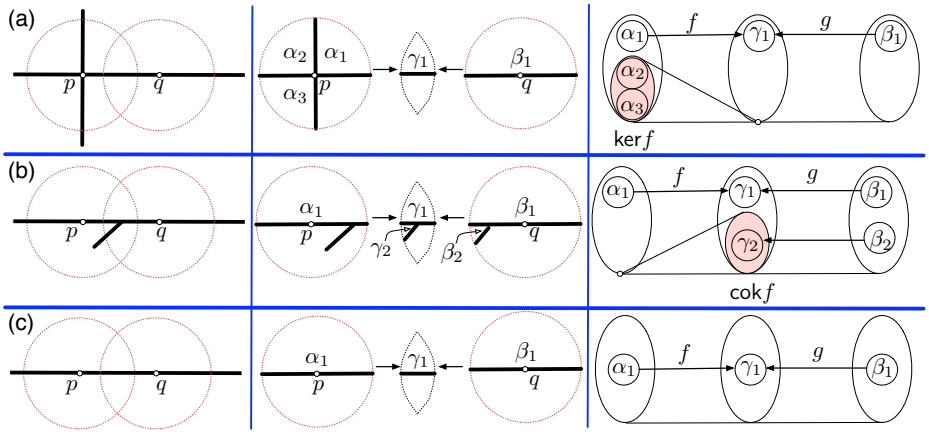

Figure 5: Let $f = \phi^{\mathbb{X}}(p, q, r)$ and $g = \phi^{\mathbb{X}}(q, p, r)$. The local homology classes are labeled in their corresponding locations. (a) p and q do not have the same local structure at radius r since $\ker f \neq 0$. (b) p and q do not have the same local structure at radius r since $\operatorname{cok} f \neq 0$. (c) p and q have the same local structure at radius r since $\ker f = \operatorname{cok} f = 0$ and $\ker g = \operatorname{cok} g = 0$.

For example, consider the space $\mathbb{X}$ drawn in the plane as shown in Figures 5 (a), (b), and (c). For each pair (p, q) of points shown in the three parts of the figure, we let $f = \phi^{\mathbb{X}}(p, q, r)$ and $g = \phi^{\mathbb{X}}(q, p, r)$. Then the points p and q are considered to have the same local structure if f and g are both isomorphisms; equivalently, if $\ker f = \operatorname{cok} f = 0$ and if $\ker g = \operatorname{cok} g = 0$. In part (a), $\ker(f) \neq 0$, since the classes α_2 and α_3 go to zero when passing to the intersection. In part (b), there is a class $\gamma_2 \in \operatorname{cok} f$. The maps f and g in part (c) are both isomorphisms.

Returning to the general case, we use these maps to impose an equivalence relation on $\mathbb{R}^N$.

Definition 3.1 (Local equivalence)**.** *Two points x and y are said to have equivalent local structure at radius r, denoted $x \sim_r y$, iff there exists a chain of points $x = x_0, x_1, \ldots, x_m = y$ from $\mathbb{X}$ such that, for each $1 \leq i \leq m$, the maps $\phi^{\mathbb{X}}(x_{i-1}, x_i, r)$ and $\phi^{\mathbb{X}}(x_i, x_{i-1}, r)$ are both isomorphisms.*

In other words, x and y have the same local structure at this radius iff they can be connected by a chain of points which are pairwise close enough and whose local homology groups at radius r transfer isomorphically into each other via the intersection maps.

Different choices of r will of course lead to different equivalence classes. For example, consider the space $\mathbb{X}$ drawn in the plane as shown in the left half of Figure 6 (a). At the radius drawn, point z is equivalent to the cross point and is not equivalent to either the point x or y. Note that some points from the ambient space will now be considered equivalent to x and y, and some others will be equivalent to z. On the other hand, a smaller choice of radius would result in all three of x, y, and z belonging to the same equivalence class.

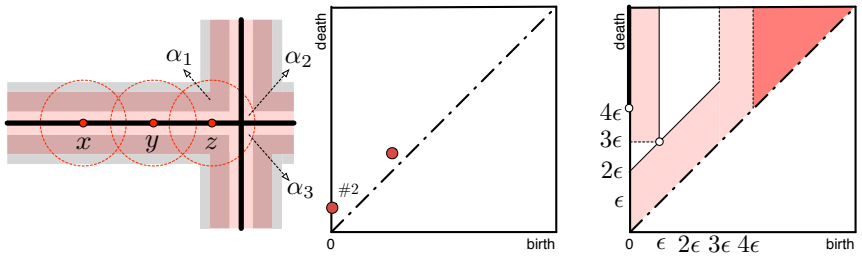

Figure 6: (a) Illustration of equivalence relation: left, $x \sim_r y$, $y \nsim_r z$; right, the 1-dim persistence diagram, for the kernel of the map going from the z ball into its intersection with the y ball. A number, i.e., $\#2$, labeling a point in the persistence diagram indicates its multiplicity. (b) Regions in $\mathbb{X}$-diagrams and $\mathbb{U}$-diagrams. The point in the $\mathbb{X}$-diagrams lie either along the solid black line or in the darkly shaded region. Adding the lightly shaded regions, we get the region of possible points in the $\mathbb{U}$-diagrams.

(Co)Kernel persistence. In order to relate the point cloud $\mathbb{U}$ to the equivalence relation $\sim_r$, we must first define a multi-scale version of the maps $\phi^{\mathbb{X}}(p,q,r)$; we do so by gradually thickening the space $\mathbb{X}$ using the sublevel sets of its distance function. For each $p, q \in \mathbb{R}^N$ and $r, \alpha \geq 0$, we will consider the intersection map $\phi^{\mathbb{X}}_{\alpha}(p,q,r)$, which is defined by substituting $\mathbb{X}_\alpha$ for $\mathbb{X}$ in (1). Note of course that $\phi^{\mathbb{X}}(p,q,r) = \phi^{\mathbb{X}}_0(p,q,r)$.

For the moment, we fix a choice of p, q, and r, and we use the following shorthand, $B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap B_r(p)$, $\partial B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap \partial B_r(p)$, $B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap B_r(p) \cap B_r(q)$, $\partial B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap \partial(B_r(p) \cap B_r(q))$, and we also often write $B^{\mathbb{X}}_p = B^{\mathbb{X}}_p(0)$ and $B^{\mathbb{X}}_{pq} = B^{\mathbb{X}}_{pq}(0)$. By replacing $\mathbb{X}$ with $\mathbb{U}$ in this shorthand, we also write $B^{\mathbb{U}}_p(\alpha) = \mathbb{U}_\alpha \cap B_r(p)$, and so forth.

For any pair of non-negative real values $\alpha \leq \beta$ the inclusion $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ gives rise to the following commutative diagram:

$$\begin{array}{ccc} \mathsf{H}(B^{\mathbb{X}}_p(\alpha), \partial B^{\mathbb{X}}_p(\alpha)) & \xrightarrow{\phi^{\mathbb{X}}_\alpha} & \mathsf{H}(B^{\mathbb{X}}_{pq}(\alpha), \partial B^{\mathbb{X}}_{pq}(\alpha)) \\ \downarrow & & \downarrow \\ \mathsf{H}(B^{\mathbb{X}}_p(\beta), \partial B^{\mathbb{X}}_p(\beta)) & \xrightarrow{\phi^{\mathbb{X}}_\beta} & \mathsf{H}(B^{\mathbb{X}}_{pq}(\beta), \partial B^{\mathbb{X}}_{pq}(\beta)) \end{array} \tag{2}$$

Hence there are maps $\ker \phi^{\mathbb{X}}_\alpha \to \ker \phi^{\mathbb{X}}_\beta$ and $\operatorname{cok} \phi^{\mathbb{X}}_\alpha \to \operatorname{cok} \phi^{\mathbb{X}}_\beta$. Allowing α to increase from 0 to ∞ gives rise to two persistence modules, $\{\ker \phi^{\mathbb{X}}_\alpha\}$ and $\{\operatorname{cok} \phi^{\mathbb{X}}_\alpha\}$, with diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ and $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$. Recall that a homomorphism is an isomorphism iff its kernel and cokernel are both zero. In our context then, the map $\phi^{\mathbb{X}}$ is an isomorphism iff neither $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ nor $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$ contain any points on the y-axis above 0.

Example. As shown in the left part of Figure 6 (a), x, y, and z are points sampled from a cross embedded in the plane. Taking r as drawn, the right part of Figure 6 (a) displays $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$, where $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(z,y,r)$; we now explain this diagram in some detail. The group $\mathsf{H}_1(B^{\mathbb{X}}_z, \partial B^{\mathbb{X}}_z)$ has rank three; as a possible basis we might take the three local homology classes represented by α_1, α_2, and α_3, which are pairs of segments defining the northwest-facing right angle, the northeast-facing right angle, and the southeast-facing right angle. Under the intersection map $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}_0$, the first of these classes α_1 maps to the generator of $\mathsf{H}_1(B^{\mathbb{X}}_{zy}, \partial B^{\mathbb{X}}_{zy})$, while the other two map to zero. Hence $\ker \phi^{\mathbb{X}}_0$ has rank two. As $\mathbb{X}$ starts to thicken into the ambient space, both classes in this kernel eventually die, one at the α value which fills in the northeast corner of the larger ball, and the other at the α value which fills in the southeast corner; these two values are the same here due to symmetry in the picture. At this value,

the map $\phi_\alpha^{\mathbb{X}}$ becomes an isomorphism and it remains so until the intersection of the two balls fills in completely. This gives birth to a new kernel class which subsequently dies when the larger ball finally fills in. The diagram $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$ thus contains three points; the leftmost two show that the map $\phi^{\mathbb{X}}$ is not an isomorphism, and thus that z and y do not have the same local structure at the chosen radius level.

3.2 Topological Inference Theorem

Given a point cloud $\mathbb{U}$ sampled from $\mathbb{X}$, we consider the following question: for a radius r, how can we infer whether or not any given pair of points in $\mathbb{U}$ has the same local structure at this radius? In this subsection, we prove a theorem which describes the circumstances under which we can make the above inference. Naturally, any inference will require that we use $\mathbb{U}$ to judge whether or not the maps $\phi^{\mathbb{X}}(p, q, r)$ are isomorphisms. The basic idea is that if $\mathbb{U}$ is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by $\mathbb{U}$ will be good enough approximations of the diagrams defined by $\mathbb{X}$.

(Co)Kernel stability. Again we fix p, q, and r, and write $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$. For each $\alpha \geq 0$, we let $\mathbb{U}_\alpha = d_{\mathbb{U}}^{-1}[0, \alpha]$. We consider $\phi_\alpha^{\mathbb{U}} = \phi_\alpha^{\mathbb{U}}(p, q, r)$, defined by replacing $\mathbb{X}$ with $\mathbb{U}_\alpha$ in (1), as

$$\mathsf{H}(\mathbb{U}_\alpha \cap B_r(p), \mathbb{U}_\alpha \cap \partial B_r(p)) \to \mathsf{H}(\mathbb{U}_\alpha \cap B_r(p) \cap B_r(q), \mathbb{U}_\alpha \cap \partial(B_r(p) \cap B_r(q))). \tag{3}$$

Running α from 0 to ∞, we obtain two more persistence modules, $\{\ker \phi_\alpha^{\mathbb{U}}\}$ and $\{\mathrm{cok}\, \phi_\alpha^{\mathbb{U}}\}$, with diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})$.

If $\mathbb{U}$ is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by $\mathbb{U}$ will be good approximations of the diagrams defined by $\mathbb{X}$. More precisely, we have the following theorem,

Theorem 3.1 ((Co)Kernel Diagram Stability)**.** *The bottleneck distances between the (co)kernel diagrams of $\phi^{\mathbb{U}}$ and $\phi^{\mathbb{X}}$ are upper-bounded by the Hausdorff distance between $\mathbb{U}$ and $\mathbb{X}$:*

$$d_B(\mathrm{Dgm}(\ker \phi^{\mathbb{U}}), \mathrm{Dgm}(\ker \phi^{\mathbb{X}})) \leq d_H(\mathbb{U}, \mathbb{X}), \quad d_B(\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}), \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{X}})) \leq d_H(\mathbb{U}, \mathbb{X}).$$

Proof. We prove the first inequality; the proof of the second is identical. Put $\epsilon = d_H(\mathbb{U}, \mathbb{X})$. Then, for each $\alpha \geq 0$, the inclusions $\mathbb{U}_\alpha \hookrightarrow \mathbb{X}_{\alpha+\epsilon}$ and $\mathbb{X}_\alpha \hookrightarrow \mathbb{U}_{\alpha+\epsilon}$ induce maps $\ker \phi_\alpha^{\mathbb{U}} \to \ker \phi_{\alpha+\epsilon}^{\mathbb{X}}$ and $\ker \phi_\alpha^{\mathbb{X}} \to \ker \phi_{\alpha+\epsilon}^{\mathbb{U}}$. These maps clearly commute with the module maps in the needed way, and hence we have the required ϵ-interleaving and can thus appeal to Theorem 2.1. □

Main inference result. We now suppose that we have a point sample $\mathbb{U}$ of a space $\mathbb{X}$, where the Hausdorff distance between the two is no more than some ε. In this case, we call $\mathbb{U}$ an ε-approximation of $\mathbb{X}$. Given two points $p, q \in U$ and a fixed radius r, we set $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$, and we wish to determine whether or not $\phi^{\mathbb{X}}$ is an isomorphism. Since we only have access to the point sample $\mathbb{U}$, we instead compute the diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})$; we provide an algorithm for doing this in Section 5.

Given any persistence diagram $\mathcal{D}$, which we recall is a multi-set of points in the extended plane, and two positive real numbers $a < b$, we let $\mathcal{D}(a, b)$ denote the intersection of $\mathcal{D}$ with the portion of the extended plane which lies above $y = b$ and to the left of $x = a$; note that these points correspond to classes which are born no later than a and die no earlier than b.

For a fixed choice of p, q, r, we consider the following two persistence modules: $\{\mathsf{H}(B_p^{\mathbb{X}}(\alpha), \partial B_p^{\mathbb{X}}(\alpha))\}$ and $\{\mathsf{H}(B_{pq}^{\mathbb{X}}(\alpha), \partial B_{pq}^{\mathbb{X}}(\alpha))\}$. We let $\sigma(p, r)$ and $\sigma(p, q, r)$ denote their respective feature sizes and then set $\rho(p, q, r)$ to their minimum. Geometrically, $\rho(p, q, r)$ is related to a local *reach* and the gradient of $d_{\mathbb{X}}$ (as detailed in [4]).

We now give the main theorem of this section, which states that we can use $\mathbb{U}$ to decide whether or not $\phi^{\mathbb{X}}(p, q, r)$ is an isomorphism as long as $\rho(p, q, r)$ is large enough relative to the sampling density.

Theorem 3.2 (Topological Inference Theorem)**.** *Suppose that we have an ε-approximation $\mathbb{U}$ from $\mathbb{X}$. Then for each pair of points $p, q \in \mathbb{R}^N$ such that $\rho = \rho(p, q, r) \geq 4\varepsilon$, the map $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$ is an isomorphism iff* $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})(\varepsilon, 3\varepsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})(\varepsilon, 3\varepsilon) = \emptyset$.

Proof. To simplify exposition, we will refer to points in $\mathrm{Dgm}(\ker \phi^{\mathbb{X}}) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{X}})$ and $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})$ as $\mathbb{X}$-points and $\mathbb{U}$-points, respectively.

Whenever $0 < \alpha < \beta < 4\epsilon < \rho$, the two vertical maps in diagram (2) will by definition both be isomorphisms. This is evidently an immediate consequence of the definition of the feature size. Hence the maps $\ker \phi^{\mathbb{X}}_{\alpha} \to \ker \phi^{\mathbb{X}}_{\beta}$ and $\mathrm{cok}\, \phi^{\mathbb{X}}_{\alpha} \to \mathrm{cok}\, \phi^{\mathbb{X}}_{\beta}$ must also be isomorphisms, and so, as α increases from 0 to ∞, any element of the (co)kernel of $\phi^{\mathbb{X}}$ must live until at least 4ϵ, and any (co)kernel class which is born after 0 must in fact be born after 4ϵ. In other words, any $\mathbb{X}$-point must lie either to the right of the line $x = 4\epsilon$, or along the y-axis and above the point $(0, 4\epsilon)$; see Figure 6 (b). Recall that $\phi^{\mathbb{X}}$ is an isomorphism iff $\ker \phi^{\mathbb{X}} = 0 = \mathrm{cok}\, \phi^{\mathbb{X}}$. Thus $\phi^{\mathbb{X}}$ is an isomorphism iff the black line in Figure 6 (b) contains no $\mathbb{X}$-points.

On the other hand, Theorem 3.1 requires that every $\mathbb{U}$-point must lie within ϵ of an $\mathbb{X}$-point. That is, all $\mathbb{U}$-points are contained within the two lightly shaded regions drawn in Figure 6 (b). Since the rightmost such region is more than ϵ away from the thick black line, there will be a $\mathbb{U}$-point in the left region iff there is an $\mathbb{X}$-point on the thick black line. But the $\mathbb{U}$-points within the left region are exactly the members of $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})(\epsilon, 3\epsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})(\epsilon, 3\epsilon)$.

□

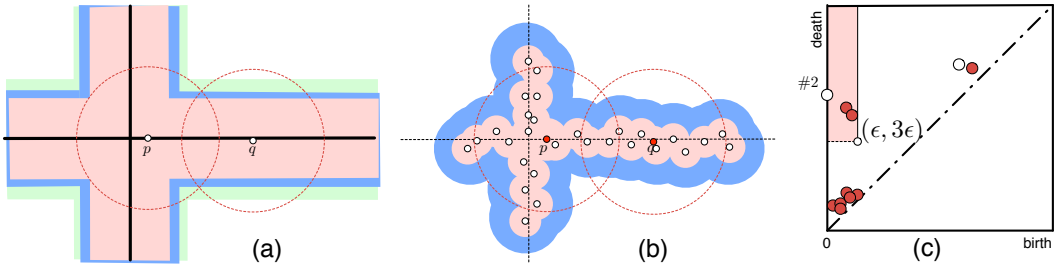

Figure 7: (a) Space $\mathbb{X}$ is the black cross, with p, q, r as drawn. (b) Space $\mathbb{U}$ is an ε-approximation of $\mathbb{X}$. (c) The kernel persistence diagram of $\phi^{\mathbb{X}}(p, q, r)$ ($\mathbb{X}$-diagram) is shown to contain black empty points; the kernel diagram of $\phi^{\mathbb{U}}(p, q, r)$ ($\mathbb{U}$-diagram) is shown to contain red filled points. Suppose only $\mathbb{U}$ is given, we can use $\mathbb{U}$-diagram to infer that points p and q are not locally equivalent.

Figure 6 (b) illustrates Theorem 3.2, that is, under certain topological conditions, $\phi^{\mathbb{X}}$ is an isomorphism if and only if certain regions in the $\mathbb{U}$-diagrams are empty. For example, suppose $\mathbb{X}$ is the cross shown in Figure 7 (a), with p, q, r as drawn. p and q are locally different at this radius level, as shown by the presence of two black empty points on the y-axis of the kernel persistence $\mathbb{X}$-diagram (Figure 7 (c)). Suppose $\mathbb{X}$ is unknown and we are only given $\mathbb{U}$, an ε-approximation of $\mathbb{X}$ (Figure 7 (b)). From the kernel $\mathbb{U}$-diagram, which has two points in the relevant rectangle, we can infer that p and q do not have the same local structure at radius level r by applying Theorem 3.2.

4 Probabilistic Inference Theorem

The topological inference of Section 3 states conditions under which the point sample $\mathbb{U}$ can be used to infer stratification properties of the space $\mathbb{X}$. The basic condition is that the Hausdorff distance between the two must be small. In this section we describe a probabilistic model for generating the point sample $\mathbb{U}$, and we provide an estimate of how large this point sample should be to infer stratification properties of the space $\mathbb{X}$ with a quantified measure of confidence. More specifically, we provide a local estimate, based on $\rho(p, q, r)$ and $\rho(q, p, r)$, of how many sample points are needed to infer the local relationship at radius level r between two fixed points p and q; this same theorem can be used to give a global estimate of the number of points needed for inference between any pair of points whose ρ-values are above some fixed low threshold.

Sampling strategy. We assume $\mathbb{X}$ to be compact. Since the stratified space $\mathbb{X}$ can contain singularities and maximal strata of varying dimensions, some care is required in the sampling design. Consider for example a sheet of area one, punctured by a line of length one. In this case, sampling from a naively constructed uniform measure on this space would result in no points being drawn from the line. This same issue arose and was dealt with in [26], although in a slightly different approach than we will develop.

A sampling strategy that will deal with the problem of varying dimensions is to use a mixture model. In the example of the sheet and line, a uniform measure would be placed on the sheet, while another uniform measure would be placed on the line, and a mixture probability is placed on the two measures; for example, each measure could be drawn with probability $1/2$. We now formalize this approach. Consider each (non-empty) i-dimensional stratum $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ of $\mathbb{X}$. All strata that are included in the closure of some higher-dimensional strata, in other words all non-maximal strata, are not considered in the model. A uniform measure is assigned to the closure of each maximal stratum, $\mu_i(\mathbb{S}_i)$, this is possible since each such closure is compact. We assume a finite number of maximal strata K and assign to the closure of each such stratum a probability $p_i = 1/K$. This implies the following density $f(x) = \frac{1}{K}\sum_{j=1}^{K} \nu_i(X = x)$, where ν_i is the density corresponding to measure μ_i. The point sample is generated from the following model: $\mathbb{U} = \{x_1, ..., .x_n\} \overset{iid}{\sim} f(x)$. We call this model M.

Lower bounds on the sample size of the point cloud. Our main theorem is the probabilistic analogue of Theorem 3.2. An immediate consequence of this theorem is that, for two points $p, q \in \mathbb{U}$, we can infer with probability at least $1 - \xi$ whether p and q are locally equivalent, $p \sim_r q$. The confidence level $1 - \xi$ will be a monotonic function of the size of the point sample. The theorem involves a parameter $v(\rho)$, for each positive ρ, which is based on the volume of the intersection of ρ-balls with $\mathbb{X}$. First we note that each maximal stratum of $\mathbb{X}$ comes with its own notion of volume: in the plane punctured by a line example, we measure volume in the plane and in the line as area and length, respectively. The volume $\mathrm{vol}\,(\mathbb{Y})$ of any subspace $\mathbb{Y}$ of $\mathbb{X}$ is the sum of the volumes of the intersections of $\mathbb{Y}$ with each maximal stratum. For $\rho > 0$, we define $v(\rho) = \inf_{x\in\mathbb{X}} \frac{\mathrm{vol}\,(B_{\rho/32}(x)\cap\mathbb{X})}{\mathrm{vol}\,(\mathbb{X})}$. We then have our main theorem,

Theorem 4.1 (Local Probabilistic Sampling Theorem)**.** *Let $\{x_1, x_2, ..., x_n\}$ be drawn from model M. Fix a pair of points $p, q \in \mathbb{R}^N$ and a positive radius r, and put $\rho = \min\{\rho(p, q, r), \rho(q, p, r)\}$. If $n \geq \frac{1}{v(\rho)}\left(\log\frac{1}{v(\rho)} + \log\frac{1}{\xi}\right)$, then, with probability greater than $1 - \xi$ we can correctly infer whether or not $\phi^{\mathbb{X}}(p, q, r)$ and $\phi^{\mathbb{X}}(q, p, r)$ are both isomorphisms.*

Proof. A finite collection $\mathbb{U} = \{x_1, x_2, ..., x_n\}$ of points in $\mathbb{R}^N$ is *ε-dense* with respect to $\mathbb{X}$ if $\mathbb{X} \subseteq \mathbb{U}^{\varepsilon}$; equivalently, $\mathbb{U}$ is an ε-cover of $\mathbb{X}$. Let $C(\varepsilon)$ be the *ε-covering number* of $\mathbb{X}$, the minimum number of sets $B_{\varepsilon} \cap \mathbb{X}$ that cover $\mathbb{X}$. Let $P(\varepsilon)$ be the *ε-packing number* of $\mathbb{X}$, the maximum number of sets $B_{\varepsilon} \cap \mathbb{X}$ that can be packed into $\mathbb{X}$ without overlap.

We consider a cover of $\mathbb{X}$ with balls of radius $\rho/16$. If there is a sample point in each $\rho/16$-ball, then $\mathbb{U}$ will be an ε-approximation of $\mathbb{X}$, with $\varepsilon \leq 4(\rho/16) = \rho/4$. This satisfies the condition of the topological inference theorem, and therefore we can infer the local structure between p and q.

The following two results from [25] will be useful in computing the number of sample points n needed to obtain, with confidence, such an ε-approximation.

Lemma 4.1 (Lemma 5.1 in [25])**.** *Let $\{A_1, A_2, ..., A_l\}$ be a finite collection of measurable sets with probability measure μ on $\cup_{i=1}^{l} A_i$, such that for all A_i, $\mu(A_i) > \alpha$. Let $\mathbb{U} = \{x_1, x_2, ..., x_n\}$ be drawn iid according to μ. If $n \geq \frac{1}{\alpha}(\log l + \log\frac{1}{\xi})$, then, with probability $1 - \xi$, $\forall i$, $\mathbb{U} \cap A_i \neq \emptyset$.*

Lemma 4.2 (Lemma 5.2 in [25])**.** *Let $C(\varepsilon)$ be the covering number of an ε-cover of $\mathbb{X}$ and $P(\varepsilon)$ be the packing number of an ε-packing, then*

$$P(2\varepsilon) \leq C(2\varepsilon) \leq P(\varepsilon).$$

Again, we consider a cover of $\mathbb{X}$ by balls of radius $\rho/16$. Let $\{y_i\}_{i=1}^{l} \in \mathbb{X}$ be the centers of the balls contained in a minimal sub-cover. Put $A_i = B_{\rho/16}(y_i) \cap \mathbb{X}$. Applying Lemma 4.1, we obtain the estimate

$$n \geq \frac{1}{\alpha}\left(\log l + \log\frac{1}{\xi}\right),$$

where l is the $\rho/16$-covering number, and $\alpha = \min_i \frac{\mathrm{vol}\,(A_i)}{\mathrm{vol}\,(\mathbb{X})}$.

Applying Lemma 4.2,

$$l = C(\rho/16) \leq P(\rho/32) \leq \frac{\mathrm{vol}\,(\mathbb{X})}{\mathrm{vol}\,(B_{\rho/32} \cap \mathbb{X})} \leq \frac{1}{v(\rho)}.$$

On the other hand, $\frac{1}{\alpha} \leq \frac{1}{v(\rho)}$ by definition, and the result follows.

□

To extend the above theorem to a more global result, one can pick a positive ρ and radius r, and consider the set of all pairs of points (p, q) such that $\rho \leq \min\{\rho(p, q, r), \rho(q, p, r\}$. Applying Theorem 4.1 uniformly to all pairs of points will give the minimum number of sample points needed to settle the isomorphism question for all of the intersection maps between all pairs.

5 Algorithm

The theorems in the previous sections give conditions under which a point cloud $\mathbb{U}$, sampled from a stratified space $\mathbb{X}$, can be used to infer the local equivalences between points on $\mathbb{X}$. We now switch gears slightly, and imagine clustering the $\mathbb{U}$-points into strata.

The basic strategy is to build a graph on the point set, with edges corresponding to positive isomorphism judgments. The connected components of this graph will then be our proposed strata.

More precisely, we build a graph where each node in the graph corresponds uniquely to a point from $\mathbb{U}$. Two points $p, q \in \mathbb{U}$ (where $||p - q|| \leq 2r$) are connected by an edge iff both $\phi^{\mathbb{X}}(p, q, r)$ and $\phi^{\mathbb{X}}(q, p, r)$ are isomorphisms, equivalently iff $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})(\epsilon, 3\epsilon)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})(\epsilon, 3\epsilon)$ are empty. The connected components of the resulting graph are our clusters. A more detailed statement of this procedure is giving in pseudo-code, see Algorithm 5.1. Note that the connectivity of the graph is encoded by a weight matrix, and our clustering strategy is based on a $0/1$-weight assignment. We discuss the robustness of our algorithm in a subsequent section.

Algorithm 5.1 Strata-Inference($\mathbb{U}, r, \varepsilon$)

for all $p, q \in \mathbb{U}$ **do**
  if $||p - q|| > 2r$ **then**
    $W(p, q) = 0$
  else
    Compute $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}(p, q, r))$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}(p, q, r))$
    Compute $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}(q, p, r))$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}(q, p, r))$
    if $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}(p, q, r))(\epsilon, 3\epsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}(p, q, r))(\epsilon, 3\epsilon) \neq \emptyset$ **then**
      $W(p, q) = 0$
    else if $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}(q, p, r))(\epsilon, 3\epsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}(q, p, r))(\epsilon, 3\epsilon) \neq \emptyset$ **then**
      $W(p, q) = 0$
    else
      $W(p, q) = 1$
    end if
  end if
end for
Compute connected components based on $\mathbf{W}$.

A crucial subroutine in the clustering algorithm is the computation of the diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})$, for $\phi^{\mathbb{U}} = \phi^{\mathbb{U}}(p, q, r)$ between all pairs $(p, q) \in \mathbb{U} \times \mathbb{U}$. We will focus our attention in this section on the computation of the (co)kernel diagrams.

To compute the diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{U}})$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}})$ we require for each $\alpha \geq 0$ a simplicial analogue of the map $\phi^{\mathbb{U}}_{\alpha} : \mathsf{H}(B^{\mathbb{U}}_{p}(\alpha), \partial B^{\mathbb{U}}_{p}(\alpha)) \to \mathsf{H}(B^{\mathbb{U}}_{pq}(\alpha), \partial B^{\mathbb{U}}_{pq}(\alpha))$. We define, for each $\alpha \geq 0$ (a) two pairs of

simplicial complexes $L_0(\alpha) \subseteq L(\alpha)$ and $K_0(\alpha) \subseteq K(\alpha)$, and (b) a relative homology map between them $\psi_\alpha : \mathsf{H}(L(\alpha), L_0(\alpha)) \to \mathsf{H}(K(\alpha), K_0(\alpha))$. Later, we give a correctness proof that $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}) = \mathrm{Dgm}(\ker \psi)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}) = \mathrm{Dgm}(\mathrm{cok}\, \psi)$.

5.1 Robustness of clustering

Two types of errors in the clustering can occur: false positives where the algorithm connects points that should not be connected and false negatives where points that should be connected are not. The current algorithm we state is somewhat brittle with respect to both false positives as well as false negatives. We will suggest a very simple adaptation of our current algorithm that should be more stable with respect to both false positives and false negatives.

The false positives are driven by the condition in Theorem 3.2 that $\rho < 4\epsilon$, so if the point cloud is not sampled fine enough we can get incorrect positive isomorphisms and therefore incorrect edges in the graph. If we use transitive closure to define the connected components this can be very damaging in practice since a false edge can collapse disjoint components into one large cluster.

The false negatives occur because our point sample $\mathbb{U}$ is not fine enough to capture chains of points that connect pairs in U through isomorphisms, there may be other points in $\mathbb{X}$ which if we had sampled then the chain would be observed. The probability of these events in theory decays exponentially as the sample size increases and the confidence parameter ξ controls these errors.

We now state a simple adaptation of the algorithm that will make it more robust. It is natural to think of the $0/1$-weight assignment on pairs of points $p, q \in \mathbb{U}$ as an association matrix $\mathbf{W}$. A classic approach for robust partitioning is via spectral graph theory [23, 21, 9]. This approach is based an eigen-decomposition of the the graph Laplacian, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with the diagonal matrix $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. The smallest nontrivial eigenvalue λ_1 of $\mathbf{W}$ is called the Fiedler constant and estimates of how well the vertex set can be partitioned [15]. The corresponding eigenvector v_1 is used to partition the vertex set. There are strong connections between spectral clustering and diffusions or random walks on graphs [9].

The problems of spectral clustering and lower dimensional embeddings have been examined from a manifold learning perspective [1, 2, 16]. The idea central to these analyses is given a point sample from a manifold construct an appropriate graph Laplacian and use its eigenvectors to embed the point cloud in a lower dimensional space. A theoretical analysis of this idea involves proving convergence of the graph Laplacian to the Laplace-Beltrami operator on the manifold and the convergence of the eigenvectors of the graph Laplacian to the eigenvalues of the Laplace-Beltrami operator. A key quantity in this analysis is the Cheeger constant which is the first nontrivial eigenvalue of the Laplace-Beltrami operator [8]. An intriguing question is whether the association matrix we construct from the point cloud can be related to the Laplacian on high forms.

5.2 Preliminaries

To construct the simplicial complexes in our algorithm, we will compute Voronoi diagrams and nerves of sets of collections derived from these Voronoi diagrams.

Voronoi diagram. Given a finite collection $\mathbb{U}$ of points in $\mathbb{R}^N$ and $u_i \in \mathbb{U}$, then the *Voronoi cell* of u_i is defined to be:

$$V_i = V(u_i) = \{x \in \mathbb{R}^N \mid ||x - u_i|| \leq ||x - u_j||, \forall u_j \in \mathbb{U}\}.$$

The set of cells V_i covers the entire space and forms the *Voronoi diagram* of $\mathbb{R}^N$, denoted as $\mathrm{Vor}\,(\mathbb{U}|\mathbb{R}^N)$. If we restrict each V_i to some subset $\mathbb{X} \subseteq \mathbb{R}^N$, then the set of cells $V_i \cap \mathbb{X}$ forms a *restricted Voronoi diagram*, denoted as $\mathrm{Vor}\,(\mathbb{U}|\mathbb{X})$. For a simplex σ with vertices in $\mathbb{U}$, we set $V_\sigma = \cap_{u_i \in \sigma} V_i$.

Nerves. The *nerve* $N(\mathcal{C})$ of a finite collection of sets $\mathcal{C}$ is defined to be the abstract simplicial complex with vertices corresponding to the sets in $\mathcal{C}$ and with simplices corresponding to all non-empty intersections among these sets, $N(\mathcal{C}) = \{S \subseteq \mathcal{C} \mid \bigcap S \neq \emptyset\}$. Every abstract simplicial complex can be geometrically realized, and therefore the concept of homotopy type makes sense. Under certain conditions, for example whenever the sets in $\mathcal{C}$ are all closed and convex subsets of Euclidean space ([14], p.59), the nerve of $\mathcal{C}$ has the same homotopy type, and thus

the same homology groups, as the union of sets in $\mathcal{C}$. This implies we can compute $\mathsf{H}(U_\alpha)$, the absolute homology of the thickened point cloud, by computing the nerve of the collection of sets $V_i \cap \mathbb{U}_\alpha$.

The nerve of the restricted Voronoi diagram $\mathrm{Vor}\,(\mathbb{U}|\mathbb{X})$ is called the *restricted Delaunay triangulation*, denoted as $\mathrm{Del}\,(\mathbb{U}|\mathbb{X})$. It contains the set of simplices σ for which $V_\sigma \cap \mathbb{X} \neq \emptyset$.

Power cells, lunes, and moons. We need to compute the relative homology groups $\mathsf{H}(B_p^{\mathbb{U}}(\alpha), \partial B_p^{\mathbb{U}}(\alpha))$ and $\mathsf{H}(B_{pq}^{\mathbb{U}}(\alpha), \partial B_{pq}^{\mathbb{U}}(\alpha))$. The direct argument used to compute absolute homology based on the nerve does not apply to computing relative homology groups since the collection of the sets $V_i \cap \partial B_p^{\mathbb{U}}(\alpha)$ and $V_i \cap \partial B_{pq}^{\mathbb{U}}(\alpha)$ need not be convex.

To get around this problem, we first define the *power cell* with respect to $B_r(p)$, $P(\alpha)$, as $P(\alpha) = \{x \in \mathbb{R}^N \mid ||x-p||^2 - r^2 \leq ||x-u||^2 - \alpha^2, \forall u \in \mathbb{U}\}$, and we set $P_0(\alpha) = B_r(p) - \operatorname{int} P(\alpha)$. Replacing p with q in this formula gives $Q(\alpha)$, the power cell with respect to $B_r(q)$. Finally, we set $Z(\alpha) = P(\alpha) \cap Q(\alpha)$, and $Z_0(\alpha) = (B_r(p) \cap B_r(q)) - \operatorname{int} Z(\alpha)$. These definitions are illustrated in Figure 8 (a). Note that $P_0(\alpha)$ and $Z_0(\alpha)$ are both contained in $\mathbb{U}_\alpha$.

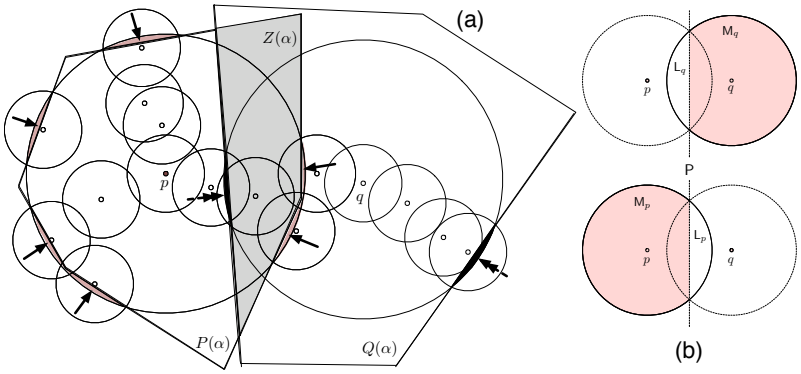

Figure 8: (a) Illustration of intersection power cell $Z(\alpha)$, as the grey shaded region. The unshaded convex regions are $P(\alpha)$ and $Q(\alpha)$ respectively. The dark pink and black shaded regions (pointed by single and double arrows) correspond to $P_0(\alpha)$ and $Q_0(\alpha)$ respectively. (b) Illustration of the lune and the moon. The shaded regions are the respective moons. The white regions within solid circles are the respective lunes.

It turns out that replacing $\partial B_p^{\mathbb{U}}(\alpha)$ with $P_0(\alpha)$ and $\partial B_{pq}^{\mathbb{U}}(\alpha)$ with $Z_0(\alpha)$ has no effect on the relative homology groups in question. That is, the spaces $(B_p^U(\alpha), \partial B_p^U(\alpha))$ and $(B_p^U(\alpha), P_0(\alpha))$ are homotopy equivalent, so are the spaces $(B_{pq}^U(\alpha), \partial B_{pq}^U(\alpha))$ and $(B_{pq}^U(\alpha), Z_0(\alpha))$. Consequently, their homology groups are isomorphic. The first part of this statement was proven in [3], and a proof of the second appears in [4]. The sets $V_i \cap P_0(\alpha)$ are convex [3]. Unfortunately, it is still possible for $V_i \cap Z_0(\alpha)$ to be non-convex, which requires a further subdivision of the Voronoi cells by bisection. Consider the hyperplane P of points in $\mathbb{R}^N$ which are equidistant from p and q. This will divide $\mathbb{R}^N$ into two half-spaces with P_p and P_q denoting the half-spaces containing p and q. Given P_p we define the *p-lune*, L_p, and *p-moon*, M_p, as follows (see Figure 8 (b)): $\mathsf{L}_p = \mathsf{P}_q \cap B_r(p), \mathsf{M}_p = \mathsf{P}_p \cap B_r(p)$.

The lune and the moon divide each Voronoi cell into two parts, $V_i^{\mathsf{L}} = V_i \cap \mathsf{L}_p$ and $V_i^{\mathsf{M}} = V_i \cap \mathsf{M}_p$. These sets are obviously convex, assuming they are non-empty, since they are each the intersection of two convex sets. It also turns out that the non-empty sets among $V_i^{\mathsf{L}} \cap Z_0(\alpha)$ and $V_i^{\mathsf{M}} \cap Z_0(\alpha)$ are convex; see [4] for a proof.

5.3 Algorithm to compute simplicial analogues

Our algorithm to compute simplicial analogues contains two steps: (a) defining the simplicial complexes and (b) defining the corresponding relative homology simplicial maps. We first define the pairs of simplicial complexes $L_0(\alpha) \subseteq L(\alpha)$ and $K_0(\alpha) \subseteq K(\alpha)$. Set $\mathcal{A}$ to be the collection of the non-empty sets among $V_i^{\mathsf{L}} \cap B_p^{\mathbb{U}}(\alpha)$ and $V_i^{\mathsf{M}} \cap B_p^{\mathbb{U}}(\alpha)$. Define $\mathcal{A}_0$ as the collection of the nonempty sets among $V_i^{\mathsf{L}} \cap P_0(\alpha)$ and $V_i^{\mathsf{M}} \cap P_0(\alpha)$. Note that $\cup\mathcal{A} = B_p^{\mathbb{U}}(\alpha)$ and $\cup\mathcal{A}_0 = P_0(\alpha)$. Taking the nerve of both collections, we define the simplicial complexes $L(\alpha) = N(\mathcal{A})$ and $L_0(\alpha) = N(\mathcal{A}_0)$. Similarly, we define $\mathcal{C}$ and $\mathcal{C}_0$ to be the collections of the non-empty sets

among, respectively, $V_i^{\mathsf{L}} \cap B_{pq}^{\mathbb{U}}(\alpha)$ and $V_i^{\mathsf{M}} \cap B_{pq}^{\mathbb{U}}(\alpha)$, and $V_i^{\mathsf{L}} \cap Z_0(\alpha)$ and $V_i^{\mathsf{M}} \cap Z_0(\alpha)$. We define $K(\alpha) = N(\mathcal{C})$ and $K_0(\alpha) = N(\mathcal{C}_0)$. See Figure 9 for an example of the simplicial complexes constructed in $\mathbb{R}^2$ for a given $\mathbb{U}$.

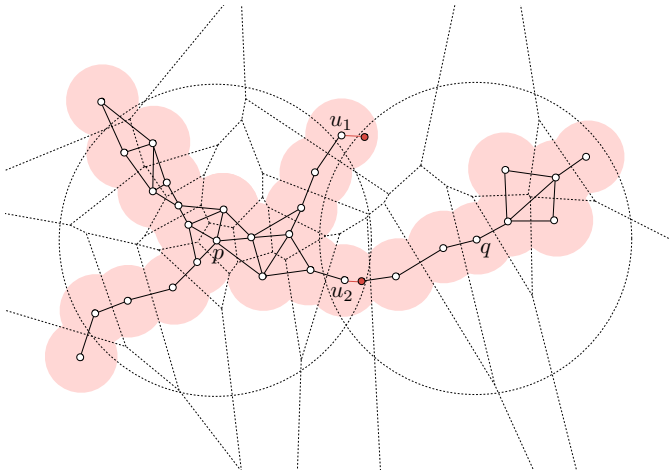

Figure 9: Illustration of the simplicial complexes constructed around two points p and q. The underlying Voronoi decomposition of the space is shown in thin dotted lines. u_1 and u_2 in $\mathbb{U}$ are the points whose restricted Voronoi regions intersect with the lune at non-convex regions.

To define the maps $\psi_\alpha : \mathsf{H}(L(\alpha), L_0(\alpha)) \to \mathsf{H}(K(\alpha), K_0(\alpha))$ we need the following technical lemma:

Lemma 5.1 (Containment Lemma)**.** *Assume that a simplex σ is in $L_0(\alpha)$. If σ is also in $K(\alpha)$, then σ is in $K_0(\alpha)$, as well.*

Proof. Recall that the lune and the moon divide each Voronoi cell into two parts, $V_i^{\mathsf{L}} = V_i \cap \mathsf{L}_p$ and $V_i^{\mathsf{M}} = V_i \cap \mathsf{M}_p$. These are defined as the *partial Voronoi cells*. For simplicity, for a simplex $\sigma \in L(\alpha)$ (similarly for a simplex in L_0, K and K_0), we define V^σ as the intersection of the partial Voronoi cells that correspond to the vertices of σ. That is, $\sigma \in L(\alpha)$ iff $V^\sigma \cap B_p^U(\alpha) \neq \emptyset$. By definition, $\sigma \in L_0(\alpha)$ iff there exists some point $x \in V^\sigma \cap P_0(\alpha)$. We must show that the set $V^\sigma \cap Z_0(\alpha)$ is non-empty. Note that $x \in P_0(\alpha)$ implies that $x \in B_r(p)$, while $x \notin \operatorname{int} P(\alpha)$ implies that $x \notin \operatorname{int} Z(\alpha)$. If $x \in B_r(q)$, then we are done, since $Z_0(\alpha) = B_r(p) \cap B_r(q) - \operatorname{int} Z(\alpha)$.

Otherwise, choose some point $y \in V^\sigma \cap U_\alpha \cap B_r(p) \cap B_r(q)$, which is possible since $\sigma \in K(\alpha)$. Since both x and y belong to the same convex set $V^\sigma \cap U_\alpha \cap B_r(p)$, there exists a directed line segment γ from x to y within this set connecting them. We imagine moving along γ and first we suppose that γ intersects $B_r(q)$ before it intersects $\operatorname{int} Q(\alpha)$. Let z be the first point of intersection. Then $z \in B_r(p) \cap B_r(q)$, $z \notin \operatorname{int} Q(\alpha)$. Therefore $z \in V^\sigma \cap Z_0(\alpha)$. On the other hand, we may prove by contradiction that it is impossible for γ to intersect $Q(\alpha)$ before it intersects $B_r(q)$. Let z' be the first point of such an intersection. Since $z' \in Q(\alpha)$, by definition $||z' - q||^2 - r^2 \leq ||z' - u_i||^2 - \alpha^2, \forall u_i \in \mathbb{U}$. Since $z' \in \mathbb{U}_\alpha, \forall u_i \in \sigma$, $||z' - u_i||^2 - \alpha^2 \leq 0$. Therefore $||z' - q||^2 - r^2 \leq ||z' - u_i||^2 - \alpha^2 \leq 0, \forall u_i \in \sigma$. Since z' is outside $B_r(q)$, $||z' - q||^2 - r^2 > 0$. This is a contradiction. □

To define ψ_α, we first construct a chain map $g = g_\alpha : \mathsf{C}(L(\alpha)) \to \mathsf{C}(K(\alpha))$ as follows. Given a simplex $\sigma \in L(\alpha)$, we define $g(\sigma) = \sigma$ if $\sigma \in K(\alpha)$, and $g(\sigma) = 0$ otherwise; we then extend g to a chain map by linearity. Using the Containment Lemma, we see that $g(\mathsf{C}(L_0(\alpha))) \subseteq \mathsf{C}(K_0(\alpha))$, and thus g descends to a relative chain map $f = f_\alpha : \mathsf{C}(L(\alpha), L_0(\alpha)) \to \mathsf{C}(K(\alpha), K_0(\alpha))$. Since f clearly commutes with all boundary operators, it induces a map on relative homology, this is our $\psi = \psi_\alpha$. To compute the diagrams involving ψ, we reduce various boundary matrices via (co)kernel persistence algorithm described in [12], in time at most cubic in the size of the simplicial complexes representing the data.

Correctness. We show that our algorithm is correct by proving the following theorem. A sketch of the proof is given here, with the details deferred to [4].

Theorem 5.1 (Correctness Theorem)**.** *The persistence diagrams involving simplicial complexes are equal to the persistence diagrams involving the point cloud, that is,* $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}) = \mathrm{Dgm}(\ker \psi)$ *and* $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}) = \mathrm{Dgm}(\mathrm{cok}\, \psi)$.

Proof sketch. To prove Theorem 5.1, we will prove, for each $\alpha \leq \beta$, that the following diagram (as well as a similar diagram involving cokernels) commutes, with the vertical maps being isomorphisms.

$$\begin{array}{ccccc} \ldots & \to & \ker \phi^{\mathbb{U}}_{\alpha} & \to & \ker \phi^{\mathbb{U}}_{\beta} \to \ldots \\ & & \uparrow \cong & & \uparrow \cong \\ \ldots & \to & \ker \psi_{\alpha} & \to & \ker \psi_{\beta} \to \ldots. \end{array} \tag{4}$$

Applying Theorem 2.2 then finishes the claim. Therefore $\mathrm{Dgm}(\ker \phi^{\mathbb{U}}) = \mathrm{Dgm}(\ker \psi)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{U}}) = \mathrm{Dgm}(\mathrm{cok}\, \psi)$.

Simple Simulations. We use a simulation on simple synthetic data with points sampled from grids to illustrate how the algorithm works. We assume we know ε and we run our algorithm for $0 \leq \alpha \leq 3\varepsilon$. As shown in Figure 10 (a), if two points x and y (also, z and w) are locally equivalent, their corresponding kernel and cokernel persistence diagrams shown in Figure 10 (b) contain the empty quadrant predicted by our theorems. On the other hand, if two points x and z are not equivalent, then the kernel persistence diagrams shown in Figure 10 (c) do not contain such empty quadrants.

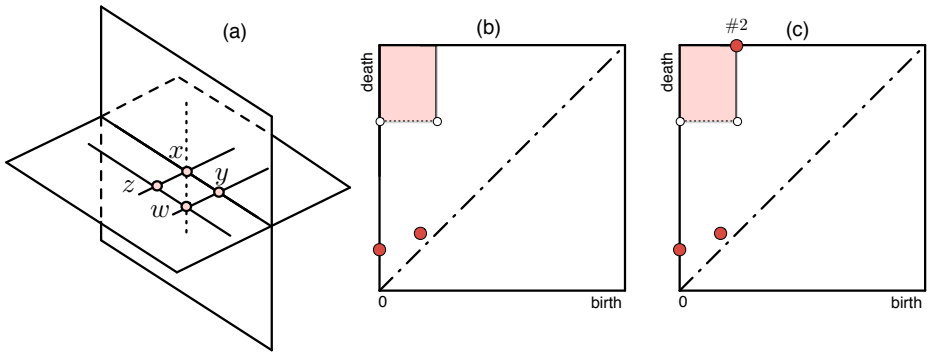

Figure 10: Points sampled from two intersecting planes. Left: points x and y belong to 1-strata, points z and w belong to 2-strata. Middle: kernel and cokernel persistent diagrams with respect to pairs x and y, z and w. Right: kernel persistent diagram with respect to pairs x and z. A number labeling a point in the persistence diagram indicates its multiplicity.

6 Discussion

We have presented a first step towards learning stratified spaces. There are several open issues of interest including: algorithmic efficiency and scaling with dimension using Rips or Witness complexes [13] instead of Delaunay triangulation, robustness of the algorithm and weighting local equivalence, and extensions to the noisy setting [25] when the mixture is concentrated around the stratified space.

Specifically, the algorithm to compute the (co)kernel diagrams from the thickened point cloud is based on an adaption of Delaunay triangulation and the power-cell construction. This algorithm should be quite slow when the dimensionality of the ambient space is high due to the runtime complexity of Delaunay triangulation. One idea to address this bottleneck is to use Rips or Witness complexes [13]. Another approach is to use dimension reduction techniques such as principal components analysis (PCA) or random projection that approximately preserve distance [10] as a preprocessing step. Another idea that may work if the ambient dimension is not too high is using faster algorithms to construct Delaunay triangulations [5].

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2002.

[2] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[3] Paul Bendich, David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy Morozov. Inferring local homology from sampled stratified spaces. In *Proceedings 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 536–546, 2007.

[4] Paul Bendich, Bei Wang, and Sayan Mukherjee. Towards stratification learning through homology inference. http://arxiv.org/abs/1008.3572, August 2010.

[5] Jean-Daniel Boissonnat, Olivier Devillers, and Samuel Hornus. Incremental construction of the delaunay triangulation and the delaunay graph in medium dimension. *Proceedings 25th Annual Symposium on Computational Geometry*, pages 208–216, 2009.

[6] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings 25th Annual Symposium on Computational Geometry*, pages 237–246, 2009.

[7] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in euclidean space. *Discrete and Computational Geometry*, 41:461–479, 2009.

[8] Jeffrey Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Problems in Analysis*, pages 195–199, Princeton, NJ, USA, 1970. Princeton University Press.

[9] Fan R.K. Chung. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, 1997.

[10] Kenneth L. Clarkson. Tighter bounds for random projections of manifolds. *Proceedings 24th Annual Symposium on Computational Geometry*, pages 39–48, 2008.

[11] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, **37**:103–120, 2007.

[12] David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Dmitriy Morozov. Persistence homology for kernels, images and cokernels. *Proceedings 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1011–1020, 2009.

[13] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Symposium on Point-Based Graphics*, pages 157–166, 2004.

[14] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[15] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633, 1975.

[16] Evarist Giné and Vladimir Koltchinskii. Empirical Graph Laplacian Approximation of Laplace-Beltrami Operators: Large Sample Results. In *Lecture Notes-Monograph Series, Vol. 51, High Dimensional Probability*, pages 238–259. Institute of Mathematical Statistics, 2006.

[17] Mark Goresky and Robert MacPherson. *Stratified Morse Theory*. Springer-Verlag, 1988.

[18] Gloria Haro, Gregory Randall, and Guillermo Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Advances in NIPS*, 17, 2005.

[19] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[20] Bruce Hughes and Shmuel Weinberger. Surgery and stratified spaces. *Surveys on Surgery Theory*, pages 311–342, 2000.

[21] R. Kannan, S. Vempala, and A. Veta. On clusterings-good, bad and spectral. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 367, Washington, DC, USA, 2000. IEEE Computer Society.

[22] Gilad Lerman and Teng Zhang. Probabilistic recovery of multiple subspaces in point clouds by geometric lp minimization, 2010.

[23] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *In Advances in Neural Information Processing*, pages 470–477. MIT Press, 2000.

[24] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley, Redwood City, California, 1984.

[25] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Computational Geometry*, 39:419–441, 2008.

[26] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning from noisy data. Manuscript, 2008.

[27] Colin Rourke and Brian Sanderson. Homology stratifications and intersection homology. *Geometry and Topology Monographs*, 2:455–472, 1999.

[28] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1945 – 1959, 2005.

[29] Shmuel Weinberger. *Chicago Lectures in Mathematics*, chapter The topological classification of stratified spaces. University of Chicago Press, Chicago, IL, 1994.

Acknowledgments

All the authors would like to thank Herbert Edelsbrunner and John Harer for useful discussions and suggestions. PB would like to thank David Cohen-Steiner and Dmitrity Morozov for helpful discussion, and SM would like to thank Shmuel Weinberger for useful comments. SM and BW would like to acknowledge the support of NIH Grants R01 CA123175-01A1 and P50 GM 081883, and SM would like to acknowledge the support of NSF Grant DMS-07-32260.