Confidence sets for persistent homology of the KDE filtration

Jaehyeok Shin
Department of Statistics
Carnegie Mellon University
jaehyeos@andrew.cmu.edu

Jisu Kim
Department of Statistics
Carnegie Mellon University
jisuk1@andrew.cmu.edu

Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University
arinaldo@cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
larrywasserman.cool@gmail.com

Abstract

When we observe a point cloud in the Euclidean space, the persistent homology of the upper level sets filtration of the density is one of the most important tools to understand topological features of the data generating distribution. The persistent homology of KDEs (kernel density estimators) for the density function is a natural way to estimate the target quantity. In practice, however, calculating the persistent homology of KDEs on d-dimensional Euclidean spaces requires a grid-approximation for the ambient space, which is computationally expensive. In this paper, we will consider the persistent homology of KDE filtrations on Rips complexes as suggested by Bobrowski et al. [2014]. We will describe a novel methodology to construct an asymptotic confidence set for the corresponding persistence diagram by using the interleaving distance and the bootstrap. Unlike existing procedures, our method does not heavily rely on grid-approximations and scales to higher dimensions.

1 Introduction

When we observe data from a distribution P, the upper level sets $D_L := \{p \geq L\}$ of the density function p reveal important topological features of the data generating distribution. For instance, density-based clustering methods [Hartigan, 1975, 1981, Cadre, 2006, Rinaldo and Wasserman, 2010] use the information about connected components of a level set to group data points in the hope that points in the same connected component share common characteristics. Rather than choosing a fixed level, a cluster tree [Kim et al., 2016, Eldridge et al., 2015, Balakrishnan et al., 2013, Chaudhuri and Dasgupta, 2010] summarizes the hierarchy of high-density clusters at all levels simultaneously.

We can investigate topological features of level sets by their corresponding homology groups. For example, the 0-th homology group of a level set contains information about connected components in the level set. By using higher order homology groups, we can further characterize each connected components. For instance, the rank of the 1-st homology group of each connected component counts the number of one-dimensional holes.

Since different level sets could show different aspects of the data generating distribution, analyzing a fixed level set might be not enough to understand the overall shape of the distribution. Alternatively, as cluster trees show clusters at all levels, we can investigate changes in shapes by looking at all

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

possible level sets simultaneously,

$$\{D_L\}_{L\in\mathbb{R}}. \tag{1}$$

Note that $D_{L_1} \subset D_{L_2}$ for any $L_1 \geq L_2$. Thus (1) is called the level sets filtration of the density function.

The persistent homology [Edelsbrunner and Harer, 2010, 2008, Zomorodian and Carlsson, 2005] quantifies topological features at multiple scales by computing a filtration of topological spaces. The persistent homology captures changes of homologies in filtrations simultaneously, see [Fasy et al., 2014, Bobrowski et al., 2014, Phillips et al., 2013, Chung et al., 2009, Bubenik, 2015].

Since the density function is unknown, the persistent homology of the density function needs to be estimated. One approach, as in Fasy et al. [2014], is to replace the level sets of unknown density function by level sets of kernel density estimator (KDE) computed on a grid of points. Another approach, as in Bobrowski et al. [2014], is to use a different approximating filtration rather than using the level sets of KDE.

Both methods have pros and cons. The first approach in Fasy et al. [2014] yields more precise estimation if we use a sufficiently fine grid approximation to calculate the persistent homology of KDEs. The first approach also has several well-studied ways of computing confidence sets which can be used to distinguish signal vs noise topological features, due to the well-known theoretical behavior of the KDE. However, a fine grid approximation could be computational intractable when the ambient space has large dimension. In contrast, the second approach in Bobrowski et al. [2014] is computationally more efficient, in particular in higher dimension, because the persistent homology is calculated on the data points only and it does not require a grid-approximation of the ambient space. This feature of the second approach makes it possible to capture heterogeneous topological features efficiently, and makes it easy to apply the method to more general settings. However, the asymptotic behavior of the approximating filtration is more complicated, and hence confidence sets for Bobrowski et al. [2014] are not yet well-studied.

The goal of this paper is to use the estimator of the persistent homology of KDE filtrations on Rips complexes proposed by Bobrowski et al. [2014] in order to construct a bootstrap-based confidence set for the corresponding persistence diagram. Our method does not heavily rely on grid-approximations and scales to higher dimensions.

1.1 Notation

Throughout the paper, we let $\mathbb{X} = \mathbb{R}^d$. We denote by $\mathcal{B}(x, r)$ the closed ball of radius $r > 0$ and center $x \in \mathbb{R}^d$. Let $\mathcal{X}_n := \{X_1, \ldots, X_n\} \subset \mathbb{X}$ be the observed data. For a function $f : \mathbb{X} \to \mathbb{R}$ and a value $L \in \mathbb{R}$, we let $\mathbb{X}_L^f := \{x \in \mathbb{X} : f(x) \geq L\}$ be the upper level set of f at level L.

2 Background

This section is a brief introduction to homology and persistent homology. We mainly follow Fasy et al. [2014] and Bobrowski et al. [2014]. We refer to Hatcher [2002] for a comprehensive explanation of homology, and to Edelsbrunner and Harer [2010] for theory and computation of persistent homology.

2.1 Homology of an abstract simplicial complex K

Definition 1. *Let $V := \{v_0, v_1, \ldots, v_n\}$ be a finite vertex set. An* abstract simplicial complex *K on V is a collection of subsets of V such that*

1. $\emptyset \in K$
2. $\{v\} \in K$ *for* $\forall v \in V$
3. *If $\sigma \in K$ and $\tau \subset \sigma$ then $\tau \subset K$*

An element σ of K is called a *simplex* or a *face*. For each $\sigma \in K$, the dimension of σ is defined by $\dim \sigma := |\sigma| - 1$, and the dimension of K is defined by $\dim K := \max\{\dim \sigma : \sigma \in K\}$.

Definition 2. *A p-chain of K is a formal sum* $\sum_{\sigma \in K_p} c_\sigma \sigma$*, where* $c_\sigma \in \mathbb{Z}_2$*,* K_p *is the collection of dimension* p *simplices.*

Definition 3. *The set of p-chains of a simplicial complex K form a p-chain group*

$$C_p(K) := \left\{ c_1\sigma_1 + \cdots + c_{n_p}\sigma_{n_p} : c_i \in \mathbb{Z}_2, \sigma_i \in K_p, n_p = |K_p| \right\}$$

Definition 4. *A boundary map* $\partial_p : C_p(K) \to C_{p-1}(K)$ *is defined by*

$$\partial_p(\sigma) = \partial_p[v_0, v_1, \ldots, v_p] := \sum_{i=0}^{p} [v_0, \ldots, \widehat{v_i}, \ldots, v_p] \text{ and extend to } C_p(K) \text{ linearly}$$

where $[v_0, \ldots, \widehat{v_i}, \ldots, v_p]$ *is a* $(p-1)$*-dimensional simplex obtained by removing* v_i *in* σ*.*

We call $\partial_p(C)$ $(\in C_{p-1}(K))$ a *boundary* of C $(\in C_p(K))$, and say C is *cycle* if $\partial_p(C) = 0$.

Definition 5. *Let* $Z_p(K) := \ker \partial_p \subset C_p(K)$ *be the p-cycle group and* $B_p(K) := \operatorname{Im} \partial_{p+1} \subset C_p(K)$ *be the p-boundary group. Now, the p-th homology group is defined by*

$$H_p(K) := Z_p(K)/B_p(K)$$

Note that it is well defined since $\partial_p \partial_{p+1} = 0$, and thus $B_p(K) \subset Z_p(K)$. Each homology class $\gamma \in H_p(K)$ corresponds to a p-dimensional cycle. Finally, we define the p-th Betti number by $\beta_p := rk(H_p(K))$. For instance, β_0 is the number of connected component of K, β_1 is the number of one-dimensional holes in K, and β_2 is the number of two-dimensional holes in K.

2.2 Computing homology from point clouds

In many applications, we cannot observe the topological space directly but only obtain (noisy) point clouds from it. Let $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_1, \ldots, r_n > 0$, then one natural way to recover the topological space of interest is to make a union of r_i-balls centered at the data points $X_1, \ldots, X_n$, defined as

$$\bigcup_{i=1}^{n} \mathcal{B}(X_i, r_i). \tag{2}$$

The corresponding simplicial complex is the *Čech complex*.

Definition 6. *Let* $\mathcal{X}_n \subset \mathbb{X}$ *and* $r \in \mathbb{R}^n$ *with* $r_1, \ldots, r_n > 0$*. The Čech complex is the set of simplices with* σ *with vertices* $X_{n_1}, \ldots, X_{n_k} \in \mathcal{X}_n$ *such that*

$$\check{C}ech(\mathcal{X}_n, r) := \left\{ \sigma = [X_{n_1}, \ldots, X_{n_k}] : \bigcap_{i=1}^{k} \mathcal{B}(X_{n_i}, r_{n_i}) \neq \emptyset \right\}$$

The homology of the union of balls in (2) can be computed by the homology of the Čech complex by the following Nerve Theorem.

Lemma 1 (Nerve Theorem)**.** *The Čech complex* $\check{C}ech(\mathcal{X}_n, r)$ *is homotopy equivalent to the union of balls* $\bigcup_{i=1}^{n} \mathcal{B}(X_i, r_i)$*.*

Computing the Čech complex requires checking whether intersections of balls $\mathcal{B}(X_{n_i}, r_{n_i})$ are empty or not. To save on computation time, we may instead check pairwise distances only and add 2- and higher-dimensional simplices whenever we can. This leads to the *Vietoris-Rips complex*.

Definition 7. *The Vietoris-Rips complex* $R(\mathcal{X}_n, r)$ *is defined by*

$$R(\mathcal{X}_n, r) := \left\{ \sigma = [X_{n_1}, \ldots, X_{n_k}] : \|X_{n_i} - X_{n_j}\| \leq r_{n_i} + r_{n_j}, \forall i \neq j \right\}$$

Note that the Čech complex and Vietoris-Rips complex have following interleaving inclusion relationship

$$\check{\text{C}}\text{ech}(\mathcal{X}_n, r) \subset R(\mathcal{X}_n, r) \subset \check{\text{C}}\text{ech}(\mathcal{X}_n, 2r)$$

In particular, when r_i's are all the same, then the constant 2 can be tightened to $\sqrt{2}$:

$$\check{\text{C}}\text{ech}(\mathcal{X}_n, r) \subset R(\mathcal{X}_n, r) \subset \check{\text{C}}\text{ech}(\mathcal{X}_n, \sqrt{2}r)$$

2.3 Persistent homology

Choosing a suitable r in $R(\mathcal{X}_n, r)$ is difficult. Instead of choosing a fixed r, we can use several $R(\mathcal{X}_n, r)$ simultaneously by using the filtration

$$\emptyset = R(\mathcal{X}_n, r^{(0)}) \subset R(\mathcal{X}_n, r^{(1)}) \subset \cdots \subset R(\mathcal{X}_n, r^{(m)}) = R(\mathcal{X}_n, \infty)$$

In general, a *filtration* is a nested sequence of topological spaces.

$$\emptyset = \mathbb{X}_0 \subset \mathbb{X}_1 \subset \cdots \subset \mathbb{X}_m = \mathbb{X}$$

For instance, in many statistical applications, we are interested in the filtration of upper level sets of a density function $f : \mathbb{X} \to \mathbb{R}$ where $\mathbb{X}_i := \mathbb{X}^f_{L_i}$ with $L_0 \geq L_1 \geq \cdots \geq L_m$. Inclusions in a filtration naturally induce maps on the corresponding homology groups :

$$0 = H_p(\mathbb{X}_0) \to H_p(\mathbb{X}_1) \to \cdots, \to H_p(\mathbb{X}_n) = H_p(\mathbb{X}) \tag{3}$$

and also induce a natural group homomorphism $i^p_{s,t} : H_p(\mathbb{X}_s) \to H_p(\mathbb{X}_t)$. In many cases, we can successfully approximate the filtration in (3) by using a suitable filtration of the corresponding Vietoris-Rips complex in Definition 7.

The persistent homology tracks when topological features are born and die. Formally, a homology class $\gamma \in H_p(\mathbb{X}_s)$ is said to be born at $\mathbb{X}_s$ if γ is not in the image of $i^p_{s-1,s}$. The same class γ born at $\mathbb{X}_s$ dies going into $\mathbb{X}_t$ if t is the smallest index such that the class γ is supported in the image of $i^p_{s-1,t}$.

Definition 8. *The* persistent homology *is the finite multi-set of pairs of births and deaths of homology classes. Each pair of birth at s and death at t of homology class γ can be visualized in the p-th* persistence diagram as a point (s, t).

2.4 Stability Theorem for functions

Let $f, g : \mathbb{X} \to \mathbb{R}$ be two functions. The stability theorem asserts that if f and g are close to each other, then their corresponding persistent homologies $\mathrm{PH}_*(f)$ and $\mathrm{PH}_*(g)$ are also close. We first define the distance between two persistent homologies by using the bottleneck distance.

Definition 9. *Let $\mathcal{X}$ be a filtration. The k-th persistence diagram of $\mathcal{X}$, denoted by $Dgm_k(\mathcal{X})$ is the set of all pairs (b, d) of birth-death times of features in $\mathrm{PH}_k(\mathcal{X})$. The bottleneck distance between the persistent homology of the filtrations, $\mathcal{X}, \mathcal{Y}$ is defined by*

$$d_B(\mathrm{PH}_k(\mathcal{X}), \mathrm{PH}_k(\mathcal{Y})) = \inf_{\gamma \in \Gamma} \sup_{p \in Dgm_k(\mathcal{X})} \|p - \gamma(p)\|_\infty,$$

where the set Γ consists of all the bijections $\gamma : Dgm_k(\mathcal{X}) \cup Diag \to Dgm_k(\mathcal{Y})$, and $Diag$ is the diagonal line $\{(x, x) : x \in \mathbb{R}\} \subset \mathbb{R}^2$.

We impose a regularity condition for the functions f and g, which is *tameness*.

Definition 10. *Let $f : \mathbb{X} \to \mathbb{R}$. Then f is* tame *if $H_k(\mathbb{X}_L)$ is of finite rank for all $k \in \mathbb{N} \cup \{0\}$ and $L \in \mathbb{R}$.*

For two tame functions f and g, their bottleneck distance is bounded by their ℓ_∞ distance, which is known as the stability theorem.

Theorem 2 (Stability Theorem)**.** *(Edelsbrunner and Harer [2010], Chazal et al. [2009]) For two tame functions $f, g : \mathbb{X} \to \mathbb{R}$,*

$$d_B(\mathrm{PH}_*(f), \mathrm{PH}_*(g)) \leq \|f - g\|_\infty.$$

The stability theorem bounds the difference between persistent homologies generated from sublevel sets or superlevel sets of functions. When it comes to comparing two persistent homologies that are not necessarily from level sets of functions, we need the Stability Theorem in more general algebraic settings. In Appendix A, we define persistence module, which is an algebraic abstraction of the persistent homology, and then state the Stability Theorems on persistence modules.

3 Definitions and Statistical model

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be an i.i.d. sample from a distribution P with Lebesgue density p. We assume that the density function p satisfy assumption 1:

Assumption 1. *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a density function of P with the following conditions:*

1. *$supp(p)$ is bounded.*
2. *p is tame (see Definition 10).*
3. *$p_{max} := \sup_{x \in \mathbb{X}} p(x) < +\infty$.*

For estimating the density p, we use the kernel density estimator (KDE), which smooths out the empirical measure by a kernel function $K : \mathbb{R}^d \to \mathbb{R}$. Let the kernel function K satisfy Assumption 2:

Assumption 2. *The kernel function $K : \mathbb{R}^d \to \mathbb{R}$ is a nonnegative function with the following conditions:*

1. *$supp(K) \subset \mathcal{B}(0, 1)$,*
2. *$\int K(x)dx = 1$.*

Then for $h > 0$, the kernel density estimator is defined as

$$\widehat{p}_h(x) := \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Let $p_h : \mathbb{R}^d \to \mathbb{R}$ be the pointwise average of the kernel density estimator, i.e. $p_h(x) := \mathbb{E}[\widehat{p}_h(x)]$. Our target of inference is the persistent homology of p_h, denoted as $\mathrm{PH}_*(p_h)$. More formally, Let the upper level set of p_h be

$$D_L := \left\{x \in \mathbb{R}^d : \ p_h(x) \geq L\right\}. \tag{4}$$

Then $\mathrm{PH}_*(p_h)$ is the persistent homology of the filtration

$$\{D_L\}_{L \in \mathbb{R}}. \tag{5}$$

4 Confidence set for persistent homology for filtration on Rips complex

In this section, we build a confidence set of the persistent homology $\mathrm{PH}_*(p_h)$ of the density level set filtration. We first consider the general form of a valid asymptotic confidence set and its equivalent condition, and then we consider two implementations using this general form.

A confidence set of the persistent homology $\mathrm{PH}_*(p_h)$ is a random set of persistent homologies that contains $\mathrm{PH}_*(p_h)$ with some probability. Specifically, for given $\alpha \in (0, 1)$, a valid $1 - \alpha$ level asymptotic confidence set of $\mathrm{PH}_*(p_h)$ is a random set $\widehat{C}_\alpha$ satisfying

$$\limsup_{n \to \infty} P(\mathrm{PH}_*(p_h) \in \widehat{C}_\alpha) \geq 1 - \alpha.$$

We construct the confidence set $\widehat{C}_\alpha$ by considering an appropriate estimator $\widehat{\mathrm{PH}_*(p_h)}$ for the persistent homology $\mathrm{PH}_*(p_h)$, and then consider all persistent homologies within c_n for some $c_n > 0$. Hence the confidence set becomes

$$\widehat{C}_\alpha = \left\{\mathcal{P} : \ d_B\left(\mathcal{P}, \widehat{\mathrm{PH}_*(p_h)}\right) \leq c_n\right\},$$

where both $\widehat{\mathrm{PH}_*(p_h)}$ and radius c_n are functions of $X_1, \ldots, X_n$. Then note that $\mathrm{PH}_*(p_h) \in \widehat{C}_\alpha$ holds if and only if

$$d_B\left(\widehat{\mathrm{PH}_*(p_h)}, \mathrm{PH}_*(p_h)\right) \leq c_n.$$

Hence $\widehat{C}_\alpha$ is a valid $1 - \alpha$ asymptotic confidence set if and only if

$$\limsup_{n \to \infty} P\left(d_B\left(\widehat{\mathrm{PH}_*(p_h)}, \mathrm{PH}_*(p_h)\right) \leq c_n\right) \geq 1 - \alpha.$$

We first construct the persistent homology estimator by using a kernel density estimator as proposed by Bobrowski et al. [2014]. Then we further approximate the estimator by considering Čech and Rips complexes, and then build the confidence sets from them.

First, we estimate the persistent homology of p_h by building up the approximating filtration where each upper level set is replaced by the union of closed balls around the sample points of high density. Formally, for $h > 0$ and $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_1, \ldots, r_n > 0$, we define the upper level set estimator $\widehat{D}_L(r)$ as

$$\widehat{D}_L(r) := \begin{cases} \bigcup_{i \in \widehat{I}_L} \mathcal{B}(X_i, r_i), \;\; \widehat{I}_L = \{i \in [n] : \widehat{p}_h(X_i) \geq L\} & \text{if } L > 0 \\ \mathbb{R}^d & \text{if } L \leq 0. \end{cases} \tag{6}$$

Then we build up the approximating filtration as

$$\left\{\widehat{D}_L(r)\right\}_{L \in \mathbb{R}}, \tag{7}$$

and let $\mathrm{PH}_*(\widehat{p}_h, r)$ be the corresponding persistent homology. Then we can use $\mathrm{PH}_*(\widehat{p}_h, r)$ as the persistent homology estimator $\widehat{\mathrm{PH}_*(p_h)}$, due to the following stability theorem.

Theorem 3. *Suppose Assumption 1 holds for the density function p and Assumption 2 holds for the kernel function K. For any given $h > 0$ and $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_i \geq h, \;\; \forall i$,*

$$d_B\left(\mathrm{PH}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_r, \tag{8}$$

where

$$\widehat{c}_r := \max_i \sup_{\|x - X_i\| \leq r_i} |\widehat{p}_h(x) - \widehat{p}_h(X_i)|. \tag{9}$$

We estimate the distance $\|\widehat{p}_h - p_h\|_\infty$ by using the bootstrap. First, we generate B bootstrap samples $\{\widetilde{X}_1^1, \ldots, \widetilde{X}_n^1\}, \ldots, \{\widetilde{X}_1^B, \ldots, \widetilde{X}_n^B\}$, by sampling with replacement from the original sample. On each bootstrap sample, let $T_i = \sqrt{nh^d}\|\widehat{p}_h - \widehat{p}_h^i\|_\infty$, where $\widehat{p}_h^i$ is the kernel density estimator computed on ith bootstrap samples $\{\widetilde{X}_1^i, \ldots, \widetilde{X}_n^i\}$. Let the bootstrap quantile $\widehat{z}_\alpha$ as

$$\widehat{z}_\alpha = \inf\left\{z : \; \frac{1}{B}\sum_{i=1}^{B} I(T_i > z) \leq \alpha\right\}. \tag{10}$$

Then we define our confidence set as

$$\widehat{C}_\alpha := \left\{\mathcal{P} : \; d_B\left(\mathcal{P}, \mathrm{PH}_*(\widehat{p}_h, r)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right\}. \tag{11}$$

This confidence set is a valid asymptotic $1 - \alpha$ confidence set, as in the following theorem:

Theorem 4. *The confidence set $\widehat{C}_\alpha$ in* (11) *is asymptotically valid and satisfies*

$$\mathbb{P}\left(d_B\left(\mathrm{PH}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \geq 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right). \tag{12}$$

Note that this is a valid confidence set, but computing $\mathrm{PH}_*(\widehat{p}_h, r)$ exactly is infeasible. Hence we further approximate this by considering the weighted Čech complex and the weighted Rips complex. For $L, h > 0$ and $r \in \mathbb{R}^n$ with $r_1, \ldots, r_n > 0$, let the samples of high density as

$$\widehat{\mathcal{X}}_n^L := \begin{cases} \{X_i : \widehat{p}_h(X_i) \geq L; 1 \leq i \leq n\}, & L > 0, \\ \mathbb{R}^d, & L \leq 0. \end{cases}$$

Consider the weighted Čech complex $\check{\mathrm{C}}\mathrm{ech}(\widehat{\mathcal{X}}_n^L, r)$ and the weighted Rips complex $R(\widehat{\mathcal{X}}_n^L, r)$. Let $\mathrm{PH}_*^{Cech}(\widehat{p}_h, r)$ and $\mathrm{PH}_*^R(\widehat{p}_h, r)$ be the corresponding persistent homology of the filtrations $\left\{\check{\mathrm{C}}\mathrm{ech}(\widehat{\mathcal{X}}_n^L, r)\right\}_{L \in \mathbb{R}}$ and $\left\{R(\widehat{\mathcal{X}}_n^L, r)\right\}_{L \in \mathbb{R}}$, respectively. Then we have the following stability theorem:

Theorem 5. *Suppose Assumption 1 holds for the density function p and Assumption 2 holds for the kernel function K. For any given $h > 0$ and $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_i \geq h, \ \forall i$,*

$$d_B\left(\mathrm{PH}_*^{Cech}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_r, \tag{13}$$

and

$$d_B\left(\mathrm{PH}_*^{R}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_{2r}. \tag{14}$$

As before, we estimate the distance $\|\widehat{p}_h - p_h\|_\infty$ by using the bootstrap, and define our confidence sets as

$$\widehat{C}_\alpha^{Cech} := \left\{\mathcal{P} : \ d_B\left(\mathcal{P}, \mathrm{PH}_*^{Cech}(\widehat{p}_h, r)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right\}, \tag{15}$$

$$\widehat{C}_\alpha^{R} := \left\{\mathcal{P} : \ d_B\left(\mathcal{P}, \mathrm{PH}_*^{R}(\widehat{p}_h, r)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_{2r}\right\}. \tag{16}$$

These confidence sets are valid asymptotic $1 - \alpha$ confidence sets, as in the following theorem:

Theorem 6. *The confidence set $\widehat{C}_\alpha^{Cech}$ in* (15) *is asymptotically valid and satisfies*

$$\mathbb{P}\left(d_B\left(\mathrm{PH}_*^{Cech}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \geq 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right). \tag{17}$$

Similarly, the confidence set $\widehat{C}_\alpha^{Rips}$ in (16) *is asymptotically valid and satisfies*

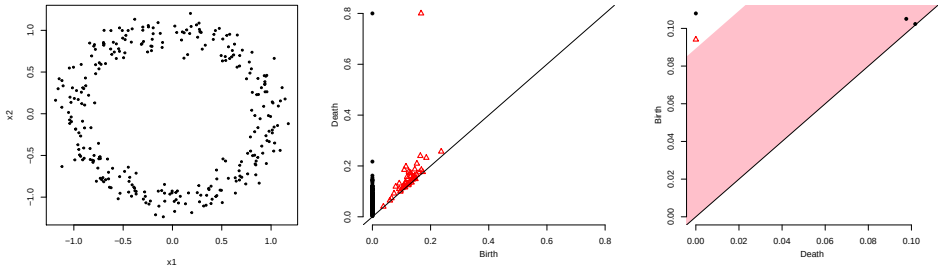
$$\mathbb{P}\left(d_B\left(\mathrm{PH}_*^{R}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_{2r}\right) \geq 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right). \tag{18}$$

Remark 7. *If we set $r_i = r, \forall i \in [n]$, we can replace $\widehat{c}_{2r}$ with $\widehat{c}_{\sqrt{2}r}$.*

5 Examples

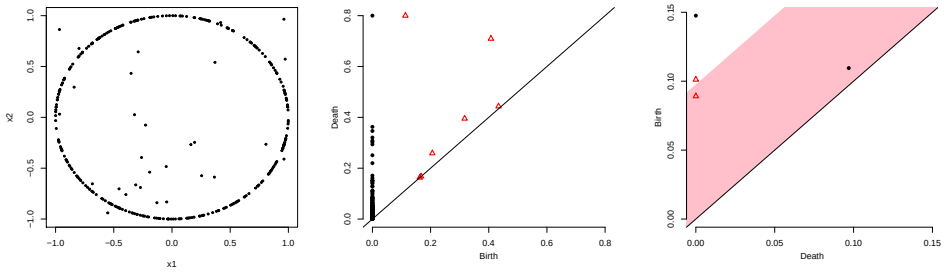
To illustrate how one can use the methods in the previous section to do statistical inference on topological features of data generating distributions, we calculate persistence diagrams of $\mathrm{PH}_*^R(\widehat{p}_h, r)$ and their confidence sets on toy examples. We make 2 synthetic data sets with circular shapes which are described in the left side of Figure 1 and 2. The right side shows persistence diagrams of $\mathrm{PH}_*^R(\widehat{p}_h, r)$. Each black dot indicates the birth and death of each 0-th homology class corresponding to each connected component. Similarly, each red triangle represents the birth and death of each 1-st homology class related to each one-dimensional hole. For all diagrams, the shaded banded regions correspond to 90% confidence sets in the sense that any homology class contained in the bands cannot be distinguished from the diagonal lines within the confidence sets. In other words, homology classes outside of band illustrate significant topological features of the underlying distribution. We refer to Fasy et al. [2014] for the detailed interpretation. In Figure 1 (c) and 2 (c), we can check there are a black dot and a red triangle outside of band which coincide to the fact that most of the data are distributed around a circle with a hole.

Persistence diagrams of $\mathrm{PH}_*^R(\widehat{p}_h, r)$ depend on choices of parameters h and $r = (r_1, \ldots, r_n)$. To choose appropriate parameters, we can select the parameter that maximizes the total number of significant homology classes which is a generally adopted strategy in TDA [Chazal et al., 2014]. In our case, we can also use another heuristic but intuitive parameter selection method based on the diagram of the Rips complex filtration $\left\{R(\mathcal{X}_n^0, r)\right\}_{r \geq 0}$. Recall that $\mathrm{PH}_*^R(\widehat{p}_h, r)$ is the persistent homology of the filtration $\left\{R(\mathcal{X}_n^L, r)\right\}_{L \in \mathbb{R}}$. Since it is based on Rips complex with radius r, $\mathrm{PH}_*^R(\widehat{p}_h, r)$ can only capture the homology classes whose birth time is smaller than r and death time is greater than r in the Rips diagram. Therefore, once the Rips diagram reveals some seemingly significant homology classes whose lifetimes are longer than the others, we can choose appropriate h and r to make sure the base line Rips complex contain the seemingly significant homology groups.



(a) (b) (c)

Figure 1: *One circle with additive noise example.* (a) Data points uniformly distributed over a circle with additive Gaussian noise. (b) Rips diagram. (c) Persistence diagram of KDE filtration on Rips complex. ($r_i = h = 0.7,\ \forall i$)



(a) (b) (c)

Figure 2: *One circle with background noise example.* (a) Data points uniformly distributed over a circle, and few outliers are added to the data set. (b) Rips diagram. (c) Persistence diagram of KDE filtration on Rips complex. ($r_i = h = 0.65,\ \forall i$)

6 Conclusions

In this paper we have developed a new methodology for constructing asymptotic confidence sets for persistence diagrams of density level sets that are computationally less expensive than existing procedures.

There are two main open questions that we will address in future work. First, while we successfully avoid evaluating the KDE on a grid of points in order to compute the persistent homology, we still need to calculate the values of the KDE on a grid to obtain the quantities $\widehat{c}_r$ in (9) and $\widehat{z}_\alpha$ in (10). As computing the persistent homology is a much bigger computational bottleneck, overall we have significantly reduced the computation time. However, the computation time still depends on the ambient dimension. To address this issue we propose to approximate $\widehat{c}_r$ and $\widehat{z}_\alpha$ using quantities that can be computed only evaluating the KDE on the sample points. In detail, we propose to replace $\widehat{c}_r$ by the similar quantity

$$\widetilde{c}_r := \max_i \sup_{X_j \in B(X_i, r_i), X_j \neq X_i} |\widehat{p}_h(X_j) - \widehat{p}_h(X_i)|.$$

Under mild conditions on the density, such as the (a, b) condition of Cuevas and Rodríguez-Casal [2004], it can be shown that $\widetilde{c}_r$ and $\widehat{c}_r$ are close with high probability. Then a weaker form of the stability theorem will guarantee that our results remain valid with $\widehat{c}_r$ replaced by $\widetilde{c}_r$. Secondly, we can replace $\widehat{z}_\alpha$ with the $(1 - \alpha)$ quantile of the the bootstrap replicates of the L_∞ norms of the differences between $\widehat{p}_h$ and its bootstrap version $\widehat{p}^*_h$, evaluated only at the sample points. In order to show that this quantity is close to $\widehat{z}_\alpha$ with high probability we will also assume the (a, b) condition.

References

S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster Trees on Manifolds. *ArXiv e-prints*, July 2013.

A. Björner. Handbook of combinatorics (vol. 2). chapter Topological Methods, pages 1819–1872. MIT Press, Cambridge, MA, USA, 1995.

Omer Bobrowski, Sayan Mukherjee, and Jonathan E Taylor. Topological consistency via kernel estimation. *arXiv preprint arXiv:1407.5272*, 2014.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1): 77–102, January 2015.

Benoît Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4):999–1023, 2006.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23*, pages 343–351. 2010.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J Guibas, and Steve Y Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 237–246. ACM, 2009.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes. *CoRR*, 2013.

Frédéric Chazal, Brittany T Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.

Moo K. Chung, Peter Bubenik, and Peter T. Kim. Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging, 21st International Conference, IPMI 2009, Williamsburg, VA, USA, July 5-10, 2009. Proceedings*, pages 386–397, 2009.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36 (02):340–354, 2004.

H. Edelsbrunner and J. Harer. Persistent homology -— a survey. In *Surveys on discrete and computational geometry*, volume 453, page 257. Amer Mathematical Society, 2008.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 588–606. PMLR, 2015.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, Aarti Singh, et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

J. A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76:388–394, 1981.

John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan, Alessandro Rinaldo, and Larry Wasserman. Statistical inference for cluster trees. In *Advances in Neural Information Processing Systems 29*, pages 1839–1847. 2016.

Jeff M. Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. *CoRR*, abs/1307.7760, 2013. URL `http://arxiv.org/abs/1307.7760`.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5): 2678–2722, 2010.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2): 249–274, February 2005.

A Stability Theorem for Persistence module

This section gives an introduction to the Stability Theorem on persistence module. We refer to Chazal et al. [2009] for more details.

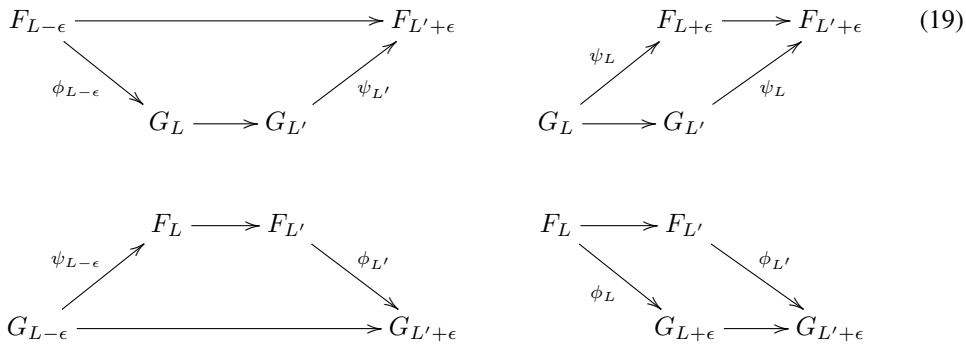
A persistence module is an algebraic abstraction of a persistent homology.

Definition 11. *[Chazal et al., 2009, Definition 2.1] A* persistence module $\mathcal{F}$ *is a family* $\{F_L\}_{L\in\mathbb{R}}$ *of* $\mathbb{Z}_2$*-vector spaces indexed by the elements of* $\mathbb{R}$*, together with a family* $\{f_L^{L'} : F_L \to F_{L'}\}_{L\leq L'}$ *of homomorphisms such that:* $\forall L \leq L' \leq L''$, $f_L^{L''} = f_{L'}^{L''} \circ f_L^{L'}$ *and* $f_L^L = id_{F_L}$.

We say that $\mathcal{F}$ is tame if F_L is a finite dimensional vector spaces for all $L \in \mathbb{R}$.

For two functions $f, g : \mathbb{X} \to \mathbb{R}$ satisfying $\|f - g\|_\infty \leq \epsilon$, their sublevel sets filtrations are nested as follows: $\forall L \in \mathbb{R}$, $\mathbb{X}_L^f \subset \mathbb{X}_{L+\epsilon}^g \subset \mathbb{X}_{L+2\epsilon}^f$. By letting $F_L = H_k(\mathbb{X}_L^f)$ and $G_L = H_k(\mathbb{X}_L^g)$, this induces the homomorphisms induced by the inclusions as $F_L \to G_{L+\epsilon} \to F_{L+2\epsilon}$. Also, the canonical inclusions $\mathbb{X}_L^f \subset \mathbb{X}_{L'}^f$ and $\mathbb{X}_L^g \subset \mathbb{X}_{L'}^g$ for $L \leq L'$ induces homomorphisms as $F_L \to F_{L'}$ and $G_L \to G_{L'}$. This homomorphisms relations can be extended to persistence modules as follows:

Definition 12. *Two persistence modules* $\mathcal{F}$ *and* $\mathcal{G}$ *are said to be strongly* ϵ*-interleaved if there exist two families of homomorphisms* $\{\phi_L : F_L \to G_{L+\epsilon}\}_{L\in\mathbb{R}}$ *and* $\{\psi_L : G_L \to F_{L+\epsilon}\}_{L\in\mathbb{R}}$ *such that the following diagrams commute for all* $L \leq L'$*:*



(19)

If two persistence modules are strongly interleaved, then their bottleneck distance are close, which is the strong stability theorem.

Theorem 8 (Strong Stability Theorem)**.** *[Chazal et al., 2009, Theorem 4.4] Let* $\mathcal{F}_\mathbb{R}$ *and* $\mathcal{G}_\mathbb{R}$ *be two tame persistence modules. If* $\mathcal{F}_\mathbb{R}$ *and* $\mathcal{G}_\mathbb{R}$ *are strongly interleaved, then* $d_B(\mathcal{F}_\mathbb{R}, \mathcal{G}_\mathbb{R}) \leq \epsilon$.

B Proofs

For bootstraping the ℓ_∞ distance $\|\widehat{p}_h - p\|_\infty$ as described in Section 4, Theorem 9 gives its validity. For its proof, see [Fasy et al., 2014, Theorem 12] and [Chazal et al., 2013, Theorem 2.1].

Theorem 9.

$$\mathbb{P}\left(\sqrt{nh^d}\|\widehat{p}_h - p_h\|_\infty \leq \widehat{z}_\alpha\right) = 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right).$$

Theorem 3. *Suppose Assumption 1 holds for the density function* p *and Assumption 2 holds for the kernel function* K*. For any given* $h > 0$ *and* $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ *with* $r_i \geq h$, $\forall i$*,*

$$d_B\left(\mathrm{PH}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_r,$$

where

$$\widehat{c}_r := \max_i \sup_{\|x - X_i\| \leq r_i} |\widehat{p}_h(x) - \widehat{p}_h(X_i)|.$$

Proof of Theorem 3. From the strong stability theorem in Theorem 8, it is sufficient to show strong ϵ-interleaving conditions in (19) on the homology level. And since D_L and $\widehat{D}_L$ are all sets, inclusion relations will carry over to the homology. Hence it is sufficient to show that enough to show that

$$D_L \subset \widehat{D}_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}, \text{ and } \widehat{D}_L \subset D_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}, \text{ for } \forall L \in \mathbb{R}.$$

For the first part, if $L \leq \|\widehat{p}_h - p_h\|_\infty$, then $D_L \subset \widehat{D}_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r} = \mathbb{R}^d$. If not, suppose $x \in D_L$, which is equivalent to $p_h(x) \geq L$. Then

$$\widehat{p}_h(x) \geq p_h(x) - \|\widehat{p}_h - p_h\|_\infty \geq L - \|\widehat{p}_h - p_h\|_\infty > 0.$$

Then from Assumption 2, $supp(K) \subset \mathcal{B}(0,1)$, hence $\widehat{p}_h(x) > 0$ implies that there exists some $X_i \in \mathcal{X}_n$ such that $\|x - X_i\| \leq h$. Then from the condition $h \leq r_i$, $\|x - X_i\| \leq r_i$ holds, and

$$\widehat{p}_h(X_i) \geq \widehat{p}_h(x) - \widehat{c}_r \geq L - \|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r.$$

Hence $x \in \widehat{D}_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}$ holds. i.e. $D_L \subset \widehat{D}_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}$ holds.

For the second part, if $L \leq \|\widehat{p}_h - p_h\|_\infty$, then $\widehat{D}_L \subset D_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r} = \mathbb{R}^d$. If not, suppose $x \in \widehat{D}_L$. Then there exists $X_i \in \mathcal{X}_n$ such that $\|x - X_i\| \leq r_i$ and $\widehat{p}_h(X_i) \geq L$. Then

$$\widehat{p}_h(x) \geq \widehat{p}_h(X_i) - \widehat{c}_r \geq L - \widehat{c}_r$$

holds, and then

$$p_h(x) \geq \widehat{p}_h(x) - \|\widehat{p}_h - p_h\|_\infty \geq L - \|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r.$$

Hence $x \in D_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}$ holds, i.e. $\widehat{D}_L \subset D_{L-\|\widehat{p}_h - p_h\|_\infty - \widehat{c}_r}$ holds.

□

Theorem 4. *The confidence set $\widehat{C}_\alpha$ in* (11) *is asymptotically valid and satisfies*

$$\mathbb{P}\left(d_B\left(\mathrm{PH}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \geq 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right).$$

Proof of Theorem 4. Theorem 3 and Theorem 9 together imply as

$$\begin{aligned}
\mathbb{P}\left(d_B\left(\mathrm{PH}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) &\geq \mathbb{P}\left(\|p_h - \widehat{p}_h\|_\infty + \widehat{c}_r \leq \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \\
&= \mathbb{P}\left(\sqrt{nh^d}\|p_h - \widehat{p}_h\|_\infty \leq \widehat{z}_\alpha\right) \\
&= 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right).
\end{aligned}$$

□

Theorem 5. *Suppose Assumption 1 holds for the density function p and Assumption 2 holds for the kernel function K. For any given $h > 0$ and $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_i \geq h$, $\forall i$,*

$$d_B\left(\mathrm{PH}_*^{Cech}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_r, \tag{20}$$

and

$$d_B\left(\mathrm{PH}_*^{R}(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \leq \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_{2r}. \tag{21}$$

Proof of Theorem 5. (Step 1)

In this proof, write $\check{C}(r)$ for $\check{C}ech(\mathcal{X}_n, r)$ and write $R(r)$ for $R(\mathcal{X}_n, r)$, for convenience.

Let $\epsilon > 0$ to be defined later, $L \in \mathbb{R}$, and $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$ with $r_1, \ldots, r_n > 0$. Triangulate $\mathbb{X}$. With taking subdivision if necessary, assume that D_L, $D_{L-\epsilon}$, $sd\left(\widehat{D}_L(r)\right)$, $sd\left(\widehat{D}_{L-\epsilon}(r)\right)$,

$\mathcal{B}(X_i, r_i)$, $B(X_i, 2r_i)$ are subcomplices. We first consider two simplicial maps from Nerve Theorem [Björner, 1995, Theorem 10.6]. We define a simplicial map $\phi_L^r : sd\left(\widehat{D}_L(r)\right) \to sd\left(\check{C}_L(r)\right)$ to be a barycentric map induced from $\sigma \mapsto \left\{X_i \in \mathcal{X}_n^L : \sigma \in B(X_i, r_i)\right\}$ (where each $\mathcal{B}(X_i, r_i)$ is understood as a simplicial subcomplex of $\mathbb{X}$). We also define a simplicial map $\psi_L^r : sd\left(\check{C}_L(r)\right) \to sd\left(\widehat{D}_L(r)\right)$ to be a barycentric map induced from $\{X_{n_1}, \ldots, X_{n_k}\} \mapsto \frac{\sum_{j=1}^k r_j X_{n_j}}{\sum_{j=1}^k r_j}$. Then the proof of [Björner, 1995, Theorem 10.6] implies that

$$\psi_L^r \circ \phi_L^r \simeq id_{\widehat{D}_L(r)} \text{ and } \phi_L^r \circ \psi_L^r \simeq id_{\check{C}_L(r)}. \tag{22}$$

Also, suppose $r' = \lambda r$ for some $\lambda \in [1, \infty)$, and suppose $L' \leq L$. Then for each $\sigma \in sd\left(\widehat{D}_L(r)\right)$, since vertices of σ can be ordered by inclusion relation, we can define its minimal vertex $\min \sigma$. Now let $\Delta_\sigma := \left\{X_i \in \mathcal{X}_n^{L'} : \min \sigma \in \mathcal{B}(X_i, r_i)\right\}$ be understood as a simplex in $\check{C}_{L'}(r')$ (it is indeed a simplex since $\min \sigma \in \mathcal{B}(X_i, r_i)$ implies $\|X_i - X_j\| \leq r_i + r_j$), then $\|\phi_L^r(\sigma)\|, \|\phi_{L'}^{r'}(\sigma)\| \subset \|\Delta_\sigma\|$. Hence for any $\gamma \in B_*\left(sd\left(\widehat{D}_L(r)\right)\right)$, $\phi_L^r(\gamma)$ and $\phi_{L'}^{r'}(\gamma)$ are homotopic to each other in $sd\left(\check{C}_{L'}(r')\right)$, and hence in $H_*\left(sd\left(\check{C}_{L'}(r')\right)\right)$,

$$(\phi_L^r)_*[\gamma] = (\phi_{L'}^{r'})_*[\gamma]. \tag{23}$$

Also, ψ_L^r satisfies that if $\sigma \in sd\left(\check{C}_L(r)\right) \cap sd\left(\check{C}_{L'}(r')\right)$ with $r' = \lambda r$ for some $\lambda \in (0, \infty)$, then

$$\psi_L^r(\sigma) = \psi_{L'}^{r'}(\sigma). \tag{24}$$

(Step 2)

We prove (20), i.e. Stability Theorem with Čech complex. Let $\epsilon := \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_r$. Our goal is to define simplicial maps $\Phi_L : D_L \to sd\left(\check{C}_{L-\epsilon}(r)\right)$ and $\Psi_L : sd\left(\check{C}_L(r)\right) \to D_{L-\epsilon}$ so that $(\Phi_L)_* : H_*(D_L) \to H_*\left(\check{C}_{L-\epsilon}(r)\right)$ and $(\Psi_L)_* : H_*\left(\check{C}_L(r)\right) \to H_*(D_{L-\epsilon})$ are homomorphisms satisfying strong ϵ-interleaving conditions in (19). Then Strong Stability Theorem (Theorem 8) implies (20).

Now we construct Φ_L and Ψ_L. Let $\imath_L^{D \to \widehat{D}} : D_L \to sd\left(\widehat{D}_{L-\epsilon}(r)\right)$, $\imath_L^{\widehat{D} \to D} : sd\left(\widehat{D}_L(2r)\right) \to D_{L-\epsilon}$ be simplicial maps induced from the inclusion maps. And then we define $\Phi_L := \phi_{L-\epsilon}^r \circ \imath_L^{D \to \widehat{D}} : D_L \to sd\left(\check{C}_{L-\epsilon}(r)\right)$ and $\Psi_L := \imath_L^{\widehat{D} \to D} \circ \psi_L^r : sd\left(\check{C}_L(r)\right) \to D_{L-\epsilon}$. For $L' \in \mathbb{R}$ with $L' \leq L$, let $\imath_{L \to L'}^D : D_L \to D_{L'}$, $\imath_{L \to L'}^C : sd\left(\check{C}_L(r)\right) \to sd\left(\check{C}_L(r)\right)$ be simplicial maps induced from the inclusion maps.

First we show that the diagram in (25) commutes,

$$\begin{array}{ccc} H_*(D_{L+\epsilon}) & \longrightarrow & H_*(D_{L'-\epsilon}) \\ & {\scriptstyle \Phi_{L+\epsilon}} \searrow \quad \nearrow {\scriptstyle \Psi_{L'}} & \\ & H_*\left(\check{C}_L(r)\right) \longrightarrow H_*\left(\check{C}_{L'}(r)\right) & \end{array} \tag{25}$$

i.e. compare $\Psi_{L'} \circ \imath_{L \to L'}^C \circ \Phi_{L+\epsilon} : D_{L+\epsilon} \to D_{L'-\epsilon}$ to inclusion map $\imath_{L+\epsilon \to L'-\epsilon}^D : D_{L+\epsilon} \to D_{L'-\epsilon}$. For $\gamma \in B_*\left(D_{L+\epsilon}\right)$, note that $\Phi_{L+\epsilon}(\gamma) = \phi_L^r(\gamma)$, so $\Psi_{L'} \circ \imath_{L \to L'}^C \circ \Phi_{L+\epsilon}(\gamma) = \psi_{L'}^r \circ \phi_L^r(\gamma)$. Then since $\phi_L^r(\gamma) \in B_*\left(sd\left(\check{C}_L(r)\right)\right) \subset B_*\left(sd\left(\check{C}_{L'}(r)\right)\right)$, (24) implies

$$\psi_{L'}^r \circ \phi_L^r(\gamma) = \psi_L^r \circ \phi_L^r(\gamma).$$

Then from (22),

$$\left(\psi_L^r \circ \phi_L^r\right)_*[\gamma] = id_{\widehat{D}_L(r)}[\gamma] = [\gamma]$$

in $H_*\left(\widehat{D}_L(r)\right)$. Since $\widehat{D}_L(r) \subset D_{L'-\epsilon}$,

$$\left(\Psi_{L'} \circ \imath_{L \to L'}^C \circ \Phi_{L+\epsilon}\right)_*[\gamma] = \left(\psi_L^r \circ \phi_L^r\right)_*[\gamma] = [\gamma] = \left(\imath_{L+\epsilon \to L'-\epsilon}^D\right)_*[\gamma]$$

in $H_*(D_{L'-\epsilon})$ as well.

Second, we show that the diagram in (26) commutes,

$$\begin{array}{ccc} & H_*\left(D_{L-\epsilon}\right) & \longrightarrow H_*\left(D_{L'-\epsilon}\right) \\ {\scriptstyle \Psi_L}\nearrow & & \nearrow{\scriptstyle \Psi_{L'}} \\ H_*\left(\check{C}_L(r)\right) & \longrightarrow H_*\left(\check{C}_{L'}(r)\right) & \end{array} \tag{26}$$

i.e. compare $\Psi_{L'} \circ \imath^C_{L\to L'} : sd(\check{C}_L(r)) \to D_{L'-\epsilon}$ to $\imath^D_{L-\epsilon\to L'-\epsilon} \circ \Psi_L : sd(\check{C}_L(r)) \to D_{L'-\epsilon}$. For $\gamma \in B_*\left(sd(R_L(r))\right)$, note that $\Psi_{L'} \circ \imath^C_{L\to L'}(\gamma) = \psi^r_{L'}(\gamma)$ and $\imath^D_{L-\epsilon\to L'-\epsilon} \circ \Psi_L(\gamma) = \psi^r_L(\gamma)$. Then since $\gamma \in B_*\left(sd(\check{C}_L(r))\right) \subset B_*\left(sd(\check{C}_{L'}(r))\right)$, (24) implies

$$\Psi_{L'} \circ \imath^C_{L\to L'}(\gamma) = \psi^r_{L'}(\gamma) = \psi^r_L(\gamma) = \imath^D_{L-\epsilon\to L'-\epsilon} \circ \Psi_L(\gamma),$$

hence $\left(\Psi_{L'} \circ \imath^C_{L\to L'}\right)_*[\gamma] = \left(\imath^D_{L-\epsilon\to L'-\epsilon} \circ \Psi_L\right)_*[\gamma]$ in $H_*(D_{L'-\epsilon})$.

Third, we show that the diagram in (27) commutes,

$$\begin{array}{ccc} & H_*\left(D_L\right) & \longrightarrow H_*\left(D_{L'}\right) \\ {\scriptstyle \Psi_{L+\epsilon}}\nearrow & & \searrow{\scriptstyle \Phi_{L'}} \\ H_*\left(\check{C}_{L+\epsilon}(r)\right) & & \longrightarrow H_*\left(\check{C}_{L'-\epsilon}(r)\right) \end{array} \tag{27}$$

i.e. compare $\Phi_{L'} \circ \imath^D_{L\to L'} \circ \Psi_{L+\epsilon} : sd\left(\check{C}_{L+\epsilon}(r)\right) \to sd\left(\check{C}_{L'-\epsilon}(r)\right)$ to inclusion map $\imath^C_{L+\epsilon\to L'-\epsilon} : sd\left(\check{C}_{L+\epsilon}(r)\right) \to sd\left(\check{C}_{L'-\epsilon}(r)\right)$. For $\gamma \in B_*\left(sd\left(\check{C}_{L+\epsilon}(r)\right)\right)$, note that $\Psi_{L+\epsilon}(\gamma) = \psi^r_{L+\epsilon}(\gamma)$, so $\Phi_{L'} \circ \imath^D_{L\to L'} \circ \Psi_{L+\epsilon}(\gamma) = \phi^r_{L'-\epsilon} \circ \psi^r_{L+\epsilon}(\gamma)$. Then since $\gamma \in B_*\left(sd\left(\check{C}_{L+\epsilon}(r)\right)\right) \subset B_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right)$, (24) implies

$$\phi^r_{L'-\epsilon} \circ \psi^r_{L+\epsilon}(\gamma) = \phi^r_{L'-\epsilon} \circ \psi^r_{L'-\epsilon}(\gamma).$$

Then from (22),

$$\left(\Phi_{L'} \circ \imath^D_{L\to L'} \circ \Psi_{L+\epsilon}\right)_*[\gamma] = \left(\phi^r_{L'-\epsilon} \circ \psi^r_{L'-\epsilon}\right)_*[\gamma] = id_{sd\left(\check{C}_{L'-\epsilon}(r)\right)}[\gamma] = [\gamma] = \left(\imath^C_{L+\epsilon\to L'-\epsilon}\right)_*[\gamma]$$

in $H_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right) \cong H_*\left(\check{C}_{L'-\epsilon}(r)\right)$.

Fourth, we show that the diagram in (28) commutes,

$$\begin{array}{ccc} H_*\left(D_L\right) & \longrightarrow H_*\left(D_{L'}\right) & \\ & {\scriptstyle \Phi_L}\searrow \quad \swarrow{\scriptstyle \Phi_{L'}} & \\ & H_*\left(\check{C}_{L-\epsilon}(r)\right) \longrightarrow H_*\left(\check{C}_{L'-\epsilon}(r)\right) & \end{array} \tag{28}$$

i.e. compare $\Phi_{L'} \circ \imath^D_{L\to L'} : D_L \to sd(\check{C}_{L'-\epsilon}(r))$ to $\imath^C_{L-\epsilon\to L'-\epsilon} \circ \Phi_L : D_L \to sd(\check{C}_{L'-\epsilon}(r))$. For $\gamma \in B_*\left(D_L\right)$, note that $\Phi_{L'} \circ \imath^D_{L\to L'}(\gamma) = \phi^r_{L'-\epsilon}(\gamma)$ and $\imath^C_{L-\epsilon\to L'-\epsilon} \circ \Phi_L(\gamma) = \phi^r_{L-\epsilon}(\gamma)$. Then from (23),

$$\left(\Phi_{L'} \circ \imath^D_{L\to L'}\right)_*[\gamma] = \left(\phi^r_{L'-\epsilon}\right)_*[\gamma] = \left(\phi^r_{L-\epsilon}\right)_*[\gamma] = \left(\imath^C_{L-\epsilon\to L'-\epsilon} \circ \Phi_L\right)_*[\gamma]$$

in $H_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right) \cong H_*\left(\check{C}_{L'-\epsilon}(r)\right)$.

(Step 3)

We prove (21), i.e. Stability Theorem with Vietoris-Rips complex. Let $\epsilon := \|\widehat{p}_h - p_h\|_\infty + \widehat{c}_{2r}$. Our goal is to define simplicial maps $\Phi_L : D_L \to sd\left(R_{L-\epsilon}(r)\right)$ and $\Psi_L : sd\left(R_L(r)\right) \to D_{L-\epsilon}$ so that $\left(\Phi_L\right)_* : H_*(D_L) \to H_*\left(R_{L-\epsilon}(r)\right)$ and $\left(\Psi_L\right)_* : H_*\left(R_L(r)\right) \to H_*(D_{L-\epsilon})$ are homomorphisms satisfying strong ϵ-interleaving conditions in (19). Then Strong Stability Theorem (Theorem 8) implies (21).

Now we construct Φ_L and Ψ_L. Let $\imath_L^{D\to\widehat{D}} : D_L \to sd\left(\widehat{D}_{L-\epsilon}(r)\right)$, $\imath_L^{C\to R} : sd\left(\check{C}_L(r)\right) \to sd\left(R_L(r)\right)$, $\imath_L^{R\to C} : sd\left(R_L\right) \to sd\left(\check{C}_L(2r)\right)$, $\imath_L^{\widehat{D}\to D} : sd\left(\widehat{D}_L(2r)\right) \to D_{L-\epsilon}$ be simplicial maps induced from the inclusion maps. And then we define $\Phi_L := \imath_{L-\epsilon}^{C\to R} \circ \phi_{L-\epsilon}^r \circ \imath_L^{D\to\widehat{D}} : D_L \to sd\left(R_{L-\epsilon}(r)\right)$ and $\Psi_L := \imath_L^{\widehat{D}\to D} \circ \psi_L^{2r} \circ \imath_L^{R\to C} : sd\left(R_L(r)\right) \to D_{L-\epsilon}$. For $L' \in \mathbb{R}$ with $L' \leq L$, let $\imath_{L\to L'}^D : D_L \to D_{L'}$, $\imath_{L\to L'}^R : sd\left(R_L(r)\right) \to sd\left(R_L(r)\right)$ be simplicial maps induced from the inclusion maps.

First we show that the diagram in (29) commutes,

$$\begin{array}{ccccc} H_*(D_{L+\epsilon}) & & \longrightarrow & & H_*(D_{L'-\epsilon}) \\ & \searrow^{\Phi_{L+\epsilon}} & & \nearrow_{\Psi_{L'}} & \\ & H_*\left(R_L(r)\right) & \longrightarrow & H_*\left(R_{L'}(r)\right) & \end{array} \tag{29}$$

i.e. compare $\Psi_{L'} \circ \imath_{L\to L'}^R \circ \Phi_{L+\epsilon} : D_{L+\epsilon} \to D_{L'-\epsilon}$ to inclusion map $\imath_{L+\epsilon\to L'-\epsilon}^D : D_{L+\epsilon} \to D_{L'-\epsilon}$. For $\gamma \in B_*\left(D_{L+\epsilon}\right)$, note that $\Phi_{L+\epsilon}(\gamma) = \phi_L^r(\gamma)$, so $\Psi_{L'} \circ \imath_{L\to L'}^R \circ \Phi_{L+\epsilon}(\gamma) = \psi_{L'}^{2r} \circ \phi_L^r(\gamma)$. Then since $\phi_L^r(\gamma) \in B_*\left(sd\left(\check{C}_L(r)\right)\right) \subset B_*\left(sd\left(\check{C}_{L'}(2r)\right)\right)$, (24) implies

$$\psi_{L'}^{2r} \circ \phi_L^r(\gamma) = \psi_L^r \circ \phi_L^r(\gamma).$$

Then from (22),

$$\left(\psi_L^r \circ \phi_L^r\right)_*[\gamma] = id_{\widehat{D}_L(r)}[\gamma] = [\gamma]$$

in $H_*\left(\widehat{D}_L(r)\right)$. Since $\widehat{D}_L(r) \subset D_{L'-\epsilon}$,

$$\left(\Psi_{L'} \circ \imath_{L\to L'}^R \circ \Phi_{L+\epsilon}\right)_*[\gamma] = \left(\psi_L^r \circ \phi_L^r\right)_*[\gamma] = [\gamma] = \left(\imath_{L+\epsilon\to L'-\epsilon}^D\right)_*[\gamma]$$

in $H_*(D_{L'-\epsilon})$ as well.

Second, we show that the diagram in (30) commutes,

$$\begin{array}{ccccc} & & H_*\left(D_{L-\epsilon}\right) & \longrightarrow & H_*\left(D_{L'-\epsilon}\right) \\ & \nearrow^{\Psi_L} & & \nearrow_{\Psi_{L'}} & \\ H_*\left(R_L(r)\right) & \longrightarrow & H_*\left(R_{L'}(r)\right) & & \end{array} \tag{30}$$

i.e. compare $\Psi_{L'} \circ \imath_{L\to L'}^R : sd(R_L(r)) \to D_{L'-\epsilon}$ to $\imath_{L-\epsilon\to L'-\epsilon}^D \circ \Psi_L : sd(R_L(r)) \to D_{L'-\epsilon}$. For $\gamma \in B_*\left(sd(R_L(r))\right)$, note that $\Psi_{L'} \circ \imath_{L\to L'}^R(\gamma) = \psi_{L'}^{2r}(\gamma)$ and $\imath_{L-\epsilon\to L'-\epsilon}^D \circ \Psi_L(\gamma) = \psi_L^{2r}(\gamma)$. Then since $\gamma \in B_*\left(sd(\check{C}_L(2r))\right) \subset B_*\left(sd(\check{C}_{L'}(2r))\right)$, (24) implies

$$\Psi_{L'} \circ \imath_{L\to L'}^R(\gamma) = \psi_{L'}^{2r}(\gamma) = \psi_L^{2r}(\gamma) = \imath_{L-\epsilon\to L'-\epsilon}^D \circ \Psi_L(\gamma),$$

hence $\left(\Psi_{L'} \circ \imath_{L\to L'}^R\right)_*[\gamma] = \left(\imath_{L-\epsilon\to L'-\epsilon}^D \circ \Psi_L\right)_*[\gamma]$ in $H_*(D_{L'-\epsilon})$.

Third, we show that the diagram in (31) commutes,

$$\begin{array}{ccccc} & & H_*\left(D_L\right) & \longrightarrow & H_*\left(D_{L'}\right) \\ & \nearrow^{\Psi_{L+\epsilon}} & & & \searrow_{\Phi_{L'}} \\ H_*\left(R_{L+\epsilon}(r)\right) & & \longrightarrow & & H_*\left(R_{L'-\epsilon}(r)\right) \end{array} \tag{31}$$

i.e. compare $\Phi_{L'} \circ \imath_{L\to L'}^D \circ \Psi_{L+\epsilon} : sd\left(R_{L+\epsilon}(r)\right) \to sd\left(R_{L'-\epsilon}(r)\right)$ to inclusion map $\imath_{L+\epsilon\to L'-\epsilon}^R : sd\left(R_{L+\epsilon}(r)\right) \to sd\left(R_{L'-\epsilon}(r)\right)$. For $\gamma \in B_*\left(sd\left(R_{L+\epsilon}(r)\right)\right)$, note that $\Psi_{L+\epsilon}(\gamma) = \psi_{L+\epsilon}^{2r}(\gamma)$, so $\Phi_{L'} \circ \imath_{L\to L'}^D \circ \Psi_{L+\epsilon}(\gamma) = \phi_{L'-\epsilon}^r \circ \psi_{L+\epsilon}^{2r}(\gamma)$. Then since $\gamma \in B_*\left(sd\left(\check{C}_{L+\epsilon}(2r)\right)\right) \subset B_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right)$ with subdivisions if necessary, (24) implies

$$\phi_{L'-\epsilon}^r \circ \psi_{L+\epsilon}^{2r}(\gamma) = \phi_{L'-\epsilon}^r \circ \psi_{L'-\epsilon}^r(\gamma).$$

Then from (22),

$$\left(\phi^r_{L'-\epsilon} \circ \psi^r_{L'-\epsilon}\right)_* [\gamma] = id_{sd\left(\check{C}_{L'-\epsilon}(r)\right)}[\gamma] = [\gamma]$$

in $H_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right)$. Since $sd\left(\check{C}_{L'-\epsilon}(r)\right) \subset sd\left(R_{L'-\epsilon}(r)\right)$,

$$\left(\Phi_{L'} \circ \imath^D_{L\to L'} \circ \Psi_{L+\epsilon}\right)_* [\gamma] = \left(\phi^r_{L'-\epsilon} \circ \psi^r_{L'-\epsilon}\right)_* [\gamma] = [\gamma] = \left(\imath^R_{L+\epsilon\to L'-\epsilon}\right)_* [\gamma]$$

in $H_*\left(sd\left(R_{L'-\epsilon}(r)\right)\right) \cong H_*\left(R_{L'-\epsilon}(r)\right)$ as well.

Fourth, we show that the diagram in (32) commutes,

$$\begin{array}{ccc} H_*\left(D_L\right) & \longrightarrow & H_*\left(D_{L'}\right) \\ & {\scriptstyle \Phi_L} \searrow \quad \swarrow {\scriptstyle \Phi_{L'}} & \\ & H_*\left(R_{L-\epsilon}(r)\right) \longrightarrow H_*\left(R_{L'-\epsilon}(r)\right) & \end{array} \tag{32}$$

i.e. compare $\Phi_{L'} \circ \imath^D_{L\to L'} : D_L \to sd(R_{L'-\epsilon}(r))$ to $\imath^R_{L-\epsilon\to L'-\epsilon} \circ \Phi_L : D_L \to sd(R_{L'-\epsilon}(r))$. For $\gamma \in B_*\left(D_L\right)$, note that $\Phi_{L'} \circ \imath^D_{L\to L'}(\gamma) = \phi^r_{L'-\epsilon}(\gamma)$ and $\imath^R_{L-\epsilon\to L'-\epsilon} \circ \Phi_L(\gamma) = \phi^r_{L-\epsilon}(\gamma)$. Then from (23),

$$\left(\Phi_{L'} \circ \imath^D_{L\to L'}\right)_* [\gamma] = \left(\phi^r_{L'-\epsilon}\right)_* [\gamma] = \left(\phi^r_{L-\epsilon}\right)_* [\gamma] = \left(\imath^R_{L-\epsilon\to L'-\epsilon} \circ \Phi_L\right)_* [\gamma]$$

in $H_*\left(sd\left(\check{C}_{L'-\epsilon}(r)\right)\right)$. Since $\check{C}_{L'-\epsilon}(r) \subset R_{L'-\epsilon}(r)$, the same relation holds in $H_*\left(sd\left(R_{L'-\epsilon}(r)\right)\right)$ as well. □

Theorem 6. *The confidence set $\widehat{C}^{Cech}_\alpha$ in* (15) *is asymptotically valid and satisfies*

$$\mathbb{P}\left(d_B\left(\mathrm{PH}^{Cech}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \ge 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right).$$

Similarly, the confidence set $\widehat{C}^{Rips}_\alpha$ in (16) *is asymptotically valid and satisfies*

$$\mathbb{P}\left(d_B\left(\mathrm{PH}^{R}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_{2r}\right) \ge 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right).$$

Proof of Theorem 6. Theorem 5 and Theorem 9 together imply as

$$\begin{aligned} \mathbb{P}\left(d_B\left(\mathrm{PH}^{Cech}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) &\ge \mathbb{P}\left(\|p_h - \widehat{p}_h\|_\infty + \widehat{c}_r \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_r\right) \\ &= \mathbb{P}\left(\sqrt{nh^d}\|p_h - \widehat{p}_h\|_\infty \le \widehat{z}_\alpha\right) \\ &= 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right), \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}\left(d_B\left(\mathrm{PH}^{R}_*(\widehat{p}_h, r), \mathrm{PH}_*(p_h)\right) \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_{2r}\right) &\ge \mathbb{P}\left(\|p_h - \widehat{p}_h\|_\infty + \widehat{c}_{2r} \le \frac{\widehat{z}_\alpha}{\sqrt{nh^d}} + \widehat{c}_{2r}\right) \\ &= \mathbb{P}\left(\sqrt{nh^d}\|p_h - \widehat{p}_h\|_\infty \le \widehat{z}_\alpha\right) \\ &= 1 - \alpha + O\left(\sqrt{\frac{1}{n}}\right), \end{aligned}$$

□