Geometric Reconstruction

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

The lecture note is largely based on [6].

There are two directions for building covers and using their nerves to exhibit the topological structure of data. First is to cover data by balls, and then use distance function frameworks. This leads to geometric inference and providing a framework to establish various theoretical results in Topological Data Analysis. Second is to use a function defined on the data and use Mapper algorithm. This leads to exploratory data analysis and visualization. See Figure 1.

We first recall the cover and the Nerve Theorem.

Definition ([12, Section 26])**.** A collection $\mathcal{A}$ of subsets of a space X is said to cover X, or to be a covering of X, if the union of the elements of $\mathcal{A}$ is equal to X. It is called an open cover of X if its elements are open subsets of X.

We let $\mathcal{U} = \{U_i\}_{i \in I}$ be a cover of $\mathbb{X}$.

Definition. The nerve $Nrv(\mathcal{U})$ of $\mathcal{U}$ is the simplicial complex whose vertices are U_i's and

$$Nrv(\mathcal{U}) := \left\{ \{U_0, \ldots, U_k\} \in \mathcal{U} : \bigcap_{i=0}^{k} U_i \neq \emptyset \right\}. \tag{1}$$

Given a cover of a data set, where each set of the cover can be, for example, a local cluster or a grouping of data points sharing some common properties, its nerve provides a compact and global combinatorial description of the relationship between these sets through their intersection patterns. See Figure 2.

The topology of the nerve is linked to underlying continuous spaces via Nerve Theorem. Under some assumptions, the nerve of a cover is homotopic equivalent to the topology of the union of sets of the cover by the following Nerve Theorem.

Theorem (Nerve Theorem [8, Corollary 4G.3][7, Section III.2])**.** *Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of a space $\mathbb{X}$ such that for any finite subset $\{U_0, \ldots, U_k\} \subset \mathcal{U}$, the intersection $\bigcap_{i=0}^{k} U_i$ is either empty or contractible. Then, the nerve $Nrv(\mathcal{U})$ is homotopic equivalent to $\mathbb{X}$.*

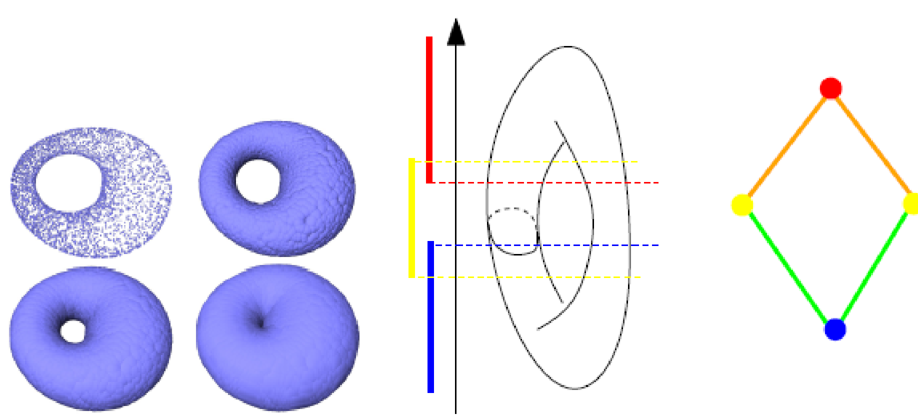

Figure 1: [1] Covering data by balls, and then use distance function frameworks (left). Using a function defined on the data and using Mapper algorithm (right).

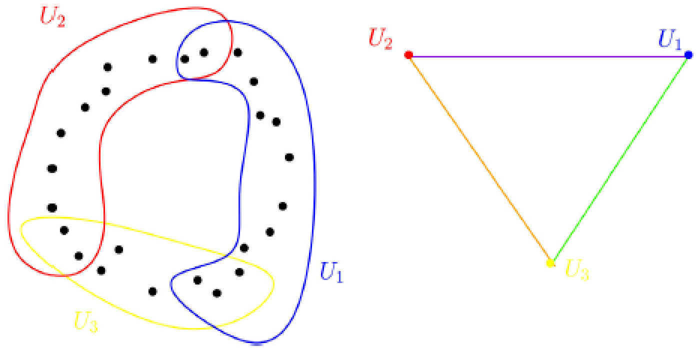

Figure 2: [6, Figure 3] Point cloud and an open cover (left), and the nerve of this cover (right).

This lecture focuses on using distance functions to do geometric inference. In this lecture note, $\mathbb{X} \subset \mathbb{R}^d$ is the target geometric structure, and $\mathcal{X} \subset \mathbb{X}$ is the data points. The general strategy to infer topological information about $\mathbb{X}$ from $\mathcal{X}$ proceeds in two steps:

1. $\mathcal{X}$ is covered by a union of balls of a fixed radius centered on the x_i's. Under some regularity assumptions on $\mathbb{X}$, one can relate the topology of this union of balls to $\mathbb{X}$.

2. From a practical and algorithmic perspective, topological features of $\mathbb{X}$ are inferred from the nerve of the union of balls, using the Nerve theorem.

We compare spaces up to homotopy equivalence:

Definition ([8, Chapter 0])**.** Let $f_0, f_1 : X \to Y$. A *homotopy* between f_0 and f_1 is a continuous function $F : X \times [0,1] \to Y$ such that for all $x \in X$, $F(x,0) = f_0(x)$ and $F(x,1) = f_1(x)$. Two functions f_0, f_1 are *homotopic* if such F exists, and we write $f_0 \simeq f_1$.

Definition ([8, Chapter 0])**.** A map $f : X \to Y$ is called a homotopy equivalence if there is a map $g : Y \to X$ such that $f \circ g \simeq id_Y$ and $g \circ f \simeq id_X$. The space X and Y are said to be homotopy equivalent or to have the same homotopy type, and write $X \simeq Y$, if such homotopy equivalence $f : X \to Y$ exists.

Distance function

Definition. Given a closed subset $A \subset \mathbb{R}^d$, the distance function d_A to A is the non-negative function defined by (see Figure 3)

$$d_A(x) \coloneqq \inf_{y \in A} d(x,y) \text{ for all } x \in \mathbb{R}^d.$$

The distance function to A is continuous and indeed 1-Lipschitz: for all $x, y \in \mathbb{R}^d$,

$$|d_A(x) - d_A(y)| \leq d(x,y).$$

Moreover, A is completely characterized by d_A since $A = d_A^{-1}(0)$.

Definition. For any non-negative real number r, the r-offset A^r of A is the r-sublevel set of d_A defined by (see Figure 3)

$$A^r = d_A^{-1}([0,r]) = \{x \in \mathbb{R}^d : d_A(x) \leq r\}.$$

Now Recall the Hausdorff distance:

Definition (Hausdorff distance [2, Definition 7.3.1])**.** Let $\mathbb{X}$ be a metric space, and $A, B \subset \mathbb{X}$ be a subset. The *Hausdorff distance* between A and B, denoted by $d_H(A,B)$, is defined as

$$d_H(A,B) \coloneqq \inf \{r > 0 : A \subset B^r \text{ and } B \subset A^r\}.$$

Indeed, the Hausdorff distance can be expressed in various equivalent ways in terms of distance functions.

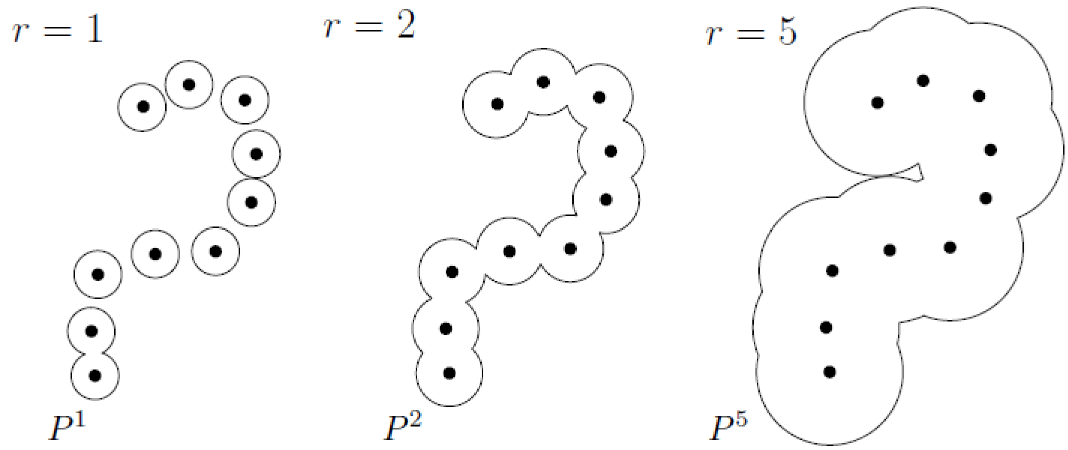

Figure 3: [3] distance function d_P and offsets P^r.

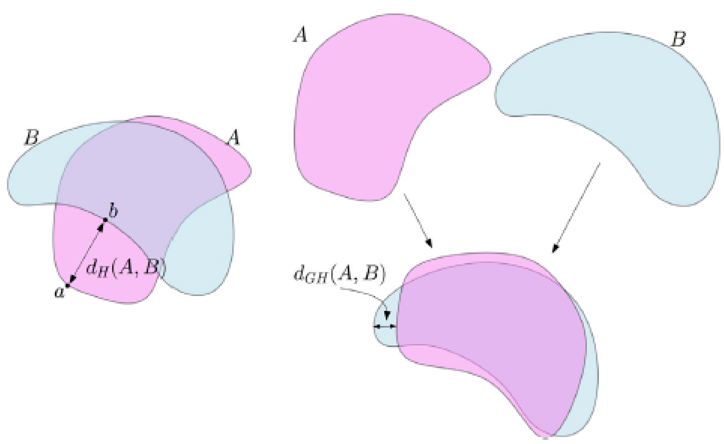

Figure 4: [6, Figure 1] Hausdorff distance $d_H(A, B)$ (left) and Gromov-Hausdorff distance $d_{GH}(A, B)$ between A and B.

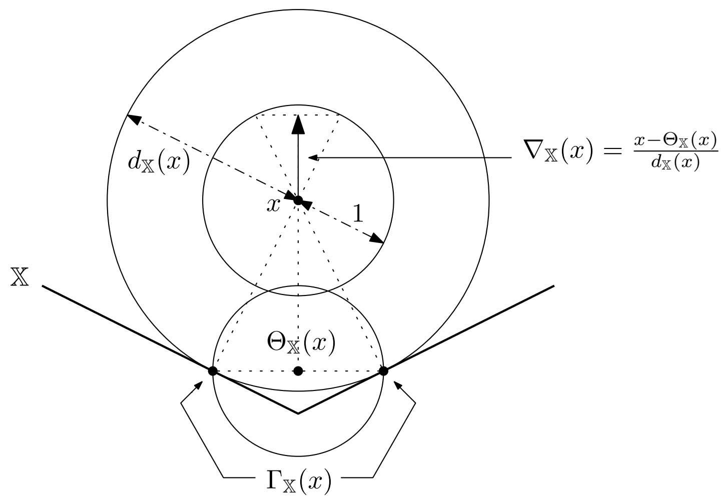

Figure 5: The graphical illustration for the generalized gradient $\nabla_A(x)$, from [4].

Proposition. *Let $A, B \subset \mathbb{R}^d$ be two closed sets.The Hausdorff distance $d_H(A, B)$ between A and B is defined by any of the following equivalent assertions:*

1. $d_H(A, B)$ is the smallest number r such that $A \subset B^r$ and $B \subset A^r$.
2. $d_H(A, B) = \max\{\sup_{x\in A} d_B(x), \sup_{x\in B} d_A(x)\}$.
3. $d_H(A, B) = \|d_A - d_B\|_\infty$.

Given a closed set $A \subset \mathbb{R}^d$, the distance function d_A is usually not differentiable. Nevertheless, it is possible to define a generalized gradient vector field $\nabla_A : \mathbb{R}^d \to \mathbb{R}^d$ for d_A that coincides with the classical gradient at the points where d_A is differentiable. Recall its definition:

For any point $x \in \mathbb{R}^d \setminus A$, let $\Gamma_A(x)$ be the set of points in A closest to x. Let $\Theta_A(x)$ be the center of the unique smallest closed ball enclosing $\Gamma_A(x)$. Then, for $x \in \mathbb{R}^d \setminus A$, the generalized gradient of the distance function d_A is defined as

$$\nabla_A(x) = \frac{x - \Theta_A(x)}{d_A(x)}, \tag{2}$$

and set $\nabla_A(x) = 0$ for $x \in A$. See Figure 5 for a graphical illustration.

The map $x \in \mathbb{R}^d \to \nabla_A(x)$ is in general not continuous. In other words, ∇_A is a discontinuous vector field. Nevertheless, it is possible to show [101, 117] that when K is compact, $x \mapsto \|\nabla_K(x)\|_2$ is a lower semi-continuous function, i.e. for any $a \in \mathbb{R}$, $\|\nabla_K\|_2^{-1}(\infty, a]$ is a cloed subset of $\mathbb{R}^d$, or equivalently, $\liminf_{x\to x_0} \|\nabla_K(x)\|_2 \geq \|\nabla_K(x_0)\|_2$.

Now recall the definition of critical points and weak feature size.

Definition ([4])**.** Let $A \subset \mathbb{R}^d$ be a closed subset.

(a) The critical point of the distance function d_A is defined as the points x for which $\nabla_A(x) = 0$. Equivalently, a point x is a critical point if and only if it lies in the convex hull of $\Gamma_A(x)$. A real $c \geq 0$ is a critical value of d_A if there exists a critical point $x \in \mathbb{R}^d$ such that $d_A(x) = c$. A regular value of d_A is a value which is not critical.

(b) The weak feature size of A, denoted as wfs(A), is the infimum of the positive critical points of d_A. If d_A does not have critical values, then wfs$(A) = \infty$.

Using the notion of critical point, some properties of distance functions are similar to the ones of differentiable functions. In particular, the sublevel sets of d_K are topological submanifolds of $\mathbb{R}^d$ and their topology can change only at critical points.

Theorem. *Let $K \subset \mathbb{R}^d$ be a compact set and let r be a regular value of d_K. The level set $d_K^{-1}(r)$ is a $(d-1)$-dimensional topological submanifold of $\mathbb{R}^d$.*

Theorem (Isotopy Lemma [4, Lemma 2.1]). *Let $K \subset \mathbb{R}^d$ be a compact set and let $r_1 < r_2$ be two real numbers such that $[r_1, r_2]$ does not contain any critical value of d_K. Then all the level sets $d_K^{-1}(r)$, $r \in [r_1, r_2]$ are homeomorphic (and even isotopic) and the set $d_K^{-1}[r_1, r_2]$ is homeomorphic to $d_K^{-1}(r_1) \times [r_1, r_2]$.*

It follows from the Isotopy Lemma that if $0 < \alpha \leq \beta < \mathrm{wfs}(K)$, then K^α and K^β are isotopic. In other words, the knowledge of K at precision, or scale, α gives the same information for any choice of $0 < \alpha < \mathrm{wfs}(K)$.

Deterministic Reconstruction

The following result allows to compare the topology of the offsets of two close compact sets with positive weak feature sizes.

Theorem. *Let $K, K' \subset \mathbb{R}^d$ be two compact sets and $\epsilon > 0$ be such that*

$$d_H(K, K') < \epsilon < \frac{1}{2} \min \{\mathrm{wfs}(K), \mathrm{wfs}(K')\}.$$

Then for any $0 < \alpha \leq 2\epsilon$, K^α and K'^α are homotopy equivalent.

Theorem (Reconstruction Theorem [4, Theorem 4.6][10, Theorem 12]). *Assume $K, K' \subset \mathbb{R}^d$ are compact sets such that K has positive μ-reach $\tau^\mu > 0$ for some $\mu \in (0, 1]$, and that*

$$d_H(K, K') = \epsilon < \frac{\mu^2}{5\mu^2 + 12} \tau^\mu.$$

Then for all $r \in (0, \mathrm{wfs}(K))$ and for all $r' \in \left[\frac{4\epsilon}{\mu^2}, \tau_\mu - 3\epsilon\right)$, $(K')^{r'}$ is homotopy equivalent to K^r.And $(K')^{r'}$ is homotopy equivalent to K as well.

Theorem (Reconstruction Theorem [13, Proposition 7.1][10, Theorem 19, 20]). *Let $\mathbb{X} \subset \mathbb{R}^d$ be a set with positive reach $\tau_{\mathbb{X}} > 0$, and let $\mathcal{X} \subset \mathbb{R}^d$ be a set of points. Let $\delta > 0$ be satisfying $\mathbb{X} \subset \bigcup_{x \in \mathcal{X}} \mathcal{B}(x, \delta)$. Suppose for some constant C, the following is satisfied:*

$$\frac{d_H(\mathbb{X}, \mathcal{X})}{\tau_{\mathbb{X}}} < C.$$

Then there exists some $r > 0$ satisfying that $\mathbb{X}$ is homotopy equivalent to $\check{C}ech_{\mathbb{R}^d}(\mathcal{X}, r)$ or $\mathrm{Rips}(\mathcal{X}, r)$.

$C = 3 - 2\sqrt{2}$ for $\check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathcal{X}, r)$ in [13, Proposition 7.1] and $C = 0.07856\ldots$ for $\mathrm{Rips}(\mathcal{X}, r)$ in [10, Theorem 20].

Probabilistic Reconstruction

Recall that $\mathbb{X} \subset \mathbb{R}^d$ is the target geometric structure, and $\mathcal{X} \subset \mathbb{X}$ is the data points. When $\mathbb{X}$ has a positive reach $\tau_{\mathbb{X}} > 0$, in terms of Reconstruction Theorem, we want to ensure that the Hausdorff distance $d_H(\mathbb{X}, \mathcal{X})$ is small enough with respect to $\tau_{\mathbb{X}}$. In the probabilistic setting where $\mathcal{X}$ is a point cloud of random samples, $d_H(\mathbb{X}, \mathcal{X})$ is also random and can be controlled via the packing number argument.

We assume that $\mathbb{X} \subset \mathbb{R}^d$ is a set with positive reach $\tau_{\mathbb{X}} > 0$, and P is a distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) = \mathbb{X}$. $X_1, \ldots, X_n$ are i.i.d. samples from P and $\mathcal{X} = \{X_1, \ldots, X_n\}$.

First recall the regularity condition on the volume growth imposed by the positive reach.

Proposition ([11, Lemma 25][13, Lemma 5.3][9, Lemma 3]). *Let $M \subset \mathbb{R}^d$ be a k-dimensional submanifold with its reach $\tau_M > 0$. Then for $p \in M$ and $r < \tau_M$, the volume of a ball $\mathrm{vol}_M(M \cap \mathcal{B}(p, r))$ is bounded as*

$$\left(1 - \frac{r^2}{4\tau_M^2}\right)^{\frac{k}{2}} r^k \omega_d \leq \mathrm{vol}_M(M \cap \mathcal{B}(p, r)) \leq \frac{d!}{k!} 2^d r^k \omega_d, \tag{3}$$

where $\omega_d := \lambda_d(\mathcal{B}(0, 1))$ is the volume of a unit ball in $\mathbb{R}^d$.

For a distribution P, we assume (a, b) assumption:

Definition. P satisfies (a, b) assumption if there exists $r_0 > 0$ such that for all $x \in \mathrm{supp}(P)$ and for all $r < r_0$,

$$P(\mathcal{B}(x, r)) \geq a r^b.$$

(a, b) assumption is a weaker than manifold assumption, as implied by the lower bound of the volume growth of (3).

Corollary. *Let $M \subset \mathbb{R}^d$ be a k-dimensional submanifold with its reach $\tau_M > 0$. P is a distribution on $\mathbb{R}^d$ with* $\text{supp}(P) = M$*, and P has a density p with respect to volume measure on M with* $\inf_{x \in M} p(x) > 0$*. Then P satisfies (a, k) assumption.*

We begin with the basic probability lemma.

Lemma ([13, Lemma 5.1])**.** *Let $\{A_i\}$ for $i = 1, \ldots, l$ be a finite collection of measurable sets of $\mathbb{X}$ and let P be a probability measure on $\mathbb{X}$ such that for all $1 \leq i \leq l$, $P(A_i) > \alpha$. Let $X_1, \ldots, X_n$ be i.i.d. from P and $\mathcal{X} = \{X_1, \ldots, X_n\}$. Then*

$$P(\forall i, \mathcal{X} \cap A_i \neq \emptyset) \geq 1 - l \exp(-n\alpha).$$

Proof. Let E_i be the event that $\mathcal{X} \cap A_i = \emptyset$, then

$$P(E_i) = (1 - P(A_i))^n \leq (1 - \alpha)^n.$$

Hence by union bound and $1 - \alpha \leq \exp(-\alpha)$,

$$P\left(\bigcup_{i=1}^{l} E_i\right) \leq \sum_{i=1}^{l} P(E_i) \leq l(1-\alpha)^n \leq l \exp(-n\alpha).$$

And then

$$P(\forall i, \mathcal{X} \cap A_i \neq \emptyset) = 1 - P\left(\bigcup_{i=1}^{l} E_i\right) \geq 1 - l \exp(-n\alpha).$$

□

Now, the idea is to take $A_i = \mathcal{B}(X_i, r)$, and then bound l.

Definition (Covering)**.** Let (X, d) be a metric space and $A \subset X$ be bounded. We say that $\{x_1, \ldots, x_n\} \subset X$ is an ϵ-covering of A if $A \subset \bigcup_{i=1}^{n} \mathcal{B}_d(x_i, \epsilon)$, i.e., for all $x \in A$, there exists x_i such that $d(x, x_i) < \epsilon$. Moreover, we say

$$N(\epsilon) = N(A, \epsilon) = \min\{n : \exists \epsilon\text{-covering of } A \text{ with size } n\}$$

is the covering number of A.

Definition (Packing)**.** Let (X, d) be a metric space and $A \subset X$ be bounded. We say that $\{x_1, \ldots, x_m\} \subset X$ is an ϵ-packing of A if $\left\{\mathcal{B}_d\left(x_i, \frac{\epsilon}{2}\right) : 1 \leq i \leq m\right\}$ are pairwise disjoint, i.e., $d(x_i, x_j) \geq \epsilon$ for all disjoint i, j. Moreover, we say

$$M(\epsilon) = M(A, \epsilon) = \max\{m : \exists \epsilon\text{-packing of } A \text{ with size } m\}$$

is the packing number of A.

Lemma.

$$M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon).$$

Proof. For $M(2\epsilon) \leq N(\epsilon)$, consider an an 2ϵ-packing $\{x_1, \ldots, x_m\}$ and an ϵ-covering $\{y_1, \ldots, y_n\}$. Then if $x_i, x_j \in \mathcal{B}_d(y_k, \epsilon)$ for $i \neq j$, then $d(x_i, x_j) \leq d(x_i, y_k) + d(y_k, x_j) < 2\epsilon$, contradicting that $d(x_i, x_j) \geq 2\epsilon$. Hence each ball $\mathcal{B}_d(y_k, \epsilon)$ can contain at most 1 point among $\{x_1, \ldots, x_m\}$, so $m \leq n$, which implies $M(2\epsilon) \leq N(\epsilon)$.

For $N(\epsilon) \leq M(\epsilon)$, we claim that any maximal ϵ-packing is an ϵ-covering. Let $\{x_1, \ldots, x_n\} \subset X$ be a maximal ϵ-packing, that is, there does not exist $y \in X$ such that $d(y, x_i) \geq \epsilon$ for all $1 \leq i \leq n$. If $A \subset \bigcup_{i=1}^{n} \mathcal{B}_d(x_i, \epsilon)$ does not hold, then $y \in A \backslash \left(\bigcup_{i=1}^{n} \mathcal{B}_d(x_i, \epsilon)\right)$ satisfies that $d(y, x_i) \geq \epsilon$ for all $1 \leq i \leq n$, contradicting the maximality of ϵ-packing. So $A \subset \bigcup_{i=1}^{n} \mathcal{B}_d(x_i, \epsilon)$, and $\{x_1, \ldots, x_n\}$ is also an ϵ-covering, and $N(\epsilon) \leq M(\epsilon)$ holds. □

When A is a manifold, then the packing number is upper bounded by by the lower bound of the volume growth of (3).

Corollary. *Let $\mathbb{X} \subset \mathbb{R}^d$ be a comapct k-dimensional submanifold with its reach $\tau_M > 0$. Then $M(\mathbb{X}, \epsilon) \leq a\epsilon^{-k}$ for some $a > 0$.*

The (a,b) assumption on the distribution also gives an upper bound on the packing number as well.

Corollary ([5, Lemma 10]). *Let P be a distribution on $\mathbb{R}^d$ with* $\mathrm{supp}(P) = \mathbb{X}$*, and assume P satisfies (a, b) assumption with $a, b > 0$. Then there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$,*

$$M(\mathbb{X}, \epsilon) \le a^{-1}\epsilon^{-b}.$$

Proof. Let ϵ_0 be satisfying that for $\epsilon < \epsilon_0$, $P(\mathcal{B}(x, r)) \ge ar^b$ holds. For $\epsilon < \epsilon_0$, let $\{\mathcal{B}(x_i, \epsilon) : 1 \le i \le l\}$ be an ϵ-packing of $\mathbb{X}$. Then $\mathcal{B}(x_i, \epsilon)$'s are disjoint, so

$$1 = P(\mathbb{X}) \ge P\left(\bigcup_{i=1}^{l} \mathcal{B}(x_i, \epsilon)\right) = \sum_{i=1}^{l} P(\mathcal{B}(x_i, \epsilon)) \ge lar^b,$$

so $l \le a^{-1}\epsilon^{-b}$. Since this holds for all ϵ-packing of $\mathbb{X}$, $M(\mathbb{X}, \epsilon) \le a^{-1}\epsilon^{-b}$ as well. □

From the Reconstruction Theorem, a sufficient condition for $\mathbb{X} \simeq \check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathcal{X}, r)$ or $\mathrm{Rips}(\mathcal{X}, r)$ is that $d_H(\mathbb{X}, \mathcal{X}) < C\tau_{\mathbb{X}}$. Hence, the probability of the homotopy equivalence can be lower bounded by the probability of Hausdorff distance $d_H(\mathbb{X}, \mathcal{X})$ being small.

Proposition ([13, Proposition 7.2][5, Theorem 2]). *Let P be a distribution on $\mathbb{R}^d$ with* $\mathrm{supp}(P) = \mathbb{X}$*, and assume P satisfies (a, b) assumption with $a, b > 0$. Let $X_1, \ldots, X_n$ be i.i.d. samples from P, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Then there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$,*

$$P(d_H(\mathbb{X}, \mathcal{X}) < \epsilon) \ge 1 - a^{-1}\epsilon^{-b}\exp(-na\epsilon^b).$$

Proof. Let ϵ_0 be satisfying that for $\epsilon < \epsilon_0$, $P(\mathcal{B}(x, r)) \ge ar^b$ holds. For $\epsilon < \epsilon_0$, let $\{\mathcal{B}(x_i, \epsilon) : 1 \le i \le l\}$ be an ϵ-covering of $\mathbb{X}$, then

$$l \le N(\mathbb{X}, \epsilon) \le M(\mathbb{X}, \epsilon) \le a^{-1}\epsilon^{-b}.$$

Now, $\forall i, \mathcal{X} \cap \mathcal{B}(x_i, \epsilon) \ne \emptyset$ implies that $d_H(\mathbb{X}, \mathcal{X}) < \epsilon$. So from the basic probability lemma,

$$\begin{aligned} P(d_H(\mathbb{X}, \mathcal{X}) < \epsilon) &\ge P(\forall i, \mathcal{X} \cap \mathcal{B}(x_i, \epsilon) \ne \emptyset) \\ &\ge 1 - a^{-1}\epsilon^{-b}\exp(-na\epsilon^b). \end{aligned}$$

□

Theorem (Reconstruction Theorem [13, Theorem 7.1][10, Theorem 19, 20]). *Let $\mathbb{X} \subset \mathbb{R}^d$ be a compact subset with positive reach $\tau_{\mathbb{X}} > 0$. Let P be a distribution on $\mathbb{R}^d$ with* $\mathrm{supp}(P) = \mathbb{X}$*, and assume P satisfies (a, b) assumption with $a, b > 0$. Let $X_1, \ldots, X_n$ be i.i.d. samples from P, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Then there exists some $r > 0$ satisfying that*

$$P\left(\mathbb{X} \simeq \check{C}ech_{\mathbb{R}^d}(\mathcal{X}, r) \text{ and } Rips(\mathcal{X}, r)\right) \ge 1 - C\exp(-nC),$$

where C depends only on on $\tau_{\mathbb{X}}, a, b$.

Proof. As soon as $d_H(\mathbb{X}, \mathcal{X}) < C'\tau_{\mathbb{X}}$, there exists some $r > 0$ satisfying that $\mathbb{X}$ is homotopy equivalent to $\check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathcal{X}, r)$ and $\mathrm{Rips}(\mathcal{X}, r)$. Hence from above,

$$\begin{aligned} P\left(\mathbb{X} \simeq \check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathcal{X}, r) \text{ and } \mathrm{Rips}(\mathcal{X}, r)\right) &\ge P(d_H(\mathbb{X}, \mathcal{X}) < C'\tau_{\mathbb{X}}) \\ &\ge 1 - a^{-1}(C'\tau_{\mathbb{X}})^{-b}\exp(-na(C'\tau_{\mathbb{X}})^b). \end{aligned}$$

□

References

[1] Jean-Daniel Boissonnat, Frédéric Cazals, Frédéric Chazal, and Julien Tierny. https://geometrica.saclay.inria.fr/team/fred.chazal/sophia2017/tdasophia2017.html, 2017.

[2] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[3] Frédéric Chazal. https://geometrica.saclay.inria.fr/team/fred.chazal/m2orsay2023.html, 2023.

[4] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in euclidean space. *Discrete & Computational Geometry*, 41(3):461–479, 2009.

[5] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015.

[6] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers Artif. Intell.*, 4:667963, 2021.

[7] Herbert Edelsbrunner and John L. Harer. *Computational topology.* American Mathematical Society, Providence, RI, 2010. An introduction.

[8] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.

[9] Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax rates for estimating the dimension of a manifold. *J. Comput. Geom.*, 10(1):42–95, 2019.

[10] Jisu Kim, Jaehyeok Shin, Frédéric Chazal, Alessandro Rinaldo, and Larry Wasserman. Homotopy Reconstruction via the Cech Complex and the Vietoris-Rips Complex. In Sergio Cabello and Danny Z. Chen, editors, *36th International Symposium on Computational Geometry (SoCG 2020)*, volume 164 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 54:1–54:19, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[11] Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry A. Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3398–3407. PMLR, 2019.

[12] James R. Munkres. *Topology.* Prentice Hall, Inc., Upper Saddle River, NJ, 2000. Second edition of [MR0464128].

[13] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.