On the estimation of persistence intensity functions and linear representations of persistence diagrams

Weichen Wu
Carnegie Mellon University

Jisu Kim
Seoul National University

Alessandro Rinaldo
The University of Texas at Austin

Abstract

Persistence diagrams are one of the most popular types of data summaries used in Topological Data Analysis. The prevailing statistical approach to analyzing persistence diagrams is concerned with filtering out topological noise. In this paper, we adopt a different viewpoint and aim at estimating the actual distribution of a random persistence diagram, which captures both topological signal and noise. To that effect, Chazal and Divol (2019) proved that, under general conditions, the expected value of a random persistence diagram is a measure admitting a Lebesgue density, called the persistence intensity function. In this paper, we are concerned with estimating the persistence intensity function and a novel, normalized version of it – called the persistence density function. We present a class of kernel-based estimators based on an i.i.d. sample of persistence diagrams and derive estimation rates in the supremum norm. As a direct corollary, we obtain uniform consistency rates for estimating linear representations of persistence diagrams, including Betti numbers and persistence surfaces. Interestingly, the persistence density function delivers stronger statistical guarantees.

1 INTRODUCTION

Topological Data Analysis (TDA) is a field at the interface of computational geometry, algebraic topology and data science whose primary objective is to extract topological and geometric features from possibly high-dimensional, noisy and/or incomplete data. See Chazal and Michel (2021b) and references therein for a recent review. The literature on the statistical analysis of TDA summaries has primarily focused on separating topological signatures from the unavoidable topological noise resulting from the data sampling process. In most cases, the primary goal of statistical inference methods for TDA is to isolate points on the sample persistence diagrams that are sufficiently far from the diagonal to be deemed statistically significant, in the sense of expressing underlying topological signals rather than randomness. This paradigm is entirely natural when the target of inference is one unobservable persistence diagram, and the sample persistent diagrams are noisy approximations to it. Towards that goal, practitioners can now deploy a variety of statistical techniques for identifying topological signals and removing topological noise with provable theoretical guarantees.

On the other hand, empirical evidence has also demonstrated that topological noise is not necessarily unstructured or uninformative and, in fact, may also carry expressive and discriminative power that can be leveraged for various machine-learning tasks. In some applications, the distribution of the topological noise itself is of interest; in cosmology, see e.g., Wilding et al. (2021). As a result, statistical summaries able to express the properties of both topological signal and topological noise in a unified manner have also been proposed and investigated: e.g., persistence images and linear functional of the persistence diagrams (Adams et al., 2017).

Recently, Chazal and Divol (2019) derived sufficient conditions to ensure that the expected persistent measure – the expected value of the random counting measure corresponding to a noisy persistent diagram

– admits a Lebesgue density, hereafter the *persistence intensity function.* The significance of this result is multifaceted. First, the persistent intensity function provides an explicit and interpretable representation of the entire distribution of the persistence homology of random filtrations. Secondly, it allows for a straightforward calculation of the expected value of any linear representation of a persistent diagram as a Lebesgue integral. Finally, the representation by the persistence intensity function is of functional, as opposed to algebraic, nature and thus can be estimated via well-established theories and methods from the non-parametric statistics functional estimation. Indeed, Chazal and Divol (2019) analyzed a kernel-based estimator of the persistence intensity function computed using a sample of i.i.d. persistence diagrams and proved its L_2 consistency. Similar results were previously established by Chen et al. (2015).

In this paper, we derive consistency rates of estimation of the persistence intensity function and of a novel variant called persistence density function in the ℓ_∞ norm based on a sample of i.i.d. persistent diagrams. As we argue below in Theorem 3.1, controlling the estimation error for the persistence intensity function in the ℓ_∞ norm is stronger than controlling the optimal transport measure OT_q for any $q > 0$ and, under mild assumptions, immediately implies uniform control and concentration of any bounded linear representation of the persistence diagram, including (persistent) Betti numbers and persistence images. Our analysis and results are different from and complement those results of Chazal and Divol (2019) and Chen et al. (2015); in particular, we seek to establish finite sample ℓ_∞ estimation guarantees, a challenging task demanding more sophisticated techniques.

We emphasize that the approach and methods considered in this paper are distinct from the prevailing practices in statistical inference for TDA, which focus on extracting topological signals. In contrast, we are interested in capturing the overall randomness of persistence diagrams and describing the topological noise arising from sampling. That is, we aim to describe the *distribution* of a random persistence diagram, not any particular realization of it or a target persistence diagram. A second notable point of departure from mainstream TDA is that we assume the availability of an i.i.d. sample of random persistent diagrams, and the accuracy of our rates improve as the number of persistent diagrams increases, not the size of the data used to compute each diagram, which we hold fixed. This type of asymptotics is of course markedly different than the one traditionally considered by the TDA literature, which has focussed on the large sample properties of *one* persistent diagram computed using one
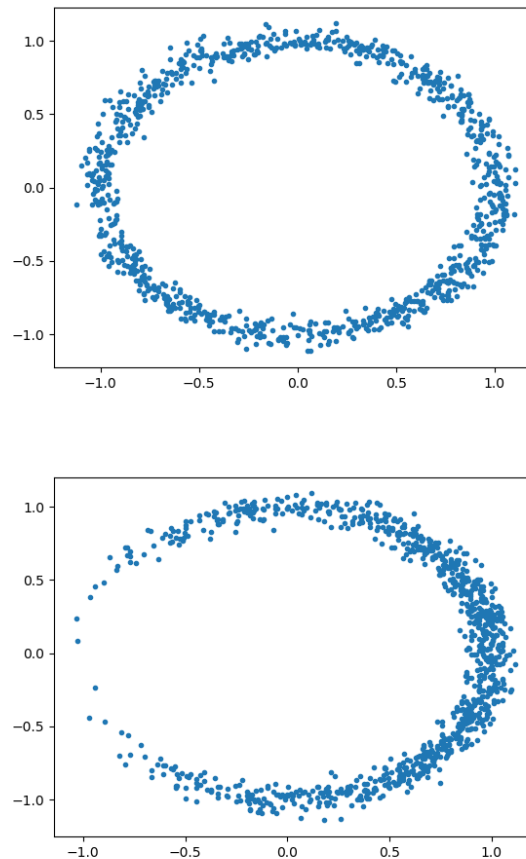

Figure 1: Sample plots for the uniform distribution (top) and power spherical distribution Cao and Aziz (2020) (bottom) on the unit circle $\mathbb{S}^1$. The parameters for the power spherical distribution are set to $\mu = \frac{\pi}{2}$ and $\kappa = 1$. Each sample contains 1000 points generated *i.i.d.* from the distributions on the unit circle corrupted by a $N(0, 0.05^2 \boldsymbol{I}_2)$ additive error.

dataset. In both regards, our perspective is rather separate from the current TDA paradigm and is not intended as an alternative framework. As an illustrative example, suppose that we are interested in the distribution of persistence diagrams originating from a uniform distribution and from a non-uniform distribution on the unit sphere, e.g. the power spherical distribution (Cao and Aziz, 2020). The target topological signature (corresponding to the homology of the unit sphere) is the same in the two cases but the topological noise is different. The sample plots in Figure 1 illustrate the difference between the two distributions, and Figure 2 shows the estimated persistence intensities (for the Vietoris-Rips filtration) based on a sample of 1000 i.i.d. persistence diagrams. The difference in the distributions of the topological noise is apparent. In contrast, such difference may not be discernible from inspecting
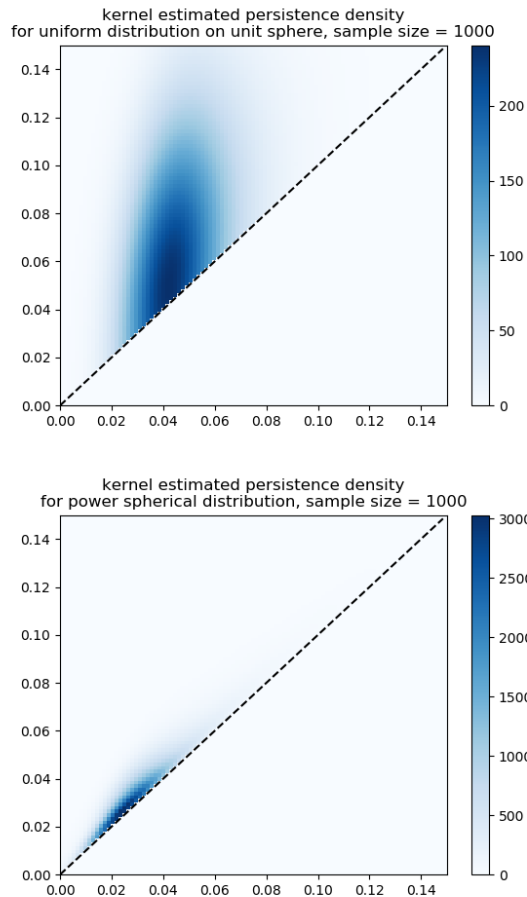
Figure 2: Estimated persistence density functions for the uniform distribution (top) and power spherical distribution (bottom) with corresponding parameters on the unit circle $\mathbb{S}^1$, based on 1000 diagrams.

individual persistence diagrams; see Figure 9 in the supplementary material.

The approach to TDA adopted in this paper conforms with common data-analytic practices used in machine learning tasks. In many applications, persistence diagrams are deployed as feature maps, embedding the data into a space with high expressive power, and are then fed as inputs to machine learning and even deep learning models. For example, an individual image, the output of one set of simulations, a time series, a protein, or a social network can each be represented as persistence diagrams. For a list of related work and comprehensive examples, we refer the reader to the review papers Hensel et al. (2021) and Barnes et al. (2021).

2 BACKGROUND AND DEFINITIONS

In this section we introduce fundamental concepts from TDA that we will use throughout. We refer the reader to (Chazal and Michel, 2021a; Chazal and Divol, 2019) for background and extensive references.

Persistence diagrams. A persistence diagram is a locally finite multiset of points $D = \{\boldsymbol{r}_i = (b_i, d_i) \mid 1 \leq i \leq N(D)\}$ belonging to the set

$$\Omega = \Omega(L) = \{(b, d) \mid 0 < b < d \leq L\} \subset \mathbb{R}^2, \quad (1)$$

consisting of all the points on the plane in the positive orthant above the identity line and of magnitude no larger than a given constant $L > 0$. The restriction that the persistence diagrams be contained in a box of side length L is a technical assumption is widely used in the TDA literature; see Divol and Lacombe (2021) and the discussion therein. To simplify our notation, we will omit the dependence on L, but we will keep track of this parameter in our error bounds. Some related quantities used throughout are

$$\partial\Omega := \{(x, x) \mid 0 \leq x \leq L\}; \qquad \overline{\Omega} := \Omega \cup \partial\Omega;$$

$$\Omega_\ell := \left\{\boldsymbol{\omega} \in \partial\Omega \big| \|\boldsymbol{x} - \partial\Omega\|_2 := \min_{x \in \Omega} \|\boldsymbol{\omega} - \boldsymbol{x}\|_2 \geq \ell\right\},$$

$$\text{for} \quad \ell \in (0, L/\sqrt{2}). \quad (2)$$

That is, $\partial\Omega$ is a segment on the diagonal in $\mathbb{R}^2$ and Ω_ℓ consists of all the points in Ω at a Euclidean distance of ℓ or smaller from it.

The expected persistent measure and its normalization. A persistence diagram $D = \{\boldsymbol{r}_i = (b_i, d_i) \in \Omega \mid 1 \leq i \leq N(D)\}$ can be equivalently represented as a counting measure μ on Ω given by

$$A \in \mathcal{B} \mapsto \mu(A) = \sum_{i=1}^{N(D)} \delta_{\boldsymbol{r}_i}(A),$$

where $\mathcal{B} = \mathcal{B}(\Omega)$ is the class of all Borel subsets of Ω and $\delta_{\boldsymbol{r}}$ denotes the Dirac point mass at $\boldsymbol{r} \in \Omega$. We will refer to μ as the *persistence measure* corresponding to D and, with a slight abuse of notation, will treat persistence diagrams as counting measures. If D is a random persistence diagram, then the associated persistence measure is also random. We will also study its *normalized measure* $\tilde{\mu}$, which is the persistence measure divided by the total number of points $N(D)$ in the persistence diagram:

$$A \in \mathcal{B} \mapsto \tilde{\mu}(A) = \frac{1}{N(D)} \sum_{i=1}^{N(D)} \delta_{\boldsymbol{r}_i}(A).$$

The normalized persistence measure may be more appropriate when the number of points $N(D)$ in the persistence diagram is not of direct interest but their spatial distribution is. This is typically the case when the persistence diagrams at hand contain many points or are obtained from large random filtrations (e.g. the Vietoris-Rips complex built on point clouds), so that the value of $N(D)$ will mostly account for noisy topological fluctuations due to sampling.

We will consider the setting in which the observed persistence diagram D is a random draw from an unknown distribution. Then, the (non-random) measures

$$A \in \mathcal{B} \mapsto \mathbb{E}[\mu](A) = \mathbb{E}[\mu(A)] \quad \text{and}$$
$$A \in \mathcal{B} \mapsto \mathbb{E}[\tilde{\mu}](A) = \mathbb{E}[\tilde{\mu}(A)]$$

are well defined. We will refer to $\mathbb{E}[\mu]$ and $\mathbb{E}[\tilde{\mu}]$ as the *expected persistence measure* and the *expected persistence probability,* respectively. Neither is a discrete measure (even though persistence measures are discrete by definitions). Of course, the expected persistence probability $\mathbb{E}[\tilde{\mu}]$ is a probability measure.

The interpretations of $\mathbb{E}[\mu]$ and of $\mathbb{E}[\tilde{\mu}]$ are straightforward: for any Borel set $A \subset \Omega$, $\mathbb{E}[\mu](A)$ is the expected number of points from the random persistence diagram falling in A, while $\mathbb{E}[\tilde{\mu}](A)$ is the probability that a random persistence diagram will intersect A. Despite their interpretability, the expected persistence measure and probability are not yet standard concepts in the practice and theory of TDA. As a result, they have not been thoroughly investigated.

The persistence intensity and density functions and linear representations. Recently, Chazal and Divol (2019) derived conditions – applicable to a wide range to problems – that ensure that the expected persistence measure $\mathbb{E}[\mu]$ and its normalization $\mathbb{E}[\tilde{\mu}]$ both admit densities with respect to the Lebesgue measure on Ω. Specifically, under fairly mild and general conditions detailed in Chazal and Divol (2019) there exist measurable functions $p : \Omega \to \mathbb{R}_{\geq 0}$ and $\tilde{p} : \Omega \to \mathbb{R}_{\geq 0}$, such that for any Borel set $A \subset \Omega$,

$$\mathbb{E}[\mu](A) = \int_A p(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \quad \mathbb{E}[\tilde{\mu}](A) = \int_A \tilde{p}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \quad (3)$$

In fact, Chazal and Divol (2019) provided explicit expressions for p and $\tilde{p}$ (see Section E.8). Notice that, by construction, $\tilde{p}$ integrates to 1 over Ω. We will refer to the functions p and $\tilde{p}$ as the *persistence intensity* and the *persistence density* functions, respectively. We remark that the notion of a persistence intensity function was originally put forward by Chen et al. (2015).

The persistence intensity and density functions "operationalize" the notions of expected persistence measure

and expected persistence probability introduced above, allowing us to evaluate, for any set $A \in \mathcal{B}$, $\mathbb{E}[\mu](A)$ and $\mathbb{E}[\tilde{\mu}](A)$ in a straightforward way as Lebesgue integrals. The main objective of the paper is to construct estimators of the persistence intensity p and persistence density $\tilde{p}$, respectively, and to provide high probability error bounds with respect to the L_∞ norm. As we show below in Theorem 3.1, L_∞-consistency for the persistence intensity function is a stronger guarantee than consistency in the OT_p metric, for any $p < \infty$.

As noted in Chazal and Divol (2019), the persistence intensity and density functions are naturally suited to compute the expected value of linear representations of random persistence diagrams. A linear representation Ψ of the persistence diagram $D = \{\boldsymbol{r}_i = (b_i, d_i) \in \Omega \mid 1 \leq i \leq N(D)\}$ with corresponding persistence measure μ is a summary statistic of D of the form

$$\Psi(D) = \sum_{i=1}^{N(D)} f(\boldsymbol{r}_i) = \int_\Omega f(\boldsymbol{u})d\mu(\boldsymbol{u}), \quad (4)$$

for a given measurable function f on Ω. (An analogous definition can be given for the normalized persistence measure $\tilde{\mu}$ instead). Then,

$$\mathbb{E}[\Psi(D)] = \int_\Omega f(\boldsymbol{u})d\mathbb{E}[\mu](\boldsymbol{u}) = \int_\Omega f(\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}, \quad (5)$$

where the second identity follows from (3). Linear representations include persistent Betti numbers, persistence surfaces (Adams et al., 2017), persistence silhouettes, Chazal et al. (2013) and persistence weighted Gaussian kernels (Kusano et al., 2016). The persistence surface is an especially popular linear representation. In detail, for a *kernel function* $K(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ and any $\boldsymbol{x} \in \mathbb{R}^2$, let $K_h(\boldsymbol{x}) = \frac{1}{h^2}K(\frac{\boldsymbol{x}}{h})$, where $h > 0$ is the bandwidth parameter[1]. The persistence surface of a persistence measure μ is defined as

$$\rho_h(\boldsymbol{u}) = \int_\Omega f(\boldsymbol{\omega})K_h(\boldsymbol{u} - \boldsymbol{\omega})\mathrm{d}\mu(\boldsymbol{\omega}), \quad (6)$$

where $f(\boldsymbol{\omega}) \colon \mathbb{R}^2 \to \mathbb{R}$ is the user-defined *weighting function,* chosen to ensure stability of the representation. Our analysis allows us to immediately obtain consistency rates for the expected persistence surface in L_∞ norm; see Theorem C.10 in the supplementary material.

Betti and the persistent Betti numbers. The **Betti number** at scale $x \in [0, L]$ is the number of persistent homologies that are in existence at "time" x. Furthermore, the **persistent Betti number** at a

---

[1](Adams et al., 2017) showed empirically that the bandwidth does not have a major influence on the efficiency of the persistence surface.

certain point $\boldsymbol{x} = (x_1, x_2) \in \Omega$ measures the number of persistent homologies that are born before x_1 and die after x_2. In our notation, given a persistence diagram D and its associated persistence measure μ, for $x \in [0, L]$ and $\boldsymbol{x} = (x_1, x_2) \in \Omega$, the corresponding Betti number and persistent Betti number are given by

$$\beta_x = \mu(B_x) \quad \text{and} \quad \beta_{\boldsymbol{x}} = \mu(B_{\boldsymbol{x}}),$$

respectively, where $B_x = [0, x) \times (x, L]$ and $B_{\boldsymbol{x}} = [0, x_1) \times (x_2, L]$. Though Betti numbers are among the most prominent and widely used TDA summaries, relatively little is known about the statistical hardness of estimating their expected values when the sample size is fixed and the number of persistence diagrams increases. Our results will yield error bounds of this type. We will also consider normalized versions of the Betti numbers defined using the persistence probability $\tilde{\mu}$ of the persistence diagram:

$$\tilde{\beta}_x = \tilde{\mu}(B_x) \quad \text{and} \quad \tilde{\beta}_{\boldsymbol{x}} = \tilde{\mu}(B_{\boldsymbol{x}}).$$

Notice that, by definition, $\tilde{\beta}_{\boldsymbol{x}} \leq 1$. While their interpretation is not as direct as the Betti numbers computed using persistence diagrams, the expected normalized (persistence) Betti numbers $\mathbb{E}[\tilde{\beta}_{\boldsymbol{x}}] = \mathbb{E}[\tilde{\mu}](B_{\boldsymbol{x}})$ are informative topological summaries while showing favorable statistical properties (see Corollary 3.8 below).

3 RESULTS

3.1 On the OT distance and the L_∞ distance between intensity functions

We first show that the topology induced by the L_∞ distance between intensity functions is stronger than the one corresponding to the *optimal transport* distance, a natural and very popular metric for persistence diagrams – and, more generally, locally finite Radon measures such as normalized persistence measures and probabilities; see, in particular, (Divol and Lacombe, 2021). In detail, for two Radon measures μ and ν supported on $\overline{\Omega}$, an *admissible transport* from μ to ν is defined as a function $\pi : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$, such that for any Borel sets $A, B \subset \overline{\Omega}$,

$$\pi(A \times \overline{\Omega}) = \mu(A), \quad \text{and} \quad \pi(\overline{\Omega} \times B) = \nu(B).$$

Let $\mathsf{adm}(\mu, \nu)$ denote all the admissible transports from μ to ν. For any $q \in \mathbb{R}^+ \cup \{\infty\}$, the q-th order Optimal Transport (OT) distance between μ and ν is defined as

$$\mathsf{OT}_q(\mu, \nu) = \left(\inf_{\pi \in \mathsf{adm}(\mu, \nu)} \int_{\overline{\Omega} \times \overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \right)^{\frac{1}{q}}.$$

The OT distance is widely used for good reasons: by transporting from and to the diagonal $\partial\Omega$, it captures the distance between two measures that have potentially different total masses, taking advantage of the fact that points on the diagonal have arbitrary multiplicity in persistent diagrams. It also proves to be stable with respect to perturbations of the input to TDA algorithms. It turns out that the L_∞ distance between intensity functions provides a tighter control on the difference between two persistent measures. Below, for a real-valued function on Ω, we let $\|f\|_\infty = \sup_{\boldsymbol{x} \in \Omega} |f(\boldsymbol{x})|$ be its L_∞ norm.

Theorem 3.1. *Let μ, ν be two expected persistent measures on Ω with intensity functions p_μ and p_ν respectively. Then*

$$\mathsf{OT}_q^q(\mu, \nu) \leq \frac{2}{(q+1)(q+2)} \left(\frac{L}{\sqrt{2}} \right)^{q+2} \|p_\mu - p_\nu\|_\infty.$$

Furthermore, there exist two sequences of expected persistence measures $\{\mu_n\}_{n \in \mathbb{N}}$ and $\{\nu_n\}_{n \in \mathbb{N}}$ with intensity functions $\{p_{\mu_n}\}_{n \in \mathbb{N}}$ and $\{p_{\nu_n}\}_{n \in \mathbb{N}}$ respectively such that, as $n \to \infty$,

$$\mathsf{OT}_q(\mu_n, \nu_n) \to 0, \quad \text{while} \quad \|p_{\mu_n} - p_{\nu_n}\|_\infty \to \infty.$$

The bottleneck distance For the case of $q = \infty$, which corresponds to the bottleneck distance when applied to persistence diagrams, there can be no meaningful upper bound of the form of Theorem 3.1: we show in Section E.1 of the supplementary material that there exist two sequences of measures such that their bottleneck distance converges to a finite number while the L_∞ distance between their intensity functions vanishes. Existing contributions in the optimal transport literature (Peyre, 2018; Nietert et al., 2021) also upper bound the optimal transport distance by a Sobolev-type distance between density functions. Notably, these bounds require, among other things, the measures to have common support and the same total mass, two conditions not assumed in Theorem 3.1.

3.2 Non-parametric estimation of the persistent intensity and density functions

In this section, we analyze the performance of kernel-based estimators of the persistent intensity function and the persistent density function in the same setting considered by Chazal and Divol (2019) and Chen et al. (2015), where we observe n i.i.d. persistent measures (i.e. diagrams) $\mu_1, \mu_2, \ldots, \mu_n$. The proposed procedures are inspired by kernel density estimators for probability densities traditionally used in the non-parametric statistics literature; see, e.g., Giné and Nickl (2021). Specifically, we consider the following estimator

for p and $\tilde{p}$, respectively:

$$\boldsymbol{\omega} \in \mathbb{R}^2 \mapsto \hat{p}_h(\boldsymbol{\omega}) := \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} K_h(\boldsymbol{x} - \boldsymbol{\omega}) \mathrm{d}\mu_i(\boldsymbol{x}); \quad (7a)$$

$$\boldsymbol{\omega} \in \mathbb{R}^2 \mapsto \check{p}_h(\boldsymbol{\omega}) = \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} K_h(\boldsymbol{x} - \boldsymbol{\omega}) \mathrm{d}\tilde{\mu}_i(\boldsymbol{x}), \quad (7b)$$

where $K(\cdot)$ is a *kernel function,* which we assume to satisfy standard conditions used in non-parametric literature, discussed in detail in Section C.2 of the supplementary material.

Assumptions. We require several regularity conditions. Furthermore, we will implicitly assume throughout that both p and $\tilde{p}$ (see 3) are well-defined as densities with respect to the Lebesgue measure, though this is not strictly necessary for our main results, Theorems 3.4 and 3.7.

First, we assume a uniform bound on the q-th order total persistence of the μ_i's, though not on the total number of points in the persistence diagrams. This can be thought of as a basic moment existence condition that, as elucidated in Cohen-Steiner et al. (2010) and discussed in Divol and Polonik (2019) and Divol and Lacombe (2021), is a relatively mild assumption satisfied by a broad variety of data-generating mechanisms.

Assumption 3.2 (Bounded total persistence)**.** *There exists a constant $M > 0$, such that, for the value of q as in Assumption 3.3, it holds that, almost surely,*

$$\max_{i=1,\ldots,n} \int_{\Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q \mathrm{d}\mu_i(\omega) < M.$$

We will denote with $\mathcal{Z}_{L,M}^q$ the set of persistent measures on Ω_L satisfying Assumption 3.2.

Next, we impose key boundedness conditions on p and $\tilde{p}$, which are needed to apply a concentration inequality for empirical processes that deliver uniform control over the variances of these estimators.

Assumption 3.3 (Boundedness)**.** *For some $q > 0$, let $\bar{p}(\boldsymbol{\omega}) := \|\boldsymbol{\omega} - \partial\Omega\|_2^q p(\boldsymbol{\omega})$. Then,*

$$\|\bar{p}\|_\infty = \sup_{\omega \in \Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q p(\boldsymbol{\omega}) < \infty \quad and$$

$$\|\tilde{p}\|_\infty = \sup_{\omega \in \Omega} \tilde{p}(\boldsymbol{\omega}) < \infty.$$

We remark that a bound on the L_∞ norm of the intensity function p is not a realistic assumption because the total mass of the persistence measure may not be uniformly bounded in several common datagenerating mechanisms. Indeed, in light of existing results, this condition would be likely violated in many scenarios; see e.g. Divol and Polonik (2019)). Thus,

we only require that the weighted intensity function $\bar{p}(\boldsymbol{\omega}) = \|\boldsymbol{\omega} - \partial\Omega\|_2^q p(\boldsymbol{\omega})$ has finite L_∞ norm. Still, it is not a priori clear that Assumption 3.3 for $\bar{p}$ itself is realistic; in the supplementary material, we prove that this is indeed the case for the Vietoris-Rips filtration built on i.i.d. samples. On the other hand, assuming that the persistence density $\tilde{p}$ is uniformly bounded poses no problems. For a formal argument, see Theorems C.1 and C.2 in the supplementary material. This fact is the primary reason why the persistence probability density function – unlike the persistence intensity function – can be estimated uniformly well over the entire set Ω - see (3.4) below. We refer readers to Section C.1 of the supplementary materials for details and a discussion on this subtle but consequential point.

One of the main results of the paper are high probability uniform bounds on the fluctuations of the kernel estimators around their expected values. For a fixed value of the bandwidth h, they imply that the estimators $\hat{p}_h$ and $\check{p}_h$ concentrate around their expected value at a parametric rate $1/\sqrt{n}$.

Theorem 3.4. *Suppose that Assumptions 3.2 and 3.3 hold. Then,*

(a) *there exist positive constants C_1, C_2 depending on $M, \|K\|_\infty, \|K\|_2, \|\bar{p}\|_\infty$ and q such that for any $\delta \in (0,1)$, it can be guaranteed with probability at least $1 - \delta$ that*

$$\sup_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^q |\hat{p}_h(\boldsymbol{\omega}) - \mathbb{E}\hat{p}_h(\boldsymbol{\omega})|$$

$$\leq \max\left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\},$$

where $\ell_\omega := \|\boldsymbol{\omega} - \partial\Omega\|_2 - h$;

(b) *there exist positive constants C_1, C_2 depending on $M, \|K\|_\infty, \|K\|_2, \|\tilde{p}\|_\infty$ and q such that for any $\delta \in (0,1)$, it can be guaranteed with probability at least $1 - \delta$ that*

$$\sup_{\boldsymbol{\omega} \in \Omega} |\check{p}_h(\boldsymbol{\omega}) - \mathbb{E}\check{p}_h(\boldsymbol{\omega})|$$

$$\leq \max\left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\}.$$

Remark. The dependence of the constants on problem related parameters is made explicit in the proofs; see the supplementary material.

There is an important difference between the two bounds in Theorem 3.4: while the variation of $\check{p}_h(\boldsymbol{\omega})$ is uniformly bounded everywhere on Ω, the variation of $\hat{p}_h(\boldsymbol{\omega})$ is uniformly bounded only when $\boldsymbol{\omega}$ is at least $2h$ away from the diagonal $\partial\Omega$, and may increase as $\boldsymbol{\omega}$

approaches the diagonal. The difficulty in controlling $\hat{p}_h$ near the diagonal stems from the fact that we only assume the total persistence to be bounded; in other words, the number of points near the diagonal in the sample persistent diagrams can be prohibitively large, since their contribution to the total persistence is negligible. This is expected in noisy settings where the sampling process will result in topological noise consisting of many points in the persistence diagram near the diagonal. We do not know whether this limitation of the estimator $\hat{p}_h$ is intrinsic to the problem or instead an artifact of our proof techniques. Nonetheless, the above result suggests that to achieve uniform control over Ω, relying on density-based rather than intensity-based representations of the persistent measures may be preferable.

Bias-variance trade-off and minimax lower bound. In order to measure how well $\hat{p}_h$ and $\check{p}_h$ concentrate not just around their expectations but around the target densities p and $\tilde{p}$, respectively, we will need to further control their biases, as a function of the bandwidth h. To that effect, we require some degree of smoothness of both p and $\tilde{p}$, as it is standard in non-parametric density estimation. We refer the reader to the appendix for the definition of smooth function spaces.

Assumption 3.5 (Smoothness)**.** *The persistence intensity function p and persistence probability density function $\tilde{p}$ are Hölder smooth of the order of $s > 0$ with parameters L_p and $L_{\tilde{p}}$, respectively.*

Using the above assumption and standard arguments, we obtain that, uniformly over $\boldsymbol{\omega} \in \Omega$, $|\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega})|$ and $|\mathbb{E}[\check{p}_h(\boldsymbol{\omega})] - \tilde{p}(\boldsymbol{\omega})|$ are both of order h^2. See Theorem C.6 in the appendix. Next, assuming that the number n of persistent diagram grows unbounded, it follows from Theorems C.6 and 3.4 that setting the bandwidth to be $h \asymp n^{-\frac{1}{2(s+1)}}$ will optimize the bias-variance trade-off, yielding high-probability estimation errors

$$\sup_{\boldsymbol{w} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^q |\hat{p}_h(\boldsymbol{\omega}) - p(\boldsymbol{\omega})| \lesssim O\left(n^{-\frac{s}{2(s+1)}}\right), \quad \text{and}$$

$$\sup_{\boldsymbol{w} \in \Omega} |\check{p}_h(\boldsymbol{\omega}) - \tilde{p}(\boldsymbol{\omega})| \lesssim O\left(n^{-\frac{s}{2(s+1)}}\right).$$

In our next result, we show that the above rate is minimax optimal for the persistence density function. For brevity, we here omit a similar result for the persistence intensity function (see Theorem C.9 in the supplementary material).

Theorem 3.6. *Let $\mathscr{F}$ denote the set of functions on Ω with Besov norm bounded by $B > 0$:*

$$\mathscr{F} = \{f : \Omega \to \mathbb{R}, \|f\|_{B_{\infty,\infty}^s} \leq B\}.$$

Then,

$$\inf_{\check{p}_n} \sup_P \mathbb{E}_{\mu_1,\ldots,\mu_n \overset{i.i.d.}{\sim} P} \|\check{p}_n - \tilde{p}\|_\infty \geq O(n^{-\frac{s}{2(s+1)}}),$$

where the infimum is taken over estimator $\check{p}_n$ mapping $\mu_1, \ldots, \mu_n$ to an intensity function in $\mathscr{F}$, the supremum is over the set of all probability distributions on $\mathcal{Z}_{L,M}^q$ and $\tilde{p}$ is the intensity function of $\mathbb{E}_P[\tilde{\mu}]$.

3.3 Kernel-based estimators for linear functionals of the persistent measure

The kernel estimators (7) can serve as a basis for estimating bounded linear representations of the expected persistence measure $\mathbb{E}[\mu]$ and its normalized counterpart $\mathbb{E}[\tilde{\mu}]$. Specifically, for $R > 0$, let $\mathscr{F}_{2h,R}$ and $\widetilde{\mathscr{F}}_R$ denote the set of linear representations of the form

$$\mathscr{F}_{2h,R} = \left\{\Psi = \int_{\Omega_{2h}} f \mathrm{d}\mathbb{E}[\mu] \middle| f : \Omega_{2h} \to \mathbb{R}_{\geq 0}, \right.$$

$$\left. \int_{\Omega_{2h}} \ell_{\boldsymbol{\omega}}^{-q} f(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \leq R\right\}, \quad \text{and}$$

$$\widetilde{\mathscr{F}}_R = \left\{\Psi = \int_\Omega f \mathrm{d}\mathbb{E}[\tilde{\mu}] \middle| f : \Omega \to \mathbb{R}_{\geq 0}, \int_\Omega f(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \leq R\right\}.$$

Then, any linear representations $\Psi \in \mathscr{F}_{2h,R}$ and $\widetilde{\Psi} \in \widetilde{\mathscr{F}}_R$ can be estimated by

$$\hat{\Psi}_h = \int_{\Omega_{2h}} f(\boldsymbol{\omega}) \hat{p}_h(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \text{ and } \check{\Psi}_h = \int_\Omega f(\boldsymbol{\omega}) \check{p}_h(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}.$$
$$(8)$$

As a direct corollary of Theorem C.6, we obtain the following uniform high-probability bound on the variance of $\hat{\Psi}_h$ and $\check{\Psi}_h$, which, for fixed h, yield $1/\sqrt{n}$ rates. In the supplementary material, we also show that, not surprisingly, the biases of both estimators are of order h^s under Assumption 3.5; see Theorem C.7 in the supplementary material.

Theorem 3.7. *Assume that Assumptions 3.2 and 3.3 hold. Then,*

(a) there exist constants C_1, C_2 depending on $M, \|K\|_\infty, \|K\|_2, \|\bar{p}\|_\infty$ and q such that for any $\delta \in (0,1)$, it can be guaranteed with probability at least $1 - \delta$ that

$$\sup_{\Psi \in \mathscr{F}_{2h,R}} \left|\hat{\Psi}_h - \mathbb{E}[\hat{\Psi}_h]\right|$$

$$\leq R \cdot \max\left\{C_1 \frac{1}{nh^2} \log\frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log\frac{1}{\delta h^2}}\right\};$$

(b) there exist constants C_1, C_2 depending on $M, \|K\|_\infty, \|K\|_2, \|\tilde{p}\|_\infty$ and q such that for any

$\delta \in (0, 1)$, *it can be guaranteed with probability at least $1 - \delta$ that*

$$\sup_{\Psi \in \widetilde{\mathscr{F}}_R} \left| \check{\Psi}_h - \mathbb{E}[\check{\Psi}_h] \right|$$

$$\leq R \cdot \max \left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\}.$$

It is important to highlight the fact that the above bounds hold uniformly over the choice of linear representations under only mild integrability assumptions. We again stress the difference between the two upper bounds: part (a) shows that for a linear functional of the original persistent measure to have controlled variation, we need to be at least $2h$ away from the diagonal $\partial\Omega$, a requirement that is not necessary for linear functionals of the normalized persistent measure, as is shown in part (b).

We apply our results to the analysis of persistence surfaces and persistence Betti numbers. Due to space limitations, in the main text we focus on the latter and refer the reader to Theorem C.10 in the supplementary material for novel error rates in estimating persistent surfaces. For any $\boldsymbol{x} \in \Omega$, the persistent Betti number $\beta_{\boldsymbol{x}}$ can be estimated in a straightforward way by integrating $\hat{p}_h$ or $\check{p}_h$ over $B_{\boldsymbol{x}}$ (which we recall we define to be $B_{\boldsymbol{x}} = [0, x_1) \times (x_2, L]$):

$$\hat{\beta}_{\boldsymbol{x},h} = \int_{B_{\boldsymbol{x}}} \hat{p}_h(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \quad \check{\beta}_{\boldsymbol{x},h} = \int_{B_{\boldsymbol{x}}} \check{p}_h(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}. \quad (9)$$

An immediate application of 3.7 yields $1/\sqrt{n}$ high-probability concentration rates.

Corollary 3.8. *There exist a constants C_1, C_2 depending on $M, \|K\|_\infty, \|K\|_2, \|\bar{p}\|_\infty, \|\tilde{p}\|_\infty$ and $q > 2$ such that, for any $\delta \in (0, 1)$,*

(a) for the persistent Betti numbers computed using $\hat{p}_h$,

$$\sup_{\boldsymbol{x} \in \Omega : \ell_{\boldsymbol{x}} > h} \ell_{\boldsymbol{x}}^{q-2} \left| \hat{\beta}_{\boldsymbol{x},h} - \mathbb{E}[\hat{\beta}_{\boldsymbol{x},h}] \right|$$

$$\leq \max \left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\}$$

(b) for the persistent Betti numbers computed using $\check{p}_h$,

$$\sup_{\boldsymbol{x} \in \Omega} \left| \check{\beta}_{\boldsymbol{x},h} - \mathbb{E}[\check{\beta}_{\boldsymbol{x},h}] \right|$$

$$\leq \frac{L^2}{4} \max \left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\}$$

It is also straightforward to see that the biases of both $\hat{\beta}_{\boldsymbol{x},h}$ and $\check{\beta}_{\boldsymbol{x},h}$ are of order h^s, uniformly in $\boldsymbol{x}$; see Corollary C.8 in the supplementary material.

As noted before, the concentration rates of the estimator of the persistence Betti numbers based on the persistence density hold uniformly over Ω, thus suggesting that the kernel-based estimator $\hat{p}_h$ *will not be guaranteed to* yield a stable estimation of the Betti number $\beta_{\boldsymbol{x}}$. As remarked above, this issue arises as the intensity function may not be uniformly bounded near the diagonal. Indeed, in the supplementary material, we describe an alternative proof technique based on an extension of the standard VC inequality and arrive at a very similar rate. On the other hand, this issue does not affect the normalized Betti numbers $\tilde{\beta}_{\boldsymbol{x}}$.

An important consequence of the previous result is a *uniform* error bound for the *expected normalized Betti curve*

$$x \in (0, L) \mapsto \mathbb{E}[\tilde{\beta}_x] = \mathbb{E}[\tilde{\mu}](B_x) = \int_{B_x} \tilde{p}(\boldsymbol{\omega}) d\boldsymbol{\omega},$$

where $B_x = [0, x) \times (x, L]$ and $\tilde{\mu}$ is the normalized persistent measure corresponding to a random persistence diagram. In detail, for constants $C_1, C_2 > 0$ depending on the model parameters, with probability at least $1 - \delta$,

$$\sup_{x \in (0, L)} \left| \check{\beta}_{x,h} - \mathbb{E}[\tilde{\beta}_x]] \right|$$

$$\leq \max \left\{ C_1 \frac{1}{nh^2} \log \frac{1}{\delta h^2}, C_2 \sqrt{\frac{1}{nh^2}} \sqrt{\log \frac{1}{\delta h^2}} \right\} + h^s$$

To the best of our knowledge, this is the first result of this kind, as typically one can only establish pointwise and not uniform consistency of Betti numbers.

4 NUMERICAL ILLUSTRATIONS

To illustrate our methodology and highlight the differences between the persistence intensity and density functions, we compare the persistence intensity and density functions of 1000 data points drawn from the uniform distribution and the power sphere distribution Cao and Aziz (2020) on the unit circle $\mathbb{S}^1$. The density functions shown in Figure 2 illustrate a clear difference between the structure of topological noise generated by the two distributions. We include plots of the sample points, sample persistence diagrams and kernel-based estimators of persistence intensity functions in Section F of the supplementary material.

We also consider the MNIST handwritten digits dataset and the ORBIT5K dataset. The ORBIT5K dataset contains independent simulations for the

linked twist map, dynamical systems for fluid flow as described in Adams et al. (2017); see also Appendix G.2 of Kim et al. (2020). In Section F of the supplementary material, we show the estimated persistence intensity and density functions computed from persistence diagrams obtained over a varying number of random samples from the ORBIT5K datasets, for different model parameters. The figures confirm our theoretical finding that the values of the persistence density function near the diagonal are not as high (on a relative scale) as those of the persistence intensity function. An analogous conclusion can be reached when inspecting the persistence intensity and density functions for different draws of the MNIST datasets for the digits 4 and 8. We further include plots of the average Betti and normalized Betti curves from the ORBIT5K dataset, along with the curves of the empirical point-wise 5% and 95% quantiles. These plots reveal the different scales of the Betti curves and normalized Betti curves, and of their uncertainty. All of our code can be accessed through `https://github.com/Weichen-Wu-CMU/estimation_of_persistence_intensity_function`.

5 DISCUSSION

In this paper, we have taken the first step towards developing a new set of methods and theories for statistical inference for TDA based on i.i.d. samples of persistence diagrams. Our main focus is on the estimation of the persistence intensity function Chazal and Divol (2019); Chen et al. (2015), a TDA summary of a functional type that encodes the entire distribution of a random persistence diagram and is naturally suited to handle linear representations. We have analyzed a simple kernel estimator and derived uniform consistency rates that hold under very mild assumptions. We also propose the persistence density function, a novel functional TDA summary that enjoys stronger statistical guarantees.

A notable advantage of deploying persistence intensity and density functions to quantify the difference between distributions of persistence diagrams compared to more traditional approaches based on optimal transport distances is that our methodology is computationally feasible. Indeed, computing kernel-based estimators of the persistence intensity and density functions is a straightforward task even with very large sample sizes, and so is to evaluate any L_p distanced between them. In contrast, computing optimal transport distances between many persistence diagrams is typically computationally prohibitive.

There remain various open problems worth pursuing. A natural direction is the study of the topology over the space of normalized persistence measures. For example, based on our results from section 3, one may expect the normalized persistence measure not to be continuous for the vague topology with respect to the Hausdorff distance. Similarly, it would also be interesting to further investigate the topology induced by convergence of the persistence densities in the L_∞ norm. From the statistical side, our results guarantee the consistency of the proposed estimators. However, in order to carry out statistical inference, it is necessary to develop more sophisticated procedures that quantify the uncertainty of our estimators. Toward that goal, it would be interesting to develop bootstrap or other resampling-based methods for constructing confidence bands for both the persistence intensity and density functions.

References

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18, 2017.

Danielle Barnes, Luis Polanco, and A. Jose Perea. A comparative study of machine learning methods for persistence diagrams. *Frontiers in Artificial Intelligence*, 4, 2021.

Nicola De Cao and Wilker Aziz. The power spherical distribution, 2020.

Frédéric Chazal and Vincent Divol. The density of expected persistence diagrams and its kernel based estimation. *Journal of Computational Geometry*, 10 (2), 2019.

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

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry A. Wasserman. Stochastic convergence of persistence landscapes and silhouettes. *Proceedings of the thirtieth annual symposium on Computational geometry*, 2013.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021b.

Yen-Chi Chen, Daren Wang, Alessandro Rinaldo, and Larry Wasserman. Astatistical analysis of persistence intensity functions. *https://arxiv.org/pdf/1510.02502.pdf*, 2015.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have l p-stable persistence. *Foundations of computational mathematics*, 10(2):127–139, 2010.

Vincent Divol and Théo Lacombe. Estimation and quantization of expected persistence diagrams. In *International Conference on Machine Learning*, pages 2760–2770. PMLR, 2021.

Vincent Divol and Wolfgang Polonik. On the choice of weight functions for linear representations of persistence diagrams. *Journal of Applied and Computational Topology*, 3(3):249–283, 2019.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge university press, 2021.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4, 2021.

Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frederic Chazal, and Larry Wasserman. Pllay: Efficient topological layer based on persistent landscapes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15965–15977. Curran Associates, Inc., 2020.

Genki Kusano, Yasuaki Hiraoka, and Kenji Fukumizu. Persistence weighted gaussian kernel for topological data analysis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2004–2013, 20–22 Jun 2016.

Frank Morgan. *Geometric measure theory: a beginner's guide.* Academic press, 2016.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In *International Conference on Machine Learning*, pages 8172–8183. PMLR, 2021.

Rémi Peyre. Comparison between w_2 distance and $\dot{H}^{-1}$ norm, and localization of wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

Georg Wilding, Keimpe Nevenzeel, Rien van de Weygaert, Gert Vegter, Pratyush Pranav, Bernard J T Jones, Konstantinos Efstathiou, and Job Feldbrugge. Persistent homology of the cosmic web – I. Hierarchical topology in ΛCDM cosmologies. *Monthly Notices of the Royal Astronomical Society*, 507(2):2968–2990, 2021.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

On the estimation of persistence intensity functions and linear representations of persistence diagrams: Supplementary Materials

A NOTATION

We use boldface small letters like $\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{\omega}$ to denote points in $\mathbb{R}^2$ and sub-scripted letters like x_1, x_2 to denote their entries. Boldface capital letters like $\boldsymbol{X}, \boldsymbol{Y}$ would be used to denote points on a Riemann manifold. For any positive integer n, the symbol $[n]$ refers to the set of all positive integers no larger than n, i.e., $[n] \coloneqq \{1, 2, \ldots, n\}$. For any set S, the symbol 2^S represents the power set of S, which contains all subsets of S as its elements. The set of all non-negative real numbers would be denoted as $\mathbb{R}_{\geq 0}$. For any function f with domain $\mathcal{A}$, the infinity norm of f is denoted as $\|f\|_\infty \coloneqq \sup_{x \in \mathcal{A}} |f(x)|$.

B BACKGROUND: THE PERSISTENCE DIAGRAM

In this section, we give a brief introduction to the persistence diagram. We refer readers to Chazal and Divol (2019) for a detailed description. Consider a random point cloud $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N) \in \mathcal{M}^N$ where $\mathcal{M}$ is a Riemann manifold; and a *filtering function* $\varphi : 2^{[N]} \times \mathcal{M}^N \to \mathbb{R}$, which satisfies

$$\varphi(J, \boldsymbol{X}) \leq \varphi(J', \boldsymbol{X}), \quad \forall J \subset J' \in 2^{[N]}, \boldsymbol{X} \in \mathcal{M}^N.$$

A simplicial complex given $\boldsymbol{X}$ and φ at level α is defined as

$$K_\alpha(\boldsymbol{X}, \varphi) = \{J \subset 2^{[N]} \mid \varphi(J, \boldsymbol{X}) \leq \alpha\}.$$

Two common examples are the *Čech complex*, where $\varphi(J, \boldsymbol{X})$ equals the radius of the circumscribed ball of $\boldsymbol{X}[J]$; and the *Vietoris-Rips complex*, where $\varphi[J, \boldsymbol{X}]$ is chosen as the maximum distance between points in $\boldsymbol{X}[J]$.

Throughout the paper, we assume that the filtering function φ takes its value in $[0, L]$. For all values $\alpha \in [0, L]$, the sequence of simplicial complexes $\{K_\alpha(\boldsymbol{X}, \varphi)\}_{\alpha \in [0, L]}$ forms a *filtration* denoted as $\mathcal{F}(\boldsymbol{X}, \varphi)$, where $K_\alpha(\boldsymbol{X}, \varphi) \subseteq K_{\alpha'}(\boldsymbol{X}, \varphi)$ whenever $\alpha \leq \alpha'$.

Persistent homology is a method for computing topological features of a simplicial complex, and can be represented by the *persistence diagram*. In the filtration $\mathcal{F}(\boldsymbol{X}, \varphi)$, for any persistent homology that begins to appear at level b and disappears at level d, we say that the homology is *born* at b and *dies* at d. With Ω defined as in (1), the persistence diagram of the point cloud $\boldsymbol{X}$ is a multiset on Ω that summarizes the birth and death times of all persistent homologies in the filtration $\mathcal{F}(\boldsymbol{X}, \varphi)$:

$$\mathsf{Dgm}(\boldsymbol{X}, \varphi) = \{(b_i, d_i) : \text{the } i\text{-th persistent homology in } \mathcal{F}(\boldsymbol{X}, \varphi)$$
$$\text{that is born at } b_i \text{ and dies at } d_i\}.$$

C SUPPORTIVE THEORETICAL RESULTS

C.1 Validation of Assumption 3.3

In this part, we provide some common data-generating mechanisms where Assumption 3.3 can be validated.

Theorem C.1. *Let q, d be two positive integers with $q > d$. Let κ be a density on $[0,1]^d$ such that $0 < \inf \kappa \le \sup \kappa < \infty$. Suppose that $\boldsymbol{X}_N$ be either a binomial process with parameters N and κ or a Poisson process of intensity $N\kappa$ in the cube $[0,1]^d$. Denote $p(\boldsymbol{u})$ as the intensity function for the k-dimensional expected persistent measure induced by the Vietoris-Rips filtration. Then when N is sufficiently large, for $\boldsymbol{u} \in \Omega$, there exists a polynomial function $\mathsf{poly}(\cdot)$, such that*

$$p(\boldsymbol{u}) \le \mathsf{poly}(N, d) \sup \kappa, \tag{10}$$

and $\bar{p}(\boldsymbol{u})$ can be correspondingly bounded.

Theorem C.2. *Let q, d be two positive integers with $q > d$. Let κ be a density on $[0,1]^{d \times N}$ such that $0 < \inf \kappa < \sup \kappa < \infty$. Suppose that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N \in [0,1]^d$ and that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N) \sim \kappa$. Denote $\tilde{p}(\boldsymbol{u})$ as the persistence density induced by the Vietoris-Rips filtration of $\boldsymbol{X}$. Then there exists a polynomial function $(\cdot)$, such that*

$$\tilde{p}(\boldsymbol{u}) \le \mathsf{poly}(N, d) \sup \kappa. \tag{11}$$

Remark C.3. The bound in (10) seems to mismatch with Assumption 3.3, since the assumption is on $\bar{p}(\boldsymbol{u}) = \|\boldsymbol{u} - \partial\Omega\|_2^q p(\boldsymbol{u})$ while (10) provides the polynomial bound on $p(\boldsymbol{u})$ directly. So one can imagine that there is no benefit on considering the assumption on $\bar{p}$ instead of p. However, though not in the formal proof, we believe that $\bar{p}(\boldsymbol{u})$ would have a polynomial bound with a slower growth order with respect to the sample size N. This is since as the sample size N grows, the corresponding persistence diagram tends to have more points that are close to the diagonal line, as observed in (Divol and Polonik, 2019). Hence the term $\|\boldsymbol{u} - \partial\Omega\|_2^q$ can suppress that effect and $\bar{p}$ can be bounded by a function with a slower growh order with respect to N.

Remark C.4. At first glance, comparing (10) and (11) seems to indicate that the persistence intensity function p and the persistence density function $\tilde{p}$ have more or less similar asymptotic properties with respect to the sample size N. However, this is mainly due to that the polynomial bound $\mathsf{poly}(N, d)$ is comprehensive; for example, N and N^{10} are both in $\mathsf{poly}(N, d)$, hence the bound $\mathsf{poly}(N, d)$ only guarantees that the function does not blow up too fast such as an exponential function. And we believe that, in fact, the growth order is different for p and $\tilde{p}$. This is mainly due to that when the sample size N is large, the persistence diagram induced by $\boldsymbol{X}$ tends to have more points $N(D)$ in the persistence diagram, and hence the normalized measure $\tilde{\mu}(A) = \frac{1}{N(D)} \sum_{i=1}^{N(D)} \delta_{\boldsymbol{r}_i}(A)$ and the corresponding persistence density function $\tilde{p}$ benefits from the term $\frac{1}{N(D)}$ lowering the growth order with respect to N. In fact, in the proofs of Theorem C.1 and C.2 in Section E.8, the bounds $\mathsf{poly}(N, d)$ are $N^5 d^3$ for (10) and $N^4 d^3$ for (11), although the bounds need not necessarily equal to the actual asymptotic orders of p and $\tilde{p}$.

C.2 Clarification of Assumptions

In this part, we provide the details in the smoothness assumption of the persistence intensity and density functions, and the regularization assumptions of the kernel function.

Hölder smoothness. Recall from Assumption 3.5 that we assume the persistence intensity function $p(\cdot)$ and the persistence density function $\tilde{p}(\cdot)$ are Hölder smooth. A function $f : \Omega \to \mathbb{R}_{\ge 0}$ is Hölder smooth with parameter L_f of oreder $s > 0$ if it is $\lfloor s \rfloor$-times continuously differentiable and that for any $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$,

$$\left| f(\boldsymbol{x}') - f(\boldsymbol{x}) - \sum_{t=1}^{\lfloor s \rfloor} \frac{1}{t!} \sum_{t_1 + t_2 = t, t_1, t_2 \ge 0} \frac{\mathrm{d}^t f(\boldsymbol{x})}{\mathrm{d} x_1^{t_1} \mathrm{d} x_2^{t_2}} (x_1' - x_1)^{t_1} (x_2' - x_2)^{t_2} \right| \le L_f \|\boldsymbol{x}' - \boldsymbol{x}\|_2^s. \tag{12}$$

Assumptions regarding the kernel function. Throughout the paper, we assume the kernel function $K(\cdot)$ satisfies some properties that are commonly used in non-parametric statistics Giné and Nickl (2021). Specifically, we make the following assumption.

Assumption C.5. *The kernel function $K : \mathbb{R}^2 \to \mathbb{R}$ satisfies the following conditions:*

(a) $K(\boldsymbol{x}) = 0$ *for all $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 > 1$;*

(b) $\|K\|_\infty \coloneqq \sup_{\boldsymbol{x}} |K(\boldsymbol{x})| < \infty$;

(c) $\int_{\mathbb{R}^2} K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$;

(d) $\|K\|_2^2 \coloneqq \int_{\mathbb{R}^2} K^2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} < \infty$.

(e) *There exists a positive integer s, such that for all non-negative integers s_1, s_2 satisfying $1 \le s_1 + s_2 < s$,*

$$\int_{\boldsymbol{x} \in \mathbb{R}^2} x_1^{s_1} x_2^{s_2} K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0.$$

(f) *K is L_K-Lipchitz with respect to the ℓ_2 norm on $\mathbb{R}^2$.*

C.3   Upper bound for the bias of kernel-based estimators

In this section, we specify the upper bounds on the bias of our kernel-based estimators for the persistence intensity/density function, the linear functionals of the persistent measure, and the persistence betti number in Theorems C.6, C.7 and Corollary C.8 respectively.

Theorem C.6. *Under Assumption 3.5, there exist constants $L_{\tilde{p}}, L_p > 0$ such that, for any $\boldsymbol{\omega} \in \Omega$,*

$$|\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega})| \le L_p h^s \int_{\|\boldsymbol{v}\|_2 \le 1} |K(\boldsymbol{v})| \|\boldsymbol{v}\|_2^s \mathrm{d}\boldsymbol{v}, \quad and$$

$$|\mathbb{E}[\check{p}_h(\boldsymbol{\omega})] - \tilde{p}(\boldsymbol{\omega})| \le L_{\tilde{p}} h^s \int_{\|\boldsymbol{v}\|_2 \le 1} |K(\boldsymbol{v})| \|\boldsymbol{v}\|_2^s \mathrm{d}\boldsymbol{v}.$$

The following theorems provide uniform bounds on the bias and variation of these kernel-based estimators.

Theorem C.7. *Under Assumption 3.5, there exist constants $L_{\tilde{p}}, L_p > 0$ such that*

$$\sup_{\Psi \in \mathscr{F}_{2h,R}} \left| \mathbb{E}[\hat{\Psi}_h] - \Psi \right| \le L_p \left(\frac{L}{\sqrt{2}} \right)^q h^s \int_{\|\boldsymbol{v}\|_2 \le 1} |K(\boldsymbol{v})| \|\boldsymbol{v}\|_2^2 \mathrm{d}\boldsymbol{v}; \quad and$$

$$\sup_{\Psi \in \widetilde{\mathscr{F}}_R} \left| \mathbb{E}[\check{\Psi}_h] - \widetilde{\Psi} \right| \le L_{\tilde{p}} h^s R \int_{\|\boldsymbol{v}\|_2 \le 1} |K(\boldsymbol{v})| \|\boldsymbol{v}\|_2^2 \mathrm{d}\boldsymbol{v}.$$

It is also straightforward to see that "the biases of both $\hat{\beta}_{\boldsymbol{x},h}$ and $\check{\beta}_{\boldsymbol{x},h}$ are of order h^s", uniformly in $\boldsymbol{x}$; see Corollary C.8 below.

Corollary C.8. *Under Assumption 3.5, it holds that*

$$\sup_{\boldsymbol{x} \in \Omega} \left| \mathbb{E}[\hat{\beta}_{\boldsymbol{x},h}] - \beta_{\boldsymbol{x}} \right| \le L_p h^s \frac{L^2}{4} \int_{\|\boldsymbol{v}\|_2 \le 1} K(\boldsymbol{v}) \|\boldsymbol{v}\|_2^2 \mathrm{d}\boldsymbol{v}, \quad and$$

$$\sup_{\boldsymbol{x} \in \Omega} \left| \mathbb{E}[\check{\beta}_{\boldsymbol{x},h}] - \tilde{\beta}_{\boldsymbol{x}} \right| \le L_{\tilde{p}} h^s \frac{L^2}{4} \int_{\|\boldsymbol{v}\|_2 \le 1} K(\boldsymbol{v}) \|\boldsymbol{v}\|_2^2 \mathrm{d}\boldsymbol{v}$$

C.4   Minimax lower bound for estimating the persistence intensity function

Below we provide a minimax lower bound on the L_∞ estimation error of the persistence intensity function by levering well-known minimax arguments for estimating a smooth probability density function based on an i.i.d. sample; see Giné and Nickl (2021) for details, as well for the definition of Besov norms.

Theorem C.9. *Let $\mathscr{F}$ denote the set of functions on Ω with Besov norm bounded by $B > 0$:*

$$\mathscr{F} = \{ f : \Omega \to \mathbb{R}, \|f\|_{B^s_{\infty,\infty}} \le B \}.$$

Then,

$$\inf_{\hat{p}_n} \sup_P \mathbb{E}_{\mu_1, \dots, \mu_n \overset{i.i.d.}{\sim} P} \sup_{\boldsymbol{\omega} \in \Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q |\hat{p}_n(\boldsymbol{\omega}) - p(\boldsymbol{\omega})| \ge O(n^{-\frac{s}{2(s+1)}}),$$

where the infimum is taken over estimator $\hat{p}_n$ mapping $\mu_1, \dots, \mu_n$ to an intensity function in $\mathscr{F}$, the supremum is over the set of all probability distribution on $\mathcal{Z}_{L,M}^q$ and p is the intensity function of $\mathbb{E}_P[\mu]$.

C.5 Estimating the persistence surface

For estimating the persistence surface in (6), we directly generate the persistence surface from the empirical averaged persistence measure $\bar{\mu}_n$ given by

$$A \in \mathcal{B} \mapsto \bar{\mu}_n(A) = \frac{1}{n} \sum_{i=1}^n \mu_i(A).$$

Since $\bar{\mu}_n$ is unbiased for $\mathbb{E}[\mu]$ and ρ is a linear transformation, $\rho_h(\bar{\mu}_n)$ is also unbiased for $\rho_h(\mathbb{E}[\mu])$. The following theorem bounds its variation.

Theorem C.10. *With the choice of the weight function*

$$f(\boldsymbol{\omega}) = \|\boldsymbol{\omega} - \partial\Omega\|_2^q,$$

when Assumptions 3.3(a) and 3.2 hold true, there exists a constant C depending on $L, M, L_K, \|K\|_\infty$ and $\|\bar{p}\|_\infty$, such that for any $\delta \in (0,1)$, it can be guaranteed with probability at least $1-\delta$ that

$$\|\rho_h(\bar{\mu}_n) - \rho_h(\mathbb{E}[\mu])\|_\infty \le C \max\left\{ \frac{1}{nh^2} \log\frac{1}{\delta h^2}, \sqrt{\frac{1}{nh^2}} \sqrt{\log\frac{1}{\delta h^2}} \right\}.$$

C.6 Estimating the persistent betti number by the empirical averaged persistence measure

As an alternative to the kernel-based estimator for the persistent betti number in (9), we can directly use the empirical persistent betti number as the estimator:

$$\bar{\beta}_{\boldsymbol{x}} = \bar{\mu}_n(B_{\boldsymbol{x}}).$$

Since $\bar{\mu}_n$ is an unbiased estimator for $\mathbb{E}[\mu]$, $\bar{\beta}_{\boldsymbol{x}}$ is an unbiased estimator for $\beta_{\boldsymbol{x}}$. As for the variation of the estimator, we provide the following theorem.

Theorem C.11. *Under Assumptions 3.5, 3.3(a) and 3.2, for any $\delta \in (0,1)$, there exists a universal constant C such that with probability at least $1-\delta$, it can be guaranteed that*

$$\sup_{\boldsymbol{x}\in\Omega_\ell} |\bar{\beta}_{\boldsymbol{x}} - \beta_{\boldsymbol{x}}| \le C\left(\frac{M\ell^{-q}}{n}\left(2\log(M\ell^{-q}n+1) + \log\frac{1}{\delta} \right) \right.$$

$$\left. + \sqrt{\min\left\{ \frac{M^2\ell^{-2q}}{n}, \frac{\sqrt{2}ML\ell^{1-2q}\|\bar{p}\|_\infty}{(q-1)_+ n} \right\} \left(\sqrt{2\log(M\ell^{-q}n+1)} + \sqrt{\log\frac{1}{\delta}} \right)} \right),$$

where $(q-1)_+ = \max\{q-1, 0\}$.

D PRELIMINARY FACTS

In this section we present and prove various auxiliary results that are needed in the proofs of the main theorems.

D.1 Preliminary facts for the proof of Theorem C.1

Bounding the weighted intensity function as in Theorem C.1 requires a detailed exploration of the persistent diagram for the Vietoris-Rips filtration. Throughout this section, we will consider the filtering function corresponding to the Vietoris-Rips filtration

$$\varphi[J](\boldsymbol{X}) = \min_{i,j\in J, i\neq j} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2.$$

Firstly, we state a form of the **area formula** given by (Morgan, 2016), which would be useful for a change of variable in deriving the intensity function for the expected persistence measure.

Theorem D.1. *Denote $\mathscr{L}^M$ as the M-dimensional Lebesgue measure and $\mathscr{H}^M$ as the M-dimensional Hausdorff measure. Consider a Lipchitz function $f : \mathbb{R}^M \to \mathbb{R}^N$ for $M \leq N$. If $h : \mathbb{R}^M \to \mathbb{R}$ is an $\mathscr{L}^M$-integrable function, then*

$$\int_{\mathbb{R}^M} h(\boldsymbol{X}) J_{\boldsymbol{X}} f(\boldsymbol{X}) \mathrm{d}\mathscr{L}^M(\boldsymbol{X}) = \int_{\mathbb{R}^N} \sum_{\boldsymbol{X} \in f^{-1}\{\boldsymbol{Y}\}} h(\boldsymbol{X}) \mathrm{d}\mathscr{H}^M \boldsymbol{Y},$$

where $J_{\boldsymbol{X}} f(\boldsymbol{X})$ is the Jacobian determinant of the function f:

$$J_{\boldsymbol{X}} f(\boldsymbol{X}) = \sqrt{\det\left(\left(\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{X}}\right)^{\top}\left(\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{X}}\right)\right)}.$$

Theorem D.1 directly implies the following corollary, the proof of which would be omitted.

Corollary D.2. *Let $\psi : \mathbb{R}^M \to \mathbb{R}^N$ be a Lipchitz bijection with $M \leq N$, and $\kappa : \mathbb{R}^N \to \mathbb{R}$ be a function which satisfies that $h := \kappa \circ \psi$ is $\mathscr{L}^M$-integrable. Then*

$$\int_{\mathbb{R}^M} \kappa \circ \psi(\boldsymbol{X}) J_{\boldsymbol{X}} \psi(\boldsymbol{X}) \mathrm{d}\mathscr{L}^M(\boldsymbol{X}) = \int_{\mathbb{R}^N} \kappa(\boldsymbol{Y}) \mathrm{d}\mathscr{H}^M(\boldsymbol{Y}).$$

The following proposition considers two kinds of partitions of the unit cube $[0,1]^{d \times N}$, with each part satisfying some desired properties.

Proposition D.3. *There exists a set S with cardinality $\mathrm{card}(S) = 4d^2$, such that for any $J_1, J_2 \subset [N]$ that satisfies $J_1 \neq J_2, |J_1| = |J_2| = 2$, bearing a zero-measured set, $[0,1]^{d \times n}$ can be partitioned as*

$$[0,1]^{d \times n} = \bigcup_{s \in S} W^s_{J_1, J_2},$$

such that within each part $W^s_{J_1, J_2}$, there exists a diffeomorphism $\Psi^s_{J_1, J_2} : W^s_{J_1, J_2} \to \mathbb{R}^2 \times [0,1]^{nd-2}$, such that:

1. *For every $\boldsymbol{X} \in W^s_{J_1, J_2}$, $\Psi^s_{J_1, J_2}(\boldsymbol{X})_1 = \varphi[J_1](\boldsymbol{X})$ and $\Psi^s_{J_1, J_2}(\boldsymbol{X})_2 = \varphi[J_2](\boldsymbol{X})$;*

2. *The Jacobian determinant $J_{\boldsymbol{X}} \Psi^s_{J_1, J_2}(\boldsymbol{X}) \geq \frac{1}{d}$.*

Proof: Let $S = [d]^2 \times \{-1, +1\}^2$, then it is easy to see that $|S| = 4d^2$. For any $J_1, J_2 \subset [n]$ with $J_1 \neq J_2$ and $|J_1| = |J_2| = 2$, let denote $J_1 = \{i_1, j_1\}$, $J_2 = \{i_2, j_2\}$ with $j_2 = \max\{j \in J_2 : j \notin J_1\}$. For any $s = (k_1, k_2, s_1, s_2) \in S$, let

$$\begin{aligned} W^s_{J_1, J_2} = \{X : &\{k_1\} = \mathrm{argmax}_k |X^k_{i_1} - X^k_{j_1}|, s_1(X^k_{j_1} - X^k_{i_1}) > 0, \\ &\{k_2\} = \mathrm{argmax}_k |X^k_{i_2} - X^k_{j_2}|, s_2(X^k_{j_2} - X^k_{i_2}) > 0.\} \end{aligned}$$

Notice here that $\{k_1\} = \mathrm{argmax}_k |X^k_{i_1} - X^k_{j_1}|$ means k_1 is the *only* index for $|X^k_{i_1} - X^k_{j_1}|$ to reach its maximum.

We begin by proving that $\{W^s_{J_1, J_2}\}_{s \in S}$ forms a partition of $[0,1]^{d \times n}$ bearing a zero-measured set. Firstly, for $s, s' \in S$ with $s \neq s'$, it is easy to see that $W^s_{J_1, J_2}$ and $W^{s'}_{J_1, J_2}$ are disjoint. Secondly, if

$$\boldsymbol{X} \in [0,1]^{d \times n} - \bigcup_{s \in S} W^s_{J_1, J_2},$$

then by definition, there exists $k, k' \in [d]$, such that $k \neq k'$ and that either

$$|X^k_{j_1} - X^k_{i_1}| = |X^{k'}_{j_1} - X^{k'}_{i_1}|$$

or

$$|X^k_{j_2} - X^k_{i_2}| = |X^{k'}_{j_2} - X^{k'}_{i_2}|.$$

Notice that for any $k, k' \in [d]$ with $k \neq k'$, the set

$$\left\{ \boldsymbol{X} : |X_{j_1}^k - X_{i_1}^k| = |X_{j_1}^{k'} - X_{i_1}^{k'}| \right\}$$
$$= \left\{ \boldsymbol{X} : X_{j_1}^k - X_{i_1}^k = |X_{j_1}^{k'} - X_{i_1}^{k'}| \right\} \cup \left\{ \boldsymbol{X} : X_{j_1}^k - X_{i_1}^k = -|X_{j_1}^{k'} - X_{i_1}^{k'}| \right\},$$

where the sets

$$\left\{ \boldsymbol{X} \in [0,1]^{d \times n} : X_{j_1}^k - X_{i_1}^k = |X_{j_1}^{k'} - X_{i_1}^{k'}| \right\} \quad \text{and}$$
$$\left\{ \boldsymbol{X} \in [0,1]^{d \times n} : X_{j_1}^k - X_{i_1}^k = -|X_{j_1}^{k'} - X_{i_1}^{k'}| \right\}$$

are a subsets of $(nd-1)$ dimensional linear manifolds in $[0,1]^{d \times n}$, and are therefore zero-measured in $\mathscr{L}^{nd}$. Similarly, we can prove that the set $[0,1]^{d \times n} - \bigcup_{s \in S} W_{J_1, J_2}^s$ is the union of a finite number of subsets of $(nd-1)$ dimensional linear manifolds in $[0,1]^{d \times n}$. Consequently,

$$\bigcup_{s \in S} W_{J_1, J_2}^s$$

is a partition of $[0,1]^{d \times n}$ bearing a zero-measured set.

Furthermore, define Ψ_{J_1, J_2}^s as

$$\Psi_{J_1, J_2}^s(\boldsymbol{X}) = \left(\varphi[J_1](\boldsymbol{X}), \varphi[J_2](\boldsymbol{X}), \{X_j^k\}_{\substack{1 \leq j \leq n \\ 1 \leq k \leq d \\ (j,k) \neq (j_1, k_1) \\ (j,k) \neq (j_2, k_2)}} \right), \quad \forall \boldsymbol{X} \in W_{J_1, J_2}^s.$$

Then we can firstly notice that

$$X_{j_1}^{k_1} = s_1 \sqrt{u_1^2 - \sum_{k \neq k_1} \left(X_{j_1}^k \right)^2} + X_{i_1}^{k_1} \quad \text{and}$$
$$X_{j_2}^{k_2} = s_2 \sqrt{u_2^2 - \sum_{k \neq k_2} \left(X_{j_2}^k \right)^2} + X_{i_2}^{k_2},$$

for $u_1 = \varphi[J_1](X)$ and $u_2 = \varphi[J_2](X)$. This validates Ψ_{J_1, J_2}^s as a diffeomorphism. The proof now boils down to bounding the Jacobian of Ψ_{J_1, J_2}^s. Towards this end, notice that the partial derivative of φ is bounded by

$$\left| \frac{\partial \varphi[J_1](\boldsymbol{X})}{\partial X_{j_1}^{k_1}} \right| = \left| \frac{\partial}{\partial X_{j_1}^{k_1}} \sqrt{\sum_{k=1}^d (X_{i_1}^k - X_{j_1}^k)^2} \right|$$
$$= \left| \frac{X_{j_1}^{k_1} - X_{i_1}^{k_1}}{\sqrt{\sum_{k=1}^d (X_{i_1}^k - X_{j_1}^k)^2}} \right|$$
$$\geq \frac{1}{\sqrt{d}},$$

where in the last line we applied the fact that

$$\left| X_{j_1}^{k_1} - X_{i_1}^{k_1} \right| = \max_{1 \leq k \leq d} |X_{j_1}^k - X_{i_1}^k| \geq \sqrt{\frac{1}{d} \sum_{k=1}^d (X_{i_1}^k - X_{j_1}^k)^2}.$$

Similarly,

$$\left| \frac{\partial \varphi[J_2](\boldsymbol{X})}{\partial X_{j_2}^{k_2}} \right| = \left| \frac{\partial}{\partial X_{j_2}^{k_2}} \sqrt{\sum_{k=1}^d (X_{i_2}^k - X_{j_2}^k)^2} \right| \geq \frac{1}{\sqrt{d}}.$$

Furthermore, since $j_2 \notin J_1$, it is easy to see that

$$\frac{\partial \varphi[J_1](\boldsymbol{X})}{\partial X_{j_2}^{k_2}} = 0.$$

Therefore, the Jacobian determinant of Ψ_{J_1,J_2}^s is bounded by

$$
\begin{aligned}
J_{\boldsymbol{X}} \Psi_{J_1,J_2}^s(\boldsymbol{X}) &= \left| \det \left(\frac{\mathrm{d}\Psi_{J_1,J_2}^s(\boldsymbol{X})}{\mathrm{d}\boldsymbol{X}} \right) \right| \\
&= \left| \det \left(\begin{pmatrix} \mathbf{I}_{nd-2} & \mathbf{0}_{(nd-2)\times 1} & \mathbf{0}_{(nd-2)\times 1} \\ \mathbf{0}_{1\times(nd-2)} & \frac{\partial \varphi[J_1](\boldsymbol{X})}{\partial X_{j_1}^{k_1}} & \frac{\partial \varphi[J_1](\boldsymbol{X})}{\partial X_{j_2}^{k_2}} \\ \mathbf{0}_{1\times(nd-2)} & \frac{\partial \varphi[J_2](\boldsymbol{X})}{\partial X_{j_1}^{k_1}} & \frac{\partial \varphi[J_2](\boldsymbol{X})}{\partial X_{j_2}^{k_2}} \end{pmatrix} \right) \right| \\
&= \left| \frac{\partial \varphi[J_1](\boldsymbol{X})}{\partial X_{j_1}^{k_1}} \cdot \frac{\partial \varphi[J_2](\boldsymbol{X})}{\partial X_{j_2}^{k_2}} \right| \geq \frac{1}{d}.
\end{aligned}
$$

This completes the proof. ∎

The following is important for representing of the persistence intensity function p and the persistence density function $\tilde{p}$.

Proposition D.4. *Bearing a zero-measured set, $[0,1]^{d\times n}$ can be partitioned as*

$$[0,1]^{d\times n} = \bigcup_{r=1}^{R} V_r,$$

such that

1. *For every $\boldsymbol{X}, \boldsymbol{X}' \in V_r$, $J_1, J_2 \subset [n]$ with $|J_1| = |J_2| = 2$, it is guaranteed that $\varphi[J_1](\boldsymbol{X}) \neq \varphi[J_2](\boldsymbol{X})$; furthermore, if $\varphi[J_1](\boldsymbol{X}) < \varphi[J_2](\boldsymbol{X})$, then $\varphi[J_1](\boldsymbol{X}') < \varphi[J_2](\boldsymbol{X}')$;*

2. *For every $\boldsymbol{X}, \boldsymbol{X} \in V_r$, $J_1, J_2, J_3, J_4 \subset [n]$ with $|J_1| = |J_2| = |J_3| = |J_4| = 2$, it is guaranteed that $\varphi[J_1](\boldsymbol{X}) - \varphi[J_2](\boldsymbol{X}) \neq \varphi[J_3](\boldsymbol{X}) - \varphi[J_4](\boldsymbol{X})$; furthermore, if $\varphi[J_1](\boldsymbol{X}) - \varphi[J_2](\boldsymbol{X}) > \varphi[J_3](\boldsymbol{X}) - \varphi[J_4](\boldsymbol{X}) > 0$, then $\varphi[J_1](\boldsymbol{X}') - \varphi[J_2](\boldsymbol{X}') > \varphi[J_3](\boldsymbol{X}') - \varphi[J_4](\boldsymbol{X}) > 0$.*

3. *For every $r \in [R]$ and $\boldsymbol{X} \in V_r$, there are N_r points in $\mathrm{Dgm}(\boldsymbol{X}, \varphi)$; furthermore, all these points can be ordered by their orthogonal distance to the diagonal, and the order is fixed for all $\boldsymbol{X} \in V_r$.*

Furthermore, the expected persistence measure $\mathbb{E}[\mu]$ and its normalized counterpart $\mathbb{E}[\tilde{\mu}]$ can be characterized such that for any Borel set $B \subset \Omega$,

$$\mathbb{E}[\mu](B) = \sum_{r=1}^{R} \sum_{i=1}^{N_r} \int_{x \in \Phi^{-1}[J_{ir}^1, J_{ir}^2](B) \cap V_r} \kappa(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} \quad and$$

$$\mathbb{E}[\tilde{\mu}](B) = \sum_{r=1}^{R} \frac{1}{N_r} \sum_{i=1}^{N_r} \int_{x \in \Phi^{-1}[J_{ir}^1, J_{ir}^2](B) \cap V_r} \kappa(\boldsymbol{X}) \mathrm{d}\boldsymbol{X}$$

, in which

$$\Phi[J_1, J_2](\boldsymbol{X}) = (\varphi[J_1](\boldsymbol{X}), \varphi[J_2](\boldsymbol{X})),$$

and J_{ir}^1, J_{ir}^2 are the simplicial complexes corresponding to the birth and death of the i-th persistence homology for all $\boldsymbol{X} \in V_r$.

Proof: For simplicity, we only give a sketch of the proof for this proposition. A weaker version of this proposition is proved in (Chazal and Divol, 2019), where the second property of the partition is not required. Therefore, the

partition we aim to construct here is a refinement of the partition given in (Chazal and Divol, 2019). In order to see that the second condition can be reached, we firstly prove that the set

$$A = \big\{ \boldsymbol{X} \in [0,1]^{d \times n} : \exists J_1, J_2, J_3, J_4 \subset [n], \text{ s.t.}$$
$$|J_1| = |J_2| = |J_3| = |J_4| = 2,$$
$$J_1 \neq J_2, J_3 \neq J_4, (J_1, J_2) \neq (J_3, J_4),$$
$$\varphi[J_1](\boldsymbol{X}) - \varphi[J_2](\boldsymbol{X}) = \varphi[J_3](\boldsymbol{X}) - \varphi[J_4](\boldsymbol{X}) \big\}$$

is zero-measured. For this step, the technique in proving Lemma 4.1 in (Chazal and Divol, 2019) can be applied to prove that A does not contain any open set, and all its points are singular.

We can further define

$$\mathcal{F}_n^2 = \{ (J_1, J_2) : J_1, J_2 \subset [n], |J_1| = |J_2| = 2, J_1 \neq J_2 \}.$$

Since A is zero-measured, we can only consider the set $[0,1]^{d \times n} - A$, on which

$$\{ \Delta \varphi[J_1, J_2](\boldsymbol{X}) := \varphi[J_1](\boldsymbol{X}) - \varphi[J_2](\boldsymbol{X}) \}_{(J_1, J_2) \in \mathcal{F}_n^2}$$

must take different values for different $(J_1, J_2) \in \mathcal{F}_n^2$. Denote these values as $r_1 < r_2 < ... < r_L$, and let $E_\ell(\boldsymbol{X})$ denote the element $(J_1, J_2) \subset \mathcal{F}_n^2$ such that $\Delta \varphi[J_1, J_2](\boldsymbol{X}) = r_\ell$. The sets $E_1(\boldsymbol{X}), E_2(\boldsymbol{X}), ..., E_L(\boldsymbol{X})$ then form a partition of $\mathcal{F}_n^2$. With similar techniques as Lemma 4.2 in (Chazal and Divol, 2019), we can prove that the map $\boldsymbol{X} \mapsto \mathcal{A}^2(\boldsymbol{X})$ is locally constant almost surely everywhere. This essentially completes the proof.

∎

The following lemma is a direct application of Proposition 4.6 in Divol and Polonik (2019), and guarantees that the number of points in the persistence diagram $\mathsf{Dgm}(\boldsymbol{X}, \varphi)$ that are far enough from the diagonal is upper bounded in terms of the expectation.

Lemma D.5. *Let κ be a probability density function on $[0,1]^d$ that satisfies $0 < \inf \kappa < \sup \kappa < \infty$. Denote $\mathbb{X}_n$ as a binomial process with parameters n and κ or a Poisson process with parameter $n\kappa$ on $[0,1]^d$. In the kth dimensional persistence diagram of the Vietoris-Rips filtration of $\mathbb{X}_n$, let N_ℓ be the number of points with persistence of at least ℓ. Then there are some universal constant C that the expectation of N_ℓ is upper bounded as*

$$\mathbb{E}[N_\ell] \leq Cn \exp\left(-Cn\ell^d\right),$$

where C is a constant depends only on k.

Proof: Let μ be the persistence measure corresponding to the k-th dimensional persistence diagram of the Vietoris-Rips filtration of $\mathbb{X}_n$. From Proposition 4.6 in Divol and Polonik (2019),

$$P\left(\mu(\mathbb{R} \times [\ell, \infty)) > t\right) \leq c_1 \exp\left(-c_2 \left(n\ell^d + (\frac{t}{n})^{1/(k+1)}\right)\right).$$

And hence the expectation of $\mu(\mathbb{R} \times [\ell, \infty))$ is bounded as

$$\begin{aligned}
\mathbb{E}[\mu(\mathbb{R} \times [\ell, \infty))] &\leq \int_0^\infty c_1 \exp\left(-c_2 \left(n\ell^d + (\frac{t}{n})^{1/(k+1)}\right)\right) dt \\
&= c_1 \exp\left(-c_2(n\ell^d)\right) \int_0^\infty \exp\left(-c_2(\frac{t}{n})^{1/(k+1)}\right) dt \\
&= c_1 \exp\left(-c_2(n\ell^d)\right) \int_0^\infty (k+1)nu^k \exp\left(-c_2 u\right) du \\
&= Cn \exp\left(-Cn\ell^d\right),
\end{aligned}$$

for some constant C that depends on k. Now, $\mathbb{R} \times [\ell, \infty)$ contains all the homological features whose persistence is at least ℓ, so

$$N_\ell \leq \mu(\mathbb{R} \times [\ell, \infty)).$$

And hence

$$\mathbb{E}[N_\ell] \leq Cn \exp\left(-Cn\ell^d\right).$$

∎

D.2 Uniform tail bounds

In this section, we provide some uniform tail bound theorems that are important for bounding the variation of estimators. We will omit the proofs of these theorems in the paper.

The Talagrand's inequality. The following form of the Talagrand's inequality was shown in (Steinwart and Christmann, 2008).

Theorem D.6. *Let $(\mathcal{Z}, \mathscr{F}, P)$ be a probability space and (T, d) be a separable metric space. Consider a function class $\mathcal{G} = \{g_t : t \in T\} \in L_0(\mathcal{Z})$, such that the function $t \mapsto g_t(z)$ is continuous in t for all $z \in \mathcal{Z}$. Furthermore, suppose that there exists a constant $B > 0, \sigma^2 > 0$ such that for all $g \in \mathcal{G}$, $\mathbb{E}[g] = 0, \mathbb{E}[g^2] \leq \sigma^2, ||g||_\infty \leq B$. Let $Z_1, Z_2, ..., Z_n \sim$ i.i.d. P, and define*

$$G = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) \right|.$$

Then for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$,

$$G \leq 4\mathbb{E}[G] + \sqrt{\frac{2\sigma^2}{n} \log \frac{1}{\delta}} + \frac{B}{n} \log \frac{1}{\delta}. \tag{13}$$

Theorem D.6 implies that the expectation of G is an important factor in bounding G. The following theorem gives and upper bound of $\mathbb{E}[G]$ by the covering number of $\mathcal{G}$.

Theorem D.7. *Under the same conditions as in Theorem D.6, if for any $\eta \in (0, B)$, there exists $A > 0$, $\nu > 0$ such that for any probability measure Q on $\mathcal{Z}$, the covering number is bounded as*

$$\mathcal{N}(\mathcal{G}, L_2(Q), \eta) \leq \left(\frac{AB}{\eta} \right)^\nu,$$

then there exists a constant C such that

$$\mathbb{E}[G] \leq C \left(\frac{\nu B}{n} \log \left(\frac{AB}{\sigma} \right) + \sqrt{\frac{\nu \sigma^2}{n} \log \left(\frac{AB}{\sigma} \right)} \right).$$

Tail bound by polynomial discrimination. As an alternative to the Talagrand's inequality, the following theorem bounds G with high probability when the function class $\mathcal{G}$ has *polynomial discrimination*. The proof applies the Bernstein's inequality and a straightforward union bound argument.

Theorem D.8. *Under the same conditions as in Theorem D.6, define*

$$\mathcal{G}(\mathbf{Z}_1^n) = \{(g(Z_1), g(Z_2), ..., g(Z_n)) : g \in \mathcal{G}\}. \tag{14}$$

If the cardinality of the set $\mathcal{G}(\mathbf{Z}_1^n)$ is bounded by

$$Card(\mathcal{G}(\mathbf{Z}_1^n)) \leq (An + 1)^\nu \tag{15}$$

for some $\nu > 0$, then there exists a universal constant C such that with probability at least $1 - \delta$,

$$G \leq C \left(\sqrt{\frac{\sigma^2}{n}} \left(\sqrt{\nu \log(An + 1)} + \sqrt{\log \frac{1}{\delta}} \right) + \frac{B}{n} \left(\nu \log(An + 1) + \log \frac{1}{\delta} \right) \right) \tag{16}$$

The following lemma shows that for persistent measures with bounded total persistence, the total mass of the set away from the diagonal $\partial \Omega$ is upper bounded.

Lemma D.9. *Let Ω_ℓ denote the set of points in Ω that are at least ℓ away from the diagonal:*

$$\Omega_\ell = \{\boldsymbol{\omega} \in \Omega : \|\boldsymbol{\omega} - \partial \Omega\|_2 \geq \ell\}.$$

Then for a persistent measure μ, if $Pers_q(\mu) \leq M$, then $\mu(\Omega_\ell) \leq M\ell^{-q}$.

The following theorem shown in Divol and Lacombe (2021) provides a standard lower bound for the minimax rate of estimating a probability density function using independent samples. This is useful for deducting the minimax rate for estimating the (weighted) intensity functions.

Theorem D.10. *Let $\mathscr{F}$ denote the set of probability density functions on $[0,1]^2$ with Bounded Besov norm:*

$$\mathscr{F} = \{f : [0,1]^2 \to \mathbb{R}, \int_{[0,1]^2} f(x)\mathrm{d}x = 1, \|f\|_{\infty,\infty}^r \le B\}.$$

Then for any estimator (measurable function)

$$\hat{f}_n : ([0,1]^2)^n \to \mathscr{F},$$

there exists $f \in \mathscr{F}$, such that if $X_1, X_2, ..., X_n \sim$ i.i.d. f, then

$$\mathbb{E}\|\hat{f}_n(X_1, X_2, ..., X_n) - f\|_\infty \ge O\left(n^{-\frac{r}{2r+2}}\right).$$

E PROOF OF THEOREMS AND SUPPORTIVE PROPOSITIONS

E.1 Proof of Theorem 3.1

In order to prove Theorem 3.1, we firstly show the following supportive lemma.

Lemma E.1. *Let Ω and $\partial\Omega$ be defined as in (1) and (2). Then for any $q > 0$,*

$$\int_\Omega \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x} = \frac{2}{(q+1)(q+2)}\left(\frac{L}{\sqrt{2}}\right)^{q+2}.$$

Proof of Lemma E.1: Take the coordinate transformation

$$\begin{cases} y_1 = \frac{x_2 - x_1}{\sqrt{2}} = \|\boldsymbol{x} - \partial\Omega\|_2; \\ y_2 = \frac{x_2 + x_1}{\sqrt{2}}. \end{cases}$$

Then it can be easily verified that the determinant of the Jacobian matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$ coordinates is 1, and that the ℓ_1 ball Ω can be represented using $\boldsymbol{y}$ coordinates by

$$\Omega = \left\{(y_1, y_2) : 0 < y_1 \le \frac{L}{\sqrt{2}}, y_1 \le y_2 \le \sqrt{2}L - y_1\right\}.$$

Therefore,

$$\begin{aligned}
\int_\Omega \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x} &= \int_0^{\frac{L}{\sqrt{2}}} \left(\int_{y_1}^{\sqrt{2}L - y_1} \mathrm{d}y_2\right) y_1^q \mathrm{d}y_1 \\
&= \int_0^{\frac{L}{\sqrt{2}}} (\sqrt{2}L - 2y_1) y_1^q \mathrm{d}y_1 \\
&= \frac{2}{(q+1)(q+2)}\left(\frac{L}{\sqrt{2}}\right)^{q+2}.
\end{aligned}$$

With this lemma, we can now prove Theorem 3.1.

Proof of Theorem 3.1: The main idea of bounding the OT distance is to construct an admissible transport between μ and ν, and then control the cost of this transport. We will separate the proof into three steps accordingly.

Step 1: Construct an admissible transport from μ to ν. Define $\hat{\pi}$ as a measure on $\overline{\Omega} \times \overline{\Omega}$ such that for any Borel sets $A, B \subset \overline{\Omega}$,

$$
\hat{\pi}(A \times B) = \int_{A \cap B \cap \Omega} \min\{p_\mu(\boldsymbol{x}), p_\nu(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x} +
$$
$$
\int_{A \cap \mathsf{Proj}_{\partial\Omega}^{-1}(B) \cap \Omega} [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \mathrm{d}\boldsymbol{x} + \int_{B \cap \mathsf{Proj}_{\partial\Omega}^{-1}(A) \cap \Omega} [p_\nu(\boldsymbol{x}) - p_\mu(\boldsymbol{x})]^+ \mathrm{d}\boldsymbol{x}. \tag{17}
$$

Here, for any set $A \subset \overline{\Omega}$,

$$
\mathsf{Proj}_{\partial\Omega}^{-1}(A) = \{\boldsymbol{\omega} \in \Omega : \mathsf{Proj}_{\partial\Omega}(\omega) \in A\}.
$$

Intuitively, $\hat{\pi}$ represents such a transport: at each point $\boldsymbol{x} \in \Omega$, if $p_\mu(\boldsymbol{x}) > p_\nu(\boldsymbol{x})$, then we transport the mass of p_ν from $\boldsymbol{x}$ to $\boldsymbol{x}$, and the remaining mass from $\boldsymbol{x}$ to its projection onto $\partial\Omega$; if $p_\nu(\boldsymbol{x}) > p_\mu(\boldsymbol{x})$, then the opposite is done.

Firstly, we prove that this is an admissible transport between μ and ν. Notice that for any Borel set $A \subset \Omega$, $A \cap \overline{\Omega} \cap \Omega = A$, $A \cap \mathsf{Proj}_{\partial\Omega}^{-1}(\overline{\Omega}) \cap \Omega = A$ and $\mathsf{Proj}_{\partial\Omega}^{-1}(A) = \emptyset$. Therefore, by taking $B = \overline{\Omega}$ in (17), we get

$$
\hat{\pi}(A \times \overline{\Omega}) = \int_A \min\{p_\mu(\boldsymbol{x}), p_\nu(\boldsymbol{x})\} \mathrm{d}\boldsymbol{x} + \int_A [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \mathrm{d}\boldsymbol{x} + 0
$$
$$
= \int_A \left\{ \min\{p_\mu(\boldsymbol{x}), p_\nu(\boldsymbol{x})\} + [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \right\} \mathrm{d}\boldsymbol{x}
$$
$$
= \int_A p_\mu(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mu(A).
$$

Similarly, we can prove that $\hat{\pi}(\overline{\Omega} \times B) = \nu(B)$ for any Borel set $B \subset \Omega$. Therefore, $\hat{\pi}$ is an admissible transport between μ and ν.

Step 2: Present $\mathrm{d}\hat{\pi}$. In order to calculate the transport cost of $\hat{\pi}$, we firstly need to present $\mathrm{d}\hat{\pi}$. For this, we would make use of *pushforward measures*. Define $\imath : \overline{\Omega} \to \overline{\Omega} \times \overline{\Omega}$ by $\imath(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{x})$, and let $\jmath : \overline{\Omega} \times \overline{\Omega} \to \overline{\Omega}$ be satisfying $\jmath \circ \imath = id$. Furthermore, let $\imath_*(\lambda_\Omega)$ be the pushforward measure on $\overline{\Omega} \times \overline{\Omega}$ generated by $\imath$. Then for any Borel sets $A, B \subset \overline{\Omega}$, one has $\imath^{-1}(A \times B) = A \cap B$, and the first term in (17) can be presented as

$$
\int_{A \cap B \cap \Omega} \min\{p_\mu(\boldsymbol{x}), p_\nu(\boldsymbol{x})\} \, d\boldsymbol{x}
$$
$$
= \int_{\imath^{-1}(A \times B)} \min\{(p_\mu \circ \jmath)(\imath(\boldsymbol{x})), (p_\nu \circ \jmath)(\imath(\boldsymbol{x}))\} \, d\lambda_\Omega(\boldsymbol{x})
$$
$$
= \int_{A \times B} \min\{(p_\mu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}), (p_\nu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y})\} \, d\imath_*(\lambda_\Omega)(\boldsymbol{x}, \boldsymbol{y}).
$$

For the second term in (17), we can similarly, define $\imath^{(1)} : \overline{\Omega} \to \overline{\Omega} \times \overline{\Omega}$ by $\imath^{(1)}(\boldsymbol{x}) = (\boldsymbol{x}, \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x}))$, let $\jmath^{(1)} : \overline{\Omega} \times \overline{\Omega} \to \overline{\Omega}$ be satisfying $\jmath^{(1)} \circ \imath^{(1)} = id$, and consider the pushforward measure $\imath_*^{(1)}(\lambda_\Omega)$. Then $(\imath^{(1)})^{-1}(A \times B) = A \cap \mathsf{Proj}_{\partial\Omega}^{-1}(B)$, and

$$
\int_{A \cap \mathsf{Proj}_{\partial\Omega}^{-1}(B) \cap \Omega} [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \, d\boldsymbol{x}
$$
$$
= \int_{(\imath^{(1)})^{-1}(A \times B)} \left[(p_\mu \circ \jmath^{(1)})(\imath^{(1)}(\boldsymbol{x})) - (p_\nu \circ \jmath^{(1)})(\imath^{(1)}(\boldsymbol{x})) \right]^+ \, d\lambda_\Omega(\boldsymbol{x})
$$
$$
= \int_{A \times B} \left[(p_\mu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ \, d\imath_*^{(1)}(\lambda_\Omega)(\boldsymbol{x}, \boldsymbol{y}).
$$

For the third term in (17), we can similarly define $\imath^{(2)} : \overline{\Omega} \to \overline{\Omega} \times \overline{\Omega}$ by $\imath^{(2)}(\boldsymbol{x}) = (\mathsf{Proj}_{\partial\Omega}(\boldsymbol{x}), \boldsymbol{x})$, let $\jmath^{(2)} : \overline{\Omega} \times \overline{\Omega} \to \overline{\Omega}$ be satisfying $\jmath^{(2)} \circ \imath^{(2)} = id$, and consider a pushforward measure $\imath_*^{(2)}(\lambda_\Omega)$. Then $(\imath^{(2)})^{-1}(A \times B) = \mathsf{Proj}_{\partial\Omega}^{-1}(A) \cap B$,

and

$$\int_{\mathrm{Proj}_{\partial\Omega}^{-1}(A)\cap B\cap\Omega} [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \, d\boldsymbol{x}$$
$$= \int_{(\imath^{(2)})^{-1}(A\times B)} \left[(p_\mu \circ \jmath^{(2)})(\imath^{(2)}(\boldsymbol{x})) - (p_\nu \circ \jmath^{(2)})(\imath^{(2)}(\boldsymbol{x})) \right]^+ d\lambda_\Omega(\boldsymbol{x})$$
$$= \int_{A\times B} \left[(p_\mu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ d\imath_*^{(1)}(\lambda_\Omega)(\boldsymbol{x}, \boldsymbol{y}).$$

Combining these results, we can obtain the following presentation of d$\hat{\pi}$:

$$d\hat{\pi} = \min\left\{ (p_\mu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}), (p_\nu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}) \right\} d\imath_*(\lambda_\Omega)$$
$$+ \left[(p_\mu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ d\imath_*^{(1)}(\lambda_\Omega)$$
$$+ \left[(p_\mu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ d\imath_*^{(2)}(\lambda_\Omega).$$

Step 3: Calculate the transportation cost of $\hat{\pi}$. Based on our presentation of d$\hat{\pi}$, the q-th order transportation cost of $\hat{\pi}$ is, by definition:

$$C_q^q(\hat{\pi}) = \int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q d\hat{\pi}(\boldsymbol{x}, \boldsymbol{y})$$
$$= \int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \min\left\{ (p_\nu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}), (p_\mu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}) \right\} d\imath_*(\lambda_\Omega)$$
$$+ \int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \left[(p_\mu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ d\imath_*^{(1)}(\lambda_\Omega)$$
$$+ \int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \left[(p_\mu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ d\imath_*^{(2)}(\lambda_\Omega). \tag{18}$$

We now explore the three terms in (18). First of all, since $\imath_*(\lambda_\Omega)$ is a pushforward measure generated by the function $\imath(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{x})$, it is easy to see that

$$\imath_*(\lambda_\Omega)(\{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega : \boldsymbol{x} \neq \boldsymbol{y}\}) = 0.$$

Therefore, the first term in (18) is simply

$$\int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \min\left\{ (p_\nu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}), (p_\mu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}) \right\} d\imath_*(\lambda_\Omega)$$
$$= \int_{(\boldsymbol{x}, \boldsymbol{y})\in\overline{\Omega}\times\overline{\Omega}, \boldsymbol{x}=\boldsymbol{y}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \min\left\{ (p_\nu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}), (p_\mu \circ \jmath)(\boldsymbol{x}, \boldsymbol{y}) \right\} d\imath_*(\lambda_\Omega)$$
$$= \int_{\boldsymbol{x}\in\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{x}\|_2^q \min\{p_\mu(\boldsymbol{x}), p_\nu(\boldsymbol{x})\} d\boldsymbol{x} = 0.$$

As for the second term, notice that $\imath_*^{(1)}(\lambda_\Omega)$ is a pushforward measure generated by the function $\imath^{(1)}(\boldsymbol{x}) = (\boldsymbol{x}, \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x}))$. Therefore by definition,

$$\imath_*^{(1)}(\lambda_\Omega)(\{(\boldsymbol{x}, \boldsymbol{y}) \in \Omega \times \Omega : \boldsymbol{y} \neq \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x})\}) = 0.$$

Hence, the second term in (18) is equal to

$$\int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \left[(p_\mu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(1)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ \mathrm{d}\imath_*^{(1)}(\lambda_\Omega)$$

$$= \int_{(\boldsymbol{x},\boldsymbol{y})\in\overline{\Omega}\times\overline{\Omega}, \boldsymbol{y}=\mathsf{Proj}_{\partial\Omega}(\boldsymbol{x})} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q$$

$$\times \left[(p_\mu \circ \jmath^{(1)})(\boldsymbol{x}, \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x})) - (p_\nu \circ \jmath^{(1)})(\boldsymbol{x}, \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x})) \right]^+ \mathrm{d}\imath_*^{(1)}(\lambda_\Omega)$$

$$= \int_{\boldsymbol{x}\in\overline{\Omega}} \|\boldsymbol{x} - \mathsf{Proj}_{\partial\Omega}(\boldsymbol{x})\|_2^q \left[(p_\mu \circ \jmath^{(1)} \circ \imath^{(1)})(\boldsymbol{x}) - (p_\nu \circ \jmath^{(1)} \circ \imath^{(1)})(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x}$$

$$= \int_\Omega \|\boldsymbol{x} - \partial\Omega\|_2^q \left[p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x}) \right]^+ \mathrm{d}\boldsymbol{x}.$$

Similarly, we can obtain that the third term of (18) is equal to

$$\int_{\overline{\Omega}\times\overline{\Omega}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^q \left[(p_\mu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) - (p_\nu \circ \jmath^{(2)})(\boldsymbol{x}, \boldsymbol{y}) \right]^+ \mathrm{d}\imath_*^{(2)}(\lambda_\Omega)$$

$$= \int_\Omega \left[p_\nu(\boldsymbol{x}) - p_\mu(\boldsymbol{x}) \right]^+ \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x}.$$

Combining these results, we obtain

$$C_q^q(\hat{\pi}) = \int_\Omega [p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})]^+ \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x} + \int_\Omega [p_\nu(\boldsymbol{x}) - p_\mu(\boldsymbol{x})]^+ \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x}$$

$$= \int_\Omega |p_\mu(\boldsymbol{x}) - p_\nu(\boldsymbol{x})| \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x}$$

$$\leq \|p_\mu - p_\nu\|_\infty \int_\Omega \|\boldsymbol{x} - \partial\Omega\|_2^q \mathrm{d}\boldsymbol{x} = \frac{2}{(q+1)(q+2)} \left(\frac{L}{\sqrt{2}} \right)^{q+2} \|p_\mu - p_\nu\|_\infty.$$

Notice that the last equality uses Lemma E.1.

Finally, since $\hat{\pi}$ is an admissible transport from μ to ν, the optimal transport distance between μ and ν, $\mathsf{OT}_q(\mu, \nu)$, should be at most $C_q(\hat{\pi})$. The bound in Theorem 3.1 follows naturally.

Example of converging OT distance while intensity functions diverge. Consider the following sequences of intensity functions

$$p_{\mu_n} = \frac{4^n}{L^2} \mathbb{1}\left\{ \|\boldsymbol{x} - \boldsymbol{u}_n\|_1 < \frac{\sqrt{2}L}{2^{n+1}} \right\}$$

$$p_{\nu_n} = \frac{4^n}{L^2} \mathbb{1}\left\{ \|\boldsymbol{x} - \boldsymbol{d}_n\|_1 < \frac{\sqrt{2}L}{2^{n+1}} \right\},$$

in which

$$\boldsymbol{u}_n = \left(\frac{\sqrt{2}L}{4}, \frac{\sqrt{2}L}{4} + \frac{\sqrt{2}L}{2^{n+1}} \right)$$

$$\boldsymbol{d}_n = \left(\frac{\sqrt{2}L}{4} - \frac{\sqrt{2}L}{2^{n+1}}, \frac{\sqrt{2}L}{4} \right).$$

Essentially, μ_n and ν_n are uniform distributions on two adjacent ℓ_1 balls. It is easy to verify that the total mass of both μ_n and ν_n is 1, and the optimal transport distance between μ_n and ν_n is upper bounded by

$$\mathsf{OT}_q(\mu_n, \nu_n) \leq \frac{L}{2^n} \to 0;$$

on the other hand, the ℓ_∞ distance between the intensity functions clearly diverges as $n \to \infty$:

$$\|p_{\mu_n} - p_{\nu_n}\|_\infty \geq |p_{\mu_n}(\boldsymbol{u}_n) - p_{\nu_n}(\boldsymbol{u}_n)| = \frac{4^n}{L^2} \to \infty.$$

∎

A remark on the bottleneck distance. We argue that there can be no meaningful upper bound for the bottleneck distance OT_∞ by the ℓ_∞ distance between the intensity or density functions. Consider the following example: define T_h as an upper-left triangle in Ω:

$$T_h := \left\{ \boldsymbol{\omega} \in \Omega \mid \|\boldsymbol{\omega} - \partial\Omega\|_2 \geq \frac{L-h}{\sqrt{2}} \right\},$$

and T_h' as a triangle tangent to the diagonal:

$$T_h' := \left\{ \boldsymbol{\omega} \in \Omega \mid \left\| \boldsymbol{\omega} - \left(\frac{L}{2}, \frac{L}{2}\right) \right\|_\infty \leq \frac{h}{2} \right\}.$$

We define μ_h as the uniform distribution on T_h, so that

$$p_{\mu_h}(\boldsymbol{\omega}) = \frac{2}{h^2} \mathbb{1}\{\boldsymbol{\omega} \in T_h\};$$

on the other hand ν is very similar to μ but has a small part of its mass on T_h':

$$p_{\nu_h}(\boldsymbol{\omega}) = \left(\frac{2}{h^2} - h\right) \mathbb{1}\{\boldsymbol{\omega} \in T_h\} + h \mathbb{1}\{\boldsymbol{\omega} \in T_h'\}.$$

As $h \to 0$, it is easy to verify that $\|p_{\mu_h} - p_{\nu_h}\|_\infty = h \to 0$, while $\mathsf{OT}(\mu_h, \nu_h) \to L/\sqrt{2}$. This is because although the densities for μ and ν becomes very close, there is always a small part of the mass of μ in T_h that has to be transported to T_h'; since the bottleneck distance only considers the *maximum* transport cost, it would converge to the limiting distance between T_h and T_h', which is $L/\sqrt{2}$. It is easy to generalize this example to the case where p_{μ_h} and p_{ν_h} are smooth.

E.2   Proof of Theorem C.6

Both theorems are classic results on the bias of kernel estimators and are proved by the smoothness of the target functions as supposed by Assumption 3.5. We here provides the proof of Theorem C.6 (a), and part (b) can be proved in a completely similar fashion.

We firstly clarify the specific smoothness condition proposed by Assumption 3.5. It guarantees Hence, we can represent the bias of $\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})]$ as an integral. Since $\bar{\mu}_n$ is an unbiased estimator for $\mathbb{E}[\mu]$,

$$\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega}) = \mathbb{E}\left[\int_{\boldsymbol{x}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\bar{\mu}_n\right] - p(\boldsymbol{\omega})$$

$$= \int_{\boldsymbol{x}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\bar{\mu}_n] - p(\boldsymbol{\omega})$$

$$= \int_{\boldsymbol{x}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - p(\boldsymbol{\omega})$$

$$= \int_{\boldsymbol{x}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) [p(\boldsymbol{x}) - p(\boldsymbol{\omega})] \mathrm{d}\boldsymbol{x},$$

where in the last line we applied the property that the kernel function $K(\cdot)$ integrals to 1. We can then apply the smoothness of $p(\cdot)$ as in (12) and obtain that

$$|\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega})|$$

$$\leq \left| \int_{\boldsymbol{x}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \sum_{t=1}^{s-1} \frac{1}{t!} \sum_{t_1 + t_2 = t} \frac{\mathrm{d}^t p(\boldsymbol{\omega})}{\mathrm{d}\omega_1^{t_1} \mathrm{d}\omega_2^{t_2}} (x_1 - \omega_1)^{t_1} (x_2 - \omega_2)^{t_2} \mathrm{d}\boldsymbol{x} \right|$$

$$+ \int_{\boldsymbol{x}} \frac{1}{h^2} \left| K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \right| L_p \|\boldsymbol{x} - \boldsymbol{\omega}\|_2^s \mathrm{d}\boldsymbol{x}$$

By taking a change of variable $\boldsymbol{v} = \frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}$, the first term can be represented as

$$\sum_{t=1}^{s-1} \frac{1}{t!} \sum_{t_1 + t_2 = t} \frac{\mathrm{d}^t p(\boldsymbol{\omega})}{\mathrm{d}\omega_1^{t_1} \mathrm{d}\omega_2^{t_2}} \int_{\|\boldsymbol{v}\|_2 \leq 1} K(\boldsymbol{v}) h^t v_1^{t_1} v_2^{t_2} \mathrm{d}\boldsymbol{v}.$$

The zero-moment condition of the kernel function in Assumption C.5 guarantees that this term equals to 0. Hence,

$$|\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega})| \leq \int_{\boldsymbol{x}} \frac{1}{h^2} \left| K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \right| L_p \|\boldsymbol{x} - \boldsymbol{\omega}\|_2^s \mathrm{d}\boldsymbol{x}$$

$$\underdashed{\boldsymbol{v} = (\boldsymbol{x} - \boldsymbol{\omega})/h} \; L_p h^s \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \|\boldsymbol{v}\|_2^s \mathrm{d}\boldsymbol{v}.$$

E.3 Proof of Theorem 3.4 (a)

A useful claim. The following claim can be applied for easing calculation in Theorem 3.4.

Claim E.2. For $q \in \mathbb{R}$ and $x \in [0, 1]$,

$$1 - x^q \leq (q \vee 1)(1 - x),$$

where $q \vee 1 = \max\{q, 1\}$.

Proof of Claim E.2. If $q \geq 1$ or $q \leq 0$, let $f(x) = 1 - x^q$. Then $f'(x) = -qx^{q-1}$ and $f''(x) = -q(q-1)x^{q-2}$, so $f''(x) \leq 0$ for $x \in [0, 1]$ and f is concave on $[0, 1]$. Then by Jensen's inequality,

$$1 - x^q = f(x) \leq f(1) + f'(1)(x - 1) = q(1 - x).$$

If $q \in [0, 1]$, then $x^q \geq x$ implies

$$1 - x^q \leq 1 - x.$$

Hence combining these gives

$$1 - x^q \leq (q \vee 1)(1 - x).$$

∎

This proof applies the Talagrand's inequality. For this purpose, we firstly define an auxiliary family of functions, and then verify the conditions in Theorems D.6 and D.7 .

Defining an auxiliary function class. Let $\mu_1, \mu_2, ..., \mu_n$ be i.i.d. random measures in $\mathcal{Z}_{L,M}^q$, $\ell_{\boldsymbol{\omega}} = \|\boldsymbol{\omega} - \partial\Omega\|_2 - h$ and $g_{\boldsymbol{\omega}}$ be defined as

$$g_{\boldsymbol{\omega}}(\mu) = \ell_{\boldsymbol{\omega}}^q \left(\int_{\Omega} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\mu - \int_{\Omega} \frac{1}{h^2} K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \right), \tag{19}$$

and K satisfy Assumption C.5. Take $\mathcal{Z} = \mathcal{Z}_{L,M}^q$, $(T, d) = (\Omega_{2h}, \|\cdot\|_2)$, and for all $\mu \in \mathcal{Z}_{L,M}^q$, define $\mathcal{G} = \{g_{\boldsymbol{\omega}} : \boldsymbol{\omega} \in \Omega_{2h}\}$. By definition, $g_{\boldsymbol{\omega}}(\mu)$ has zero mean and the variation of the kernel estimator $\hat{p}_h(\cdot)$ can be represented by

$$\sup_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^q |\hat{p}_h(\boldsymbol{\omega}) - \mathbb{E}[\hat{p}_h(\boldsymbol{\omega})]| = \sup_{\boldsymbol{\omega} \in \Omega_{2h}} \left| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\omega}}(\mu) \right|.$$

Hence, in order to apply the Talagrand's inequality, we need to bound $\|g_{\boldsymbol{\omega}}(\mu)\|_\infty$, $\mathbb{E}[g_{\boldsymbol{\omega}}(\mu)^2]$ and the covering number of $\mathcal{G}$. We provide these upper bound accordingly in the following paragraphs.

Bounding $\|g_{\boldsymbol{\omega}}(\mu)\|_\infty$ **and** $\mathbb{E}[g_{\boldsymbol{\omega}}(\mu)^2]$. Notice that since K vanishes outside the unit circle of $\mathbb{R}^2$, for any $\boldsymbol{x} \notin \Omega_{\ell_{\boldsymbol{\omega}}}$, we have $\left\|\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right\|_2 > 1$ and therefore $K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) = 0$. Hence, for all $\boldsymbol{\omega} \in \Omega_{2h}$,

$$
\begin{aligned}
|g_{\boldsymbol{\omega}}(\mu)| &= \ell_{\boldsymbol{\omega}}^q \left| \int_\Omega \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu - \int_\Omega \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \right| \\
&\le \ell_{\boldsymbol{\omega}}^q \max\left\{ \left| \int_\Omega \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu \right|, \left| \int_\Omega \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \right| \right\} \\
&= \ell_{\boldsymbol{\omega}}^q \max\left\{ \left| \int_{\Omega_{\ell_{\boldsymbol{\omega}}}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu \right|, \left| \int_{\Omega_{\ell_{\boldsymbol{\omega}}}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \right| \right\} \\
&\le \ell_{\boldsymbol{\omega}}^q \frac{\|K\|_\infty}{h^2} \max\left\{ (\mu(\Omega_{\ell_{\boldsymbol{\omega}}}), \mathbb{E}[\mu](\Omega_{\ell_{\boldsymbol{\omega}}})) \right\} \\
&\le \ell_{\boldsymbol{\omega}}^q \frac{\|K\|_\infty M}{h^2 \ell_{\boldsymbol{\omega}}^q} = \frac{\|K\|_\infty M}{h^2}
\end{aligned}
\tag{20}
$$

where in the last inequality we used Lemma D.9. On the other hand, the variance of $g_{\boldsymbol{\omega}}$ is bounded by

$$
\begin{aligned}
\mathbb{E}[g_{\boldsymbol{\omega}}(\mu)^2] &= \ell_{\boldsymbol{\omega}}^{2q} \mathbb{E} \left| \int \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu - \int \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \right|^2 \\
&\le \ell_{\boldsymbol{\omega}}^{2q} \mathbb{E} \left| \int_{\Omega_{\ell_{\boldsymbol{\omega}}}} \frac{1}{h^2} K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu \right|^2 \\
&\le \ell_{\boldsymbol{\omega}}^{2q} \mathbb{E} \left\{ \mu(\Omega_{\ell_{\boldsymbol{\omega}}}) \cdot \int_{\Omega_\ell} \frac{1}{h^4} K^2\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu \right\} \\
&= \ell_{\boldsymbol{\omega}}^{2q} \mu(\Omega_\ell) \int_{\Omega_{\ell_{\boldsymbol{\omega}}}} \frac{1}{h^4} K^2\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mathbb{E}[\mu] \\
&\le \ell_{\boldsymbol{\omega}}^{2q} \cdot \frac{M}{\ell_{\boldsymbol{\omega}}^q} \int_{\|\boldsymbol{x}-\boldsymbol{\omega}\|_2 \le h} \frac{1}{h^4} K^2\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\overset{\boldsymbol{v}=(\boldsymbol{x}-\boldsymbol{\omega})/h}{=\!=\!=\!=\!=\!=} \ell_{\boldsymbol{\omega}}^q M \int_{\|\boldsymbol{v}\|_2 \le 1} \frac{1}{h^2} K^2(\boldsymbol{v}) p(\boldsymbol{\omega}+\boldsymbol{v}h) \mathrm{d}\boldsymbol{v}
\end{aligned}
\tag{21}
$$

$$
\le \ell_{\boldsymbol{\omega}}^q M \frac{1}{h^2} \frac{\|\bar{p}\|_\infty}{\ell_{\boldsymbol{\omega}}^q} \int_{\|\boldsymbol{v}\|_2 \le 1} K^2(\boldsymbol{v}) \mathrm{d}\boldsymbol{v} = \frac{M\|\bar{p}\|_\infty \|K\|_2^2}{h^2}.
\tag{22}
$$

Bounding the covering number of $\mathcal{G}$. For any probability measure Q on $\mathcal{Z}_{L,M}^q$ and any $\eta \in (0, \frac{\|K\|_\infty M}{h^2})$, we aim to bound the covering number of $\mathcal{G}$ with respect to $L_2(Q)$ distance. This requires relating the $L_2(Q)$ distance in $\mathcal{G}$ and the ℓ_2 distance in $\mathbb{R}^2$. Specifically, for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega_{2h}$ and $\mu \in \mathcal{Z}_{L,M}^q$, we can assume without loss of generality that $\ell_{\boldsymbol{\omega}} \le \ell_{\boldsymbol{\omega}'}$. In this case, we firstly observe that

$$
\begin{aligned}
&\left| \ell_{\boldsymbol{\omega}}^q \int K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) \mathrm{d}\mu - \ell_{\boldsymbol{\omega}'}^q \int K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}'}{h}\right) \mathrm{d}\mu \right| \\
&\le \left| \int \ell_{\boldsymbol{\omega}}^q \left[K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}}{h}\right) - K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}'}{h}\right) \right] \mathrm{d}\mu \right| + \left| \int (\ell_{\boldsymbol{\omega}}^q - \ell_{\boldsymbol{\omega}'}^q) K\left(\frac{\boldsymbol{x}-\boldsymbol{\omega}'}{h}\right) \mathrm{d}\mu \right| \\
&\le \ell_{\boldsymbol{\omega}}^q \int_{\Omega_{\ell_{\boldsymbol{\omega}}}} \frac{L_k}{h} \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_2 \mathrm{d}\mu + \int_{\Omega_{\ell_{\boldsymbol{\omega}'}}} (\ell_{\boldsymbol{\omega}'}^q - \ell_{\boldsymbol{\omega}}^q) \|K\|_\infty \mathrm{d}\mu \\
&\le \ell_{\boldsymbol{\omega}}^q \frac{L_k}{h} \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_2 \mu(\Omega_{\ell_{\boldsymbol{\omega}}}) + \|K\|_\infty (\ell_{\boldsymbol{\omega}'}^q - \ell_{\boldsymbol{\omega}}^q) \mu(\Omega_{\ell_{\boldsymbol{\omega}'}}) \\
&\le \frac{ML_k}{h} \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_2 + M\|K\|_\infty \left[1 - \left(\frac{\ell_{\boldsymbol{\omega}}}{\ell_{\boldsymbol{\omega}'}}\right)^q \right].
\end{aligned}
\tag{23}
$$

Since $\ell_{\boldsymbol{\omega}} \ge \ell_{\boldsymbol{\omega}'} - \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_2$, the last term of (23) can be bounded by using Claim E.2 and $\ell_{\boldsymbol{\omega}} \ge \ell_{\boldsymbol{\omega}'} - \|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_2$

as

$$1 - \left(\frac{\ell_{\boldsymbol{\omega}}}{\ell'_{\boldsymbol{\omega}}}\right)^q \le (q \vee 1)\left(1 - \frac{\ell_{\boldsymbol{\omega}}}{\ell'_{\boldsymbol{\omega}}}\right)$$
$$\le \frac{q \vee 1}{\ell'_{\boldsymbol{\omega}}}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2$$
$$\le \frac{q \vee 1}{h}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2. \tag{24}$$

Notice that in the last line, we applied the fact that since $\boldsymbol{\omega}' \in \Omega_{2h}$, $\ell_{\boldsymbol{\omega}'} = \|\boldsymbol{\omega} - \partial\Omega\|_2 - h \ge h$.

From now on, we use q' to denote $q \vee 1$ for simplicity. Equations (23) and (24) imply that

$$\left|\ell_{\boldsymbol{\omega}}^q \int K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right)\mathrm{d}\mu - \ell_{\boldsymbol{\omega}'}^q \int K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}'}{h}\right)\mathrm{d}\mu\right| \le \frac{M(L_k + q'\|K\|_\infty)}{h}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2.$$

Therefore, the difference between $g_{\boldsymbol{\omega}}(\mu)$ and $g_{\boldsymbol{\omega}'}(\mu)$ can be bounded by

$$|g_{\boldsymbol{\omega}}(\mu) - g_{\boldsymbol{\omega}'}(\mu)| \le \left|\ell_{\boldsymbol{\omega}}^q \int \frac{1}{h^2}K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right)\mathrm{d}\mu - \ell_{\boldsymbol{\omega}'}^q \int \frac{1}{h^2}K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}'}{h}\right)\mathrm{d}\mu\right|$$
$$+ \left|\ell_{\boldsymbol{\omega}}^q \int \frac{1}{h^2}K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h}\right)\mathrm{d}\mathbb{E}[\mu] - \ell_{\boldsymbol{\omega}'}^q \int \frac{1}{h^2}K\left(\frac{\boldsymbol{x} - \boldsymbol{\omega}'}{h}\right)\mathrm{d}\mathbb{E}[\mu]\right|$$
$$\le \frac{2M(L_k + q'\|K\|_\infty)}{h^3}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2.$$

In this way, we have related the distance between $g_{\boldsymbol{\omega}}$ and $g_{\boldsymbol{\omega}'}$ to the distance between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$. Now, for any $\eta \in (0, \frac{\|K\|_\infty M}{h^2})$, we can set $\epsilon = \frac{\eta h^3}{2M(L_K + q'\|K\|_\infty)}$. It is easy to verify that

$$\epsilon < \frac{h^3}{2M(L_K + q'\|K\|_\infty)}\frac{\|K\|_\infty M}{h^2} = \frac{\|K\|_\infty}{2(L_K + q'\|K\|_\infty)}h < h.$$

Hence, we can construct a ϵ-covering of Ω_{2h} in the ℓ_2 distance, denoted as S. It is easy to show that the covering number

$$\mathscr{N}(\Omega_{2h}, \|\cdot\|_2, \epsilon) \le \frac{2L^2}{\epsilon^2}.$$

By definition, for any $\boldsymbol{\omega} \in \Omega_{2h}$, there exists $\boldsymbol{\omega}' \in S$, such that $\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2 \le \epsilon < h < \ell_{\boldsymbol{\omega}'}$. Therefore, for any measure Q on $\mathcal{Z}_{L,M}^q$,

$$\|g_{\boldsymbol{\omega}}(\mu) - g_{\boldsymbol{\omega}'}(\mu)\|_{L_2(Q)} \le \sup_{\mu \in \mathcal{Z}_{L,M}^q} |g_{\boldsymbol{\omega}}(\mu) - g_{\boldsymbol{\omega}'}(\mu)|$$
$$\le \frac{2M(L_K + q'\|K\|_\infty)}{h^3}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2 \le \frac{2M(L_K + q'\|K\|_\infty)}{h^3}\epsilon = \eta.$$

In conclusion,

$$\mathscr{N}(\mathcal{G}, L_2(Q), \eta) \le \mathscr{N}\left(\Omega_{2h}, \|\cdot\|_2, \frac{\eta h^3}{2M(L_K + q'\|K\|_\infty)}\right)$$
$$< \left(\frac{4LM(L_K + q'\|K\|_\infty)}{\eta h^3}\right)^2. \tag{25}$$

Completing the proof. With $\|g_{\boldsymbol{\omega}}(\mu)\|_\infty$, $\mathbb{E}[g_{\boldsymbol{\omega}}(\mu)^2]$ and the covering number of $\mathcal{G}$ bounded as in (20), (21) and (25), we can apply Theorems D.6 and D.7 with

$$\begin{cases} AB = \frac{4LM(L_K + q'\|K\|_\infty)}{h^3}; \\ B = \frac{\|K\|_\infty M}{h^2}; \\ \sigma^2 = \frac{M\|\bar{p}\|_\infty}{h^2}\|K\|_2^2; \\ \nu = 2. \end{cases}$$

This gives us the conclusion that with probability at least $1 - \delta$,

$$\sup_{\boldsymbol{\omega} \in \Omega_{2h}} \left| \frac{1}{n} \sum_{i=1}^{n} g_{\boldsymbol{\omega}}(\mu) \right| \lesssim \frac{2\|K\|_\infty M}{nh^2} \log \left(\frac{4L(L_K + q'\|K\|_\infty)}{\delta h^2 \|K\|_2} \sqrt{\frac{M}{\|\bar{p}\|_\infty}} \right) +$$

$$\sqrt{\frac{2M\|\bar{p}\|_\infty}{n}} \frac{\|K\|_2}{h} \sqrt{\log \left(\frac{4L(L_K + q'\|K\|_\infty)}{\delta h^2 \|K\|_2} \sqrt{\frac{M}{\|\bar{p}\|_\infty}} \right)}.$$

∎

E.4 Proof of Theorem 3.4(b)

Part (b) of Theorem 3.4 can be proved in a similar, though slightly easier, fashion to part (a). We therefore provide a sketch of the proof and omit the details.

Defining an auxiliary function class. For every $\tilde{\mu}$ and $\boldsymbol{\omega} \in \Omega$, define

$$g_{\boldsymbol{\omega}}(\tilde{\mu}) = \int_\Omega \frac{1}{h^2} K \left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h} \right) \mathrm{d}\tilde{\mu} - \int_\Omega \frac{1}{h^2} K \left(\frac{\boldsymbol{x} - \boldsymbol{\omega}}{h} \right) \mathrm{d}\mathbb{E}[\tilde{\mu}],$$

and let $\mathcal{G} = \{g_{\boldsymbol{\omega}} : \boldsymbol{\omega} \in \Omega\}$. It is easy to verify that $\mathbb{E}[g] \equiv 0$ for all $\boldsymbol{\omega} \in \Omega$, and that

$$\|\check{p}_h(\boldsymbol{\omega}) - \tilde{p}(\boldsymbol{\omega})\| = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(\mu_i) \right|.$$

Bounding $\|g\|_\infty$ and $\mathbb{E}[g^2]$. Since $\tilde{\mu}$ and $\mathbb{E}[\tilde{\mu}]$ are normalized measures with a total mass of 1, $\|g\|_\infty$ can be bounded by

$$\|g\|_\infty \leq \frac{\|K\|_\infty}{h^2};$$

in the mean time, Assumption 3.3 (b) guarantees that $\mathbb{E}[g_{\boldsymbol{\omega}}(\tilde{\mu})^2]$ can be bounded by

$$\mathbb{E}[g_{\boldsymbol{\omega}}(\tilde{\mu})^2] \leq \frac{\|\tilde{p}\|_\infty \|K\|_2^2}{h^2}.$$

Bounding the covering number of $\mathcal{G}$. We again apply the Lipchitz property of the kernel function $K(\cdot)$ to conclude that for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$,

$$|g_{\boldsymbol{\omega}}(\tilde{\mu}) - g_{\boldsymbol{\omega}'}(\tilde{\mu})| \leq \frac{2L_K}{h^3} \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2.$$

Hence, using a similar reasoning to the proof of part (a), we can bound the covering number of $\mathcal{G}$ by

$$\mathcal{N}(\mathcal{G}, L^2(Q), \eta) < \left(\frac{4LL_K}{\eta h^3} \right)^2.$$

Completing the proof. Theorem 3.4 (b) is a direct corollary of Theorems D.6 and D.7 with the following choice of parameters:

$$\begin{cases} AB = \frac{4LL_K}{h^3}; \\ B = \frac{\|K\|_\infty}{h^2}; \\ \sigma^2 = \frac{\|\tilde{p}\|_\infty}{h^2} \|K\|_2^2; \\ \nu = 2. \end{cases}$$

E.5 Proof of Theorems 3.6 and C.9

In this section, we provide the proof of Theorem C.9, which gives a minimax lower bound for estimating the weighted persistence intensity function. Theorem 3.6, which gives the minimax lower bound for estimating the persistence density function, can be proved in a similar while simpler fashion, so we omit its proof for brevity.

The main idea of this proof is to build a connection of weighted intensity function $\bar{p}(\cdot)$ and a probability density function. First of all, we can observe the conclusion of Theorem D.10 holds true also when the support for the density function is Ω instead of $[0,1]^2$. Now, notice that for any $\boldsymbol{x} \in \Omega$, we can define the following measure:

$$\mu_{\boldsymbol{x}} = M\delta_{\boldsymbol{x}} ||\boldsymbol{x} - \partial\Omega||_2^{-q}. \tag{26}$$

It is easy to verify that $\mathsf{Pers}_q(\mu_{\boldsymbol{x}}) = M$, so $\mu_{\boldsymbol{x}} \in \mathcal{Z}_{L,M}^q$. Therefore, for any estimator $\hat{p}_n : (\mathcal{Z}_{L,M}^q)^n \to \mathcal{F}$, we can construct the following estimator $\hat{f}_n$:

$$\hat{f}_n(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = \hat{p}_n(\mu_{\boldsymbol{x}_1}, \mu_{\boldsymbol{x}_2}, ..., \mu_{\boldsymbol{x}_n}).$$

Theorem D.10 states that there exists a probability density function $f : \Omega \to \mathbb{R}$ with $||f||_{\infty,\infty}^r \leq B$ such that when $X_1, X_2, ..., X_n \sim$ i.i.d. f,

$$\mathbb{E}||\hat{f}_n(X_1, X_2, ..., X_n) - f||_\infty \geq O\left(n^{-\frac{r}{2r+2}}\right).$$

We can apply the probability density function f to construct a probability measure on $\mathcal{Z}_{L,M}^q$. First, define a map $\Phi : \Omega \to \mathcal{Z}_{L,M}^q$ by $\Phi(\boldsymbol{x}) = \mu_{\boldsymbol{x}}$ in (26). Impose a measure structure on $\mathcal{Z}_{L,M}^q$ by pushforwarding the measure structure on Ω, i.e. $\mathcal{Y} \subset \mathcal{Z}_{L,M}^q$ is measurable if and only if $\Phi^{-1}(\mathcal{Y})$ is measurable in Ω. Define a probability measure P on $\mathcal{Z}_{L,M}^q$ as a pushforward measure, i.e., for any measurable set $\mathcal{Y} \subset \mathcal{Z}_{L,M}^q$,

$$P(\mathcal{Y}) = \int_{\Phi^{-1}(\mathcal{Y})} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Then from the change of variables,

$$\int_{\mathcal{Y}} g(\mu) dP(\mu) = \int_{\Phi^{-1}(\mathcal{Y})} g(\Phi(\boldsymbol{x})) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Now, the intensity for P can be represented as follows: let $p(\cdot)$ be the intensity function for $\mathbb{E}[\mu]$ when $\mu \sim P$, then for all $u \in \Omega$,

$$\bar{p}(\boldsymbol{u}) := ||\boldsymbol{u} - \partial\Omega||_2^q p(\boldsymbol{u}) = Mf(\boldsymbol{u}). \tag{27}$$

To see this fact, consider any Borel set $\mathcal{A} \subset \Omega$. By definition, the expected measure $\mathbb{E}[\mu]$ satisfies

$$\mathbb{E}[\mu](\mathcal{A}) = \mathbb{E}[\mu(\mathcal{A})] = \int_{\mathcal{Z}_{L,M}^q} \mu(\mathcal{A}) dP(\mu)$$

$$= \int_{\Phi^{-1}(\mathcal{Z}_{L,M}^q)} \Phi(\boldsymbol{x})(\mathcal{A}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \int_{\Omega} \mu_{\boldsymbol{x}}(\mathcal{A}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \int_{\Omega} M||\boldsymbol{x} - \partial\Omega||_2^{-q} \mathbf{1}\{\boldsymbol{x} \in \mathcal{A}\} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathcal{A}} M||\boldsymbol{x} - \partial\Omega||_2^{-q} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Since $\mathcal{A}$ can be any Borel set, we get $p(\boldsymbol{u}) = M||\boldsymbol{u} - \partial\Omega||_2^{-q}$ by definition, and Equation (27) follows naturally. Since the ℓ_∞ difference between $\hat{f}_n$ and f is lower bounded, we can obtain

$$\mathbb{E}_P \sup_{\boldsymbol{\omega} \in \Omega} ||\boldsymbol{\omega} - \partial\Omega||_2^q |\hat{p}_n(\boldsymbol{\omega}) - p(\boldsymbol{\omega})| = M\mathbb{E}_f ||\hat{f}_n - f||_\infty \geq O\left(n^{-\frac{r}{2r+2}}\right).$$

∎

E.6 Proof of Theorems and Corollaries regarding linear representations of the persistence measure

The theoretical results regarding linear representations of the persistence measure in Section 3.3 are rather direct applications of the theoretical results on estimating the persistence intensity and density functions. We therefore combine their proofs in this section.

Proof of Theorem C.7. First, consider the bias of $\hat{\Psi}$. For any $\Psi \in \mathscr{F}_{2h,R}$, the bias of $\hat{\Psi}$ is upper bounded by

$$
\left| \mathbb{E}[\hat{\Psi}] - \Psi \right| = \left| \int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega})(\mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega})) \mathrm{d}\boldsymbol{\omega} \right|
$$
$$
\leq \int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega}) \left| \mathbb{E}[\hat{p}_h(\boldsymbol{\omega})] - p(\boldsymbol{\omega}) \right| \mathrm{d}\boldsymbol{\omega}
$$
$$
\leq \sup_{\boldsymbol{\omega} \in \Omega_{2h}} \left| \mathbb{E}[\hat{p}(\boldsymbol{\omega})] - p(\boldsymbol{\omega}) \right| \int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega}) \mathrm{d}\omega.
$$

Then under Assumption 3.5, Theorem C.6 gives an upper bound as

$$
\left| \mathbb{E}[\hat{\Psi}] - \Psi \right| \leq L_p h^s \int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega}) \mathrm{d}\omega \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \, \|\boldsymbol{v}\|_2^2 \, \mathrm{d}v.
$$

And then, the definition of $\mathscr{F}_{2h,R}$ implies

$$
\int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega}) \mathrm{d}\omega \leq \left(\frac{L}{\sqrt{2}} \right)^q \int_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_\omega^{-q} f(\boldsymbol{\omega}) \mathrm{d}\omega \leq \left(\frac{L}{\sqrt{2}} \right)^q R,
$$

Hence it gives a further upper bound for the bias $\left| \mathbb{E}[\hat{\Psi}] - \Psi \right|$ as

$$
\left| \mathbb{E}[\hat{\Psi}] - \Psi \right| \leq L_p \left(\frac{L}{\sqrt{2}} \right)^q h^s R \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \, \|\boldsymbol{v}\|_2^2 \, \mathrm{d}v.
$$

Second, consider the bias of $\check{\Psi}$. For any $\Psi \in \widetilde{\mathscr{F}}_R$, the bias of $\check{\Psi}$ is upper bounded by

$$
\left| \mathbb{E}[\check{\Psi}] - \Psi \right| = \left| \int_{\boldsymbol{\omega} \in \Omega} f(\boldsymbol{\omega})(\mathbb{E}[\check{p}_h(\boldsymbol{\omega})] - \tilde{p}(\boldsymbol{\omega})) \mathrm{d}\boldsymbol{\omega} \right|
$$
$$
\leq \int_{\boldsymbol{\omega} \in \Omega} f(\boldsymbol{\omega}) \left| \mathbb{E}[\check{p}_h(\boldsymbol{\omega})] - \tilde{p}(\boldsymbol{\omega}) \right| \mathrm{d}\boldsymbol{\omega}
$$
$$
\leq \sup_{\boldsymbol{\omega} \in \Omega} \left| \mathbb{E}[\check{p}_h(\boldsymbol{\omega})] - \tilde{p}(\boldsymbol{\omega}) \right| \int_{\boldsymbol{\omega} \in \Omega} f(\boldsymbol{\omega}) \mathrm{d}\omega.
$$

Then under Assumption 3.5, Theorem C.6 and the definition of $\widetilde{\mathscr{F}}_R$ give the upper bound as

$$
\left| \mathbb{E}[\check{\Psi}] - \Psi \right| \leq L_{\tilde{p}} h^s R \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \, \|\boldsymbol{v}\|_2^2 \, \mathrm{d}v.
$$

Proof of Theorem 3.7. The upper bound for the variation of $\hat{\Psi}$ is a direct corollary of Theorem 3.4 (a) and the fact that

$$
\sup_{\Psi \in \mathscr{F}_{2h,R}} \left| \hat{\Psi} - \mathbb{E}[\hat{\Psi}] \right| = \sup_{\Psi \in \mathscr{F}_{2h,R}} \left| \int_{\boldsymbol{\omega} \in \Omega_{2h}} f(\boldsymbol{\omega})[\hat{p}_h(\boldsymbol{\omega}) - \mathbb{E}[\hat{p}_h](\boldsymbol{\omega})] \mathrm{d}\boldsymbol{\omega} \right|
$$
$$
\leq \int_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^{-q} f(\boldsymbol{\omega}) \mathrm{d}\omega \cdot \sup_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^q \left| \hat{p}_h(\boldsymbol{\omega}) - \mathbb{E}[\hat{p}_h](\boldsymbol{\omega}) \right|
$$
$$
\leq R \cdot \sup_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^q \left| \hat{p}_h(\boldsymbol{\omega}) - \mathbb{E}[\hat{p}_h](\boldsymbol{\omega}) \right|;
$$

The upper bound for the variation of $\check{\Psi}$ follows from Theorem 3.4 (b) and a similar relation:

$$\sup_{\check{\Psi} \in \mathscr{F}_R} \left| \check{\Psi} - \mathbb{E}[\check{\Psi}] \right| \leq R \cdot \sup_{\boldsymbol{\omega} \in \Omega} |\check{p}_h(\boldsymbol{\omega}) - \mathbb{E}[\check{p}_h(\boldsymbol{\omega})]|.$$

Proof of Corollaries C.8 and 3.8(a).   For every $\boldsymbol{x} \in \Omega_{2h}$, define

$$f_{\boldsymbol{x}}(\boldsymbol{\omega}) = \mathbb{1}\left\{\boldsymbol{\omega} \in B_{\boldsymbol{x}}\right\},$$

and let

$$\mathscr{F}_{2h,R} = \left\{ \Psi = \int_{\Omega_{2h}} f_{\boldsymbol{x}}(\boldsymbol{\omega}) \mathrm{d}\mathbb{E}[\mu] \, \middle| \, \boldsymbol{x} \in \Omega_{2h} \right\}.$$

Corollary C.8 follows from Theorem C.7 and the fact that

$$\int_{\boldsymbol{\omega} \in \Omega_{2h}} f_{\boldsymbol{x}}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \leq \frac{L^2}{4}$$

for every $\boldsymbol{x} \in \Omega_{2h}$. Similarly, Corollary 3.8 follows from Theorem 3.7 and the fact that

$$\int_{\boldsymbol{\omega} \in \Omega_{2h}} \ell_{\boldsymbol{\omega}}^{-q} f_{\boldsymbol{x}}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \leq C \ell_{\boldsymbol{x}}^{2-q},$$

for a constant C.

Proof of Corollary 3.8(b).   For every $\boldsymbol{x} \in \Omega$, we define

$$f_{\boldsymbol{x}}(\boldsymbol{\omega}) = \mathbb{1}\left\{\boldsymbol{\omega} \in B_{\boldsymbol{x}}\right\},$$

and let

$$\widetilde{\mathscr{F}}_R = \left\{ \widetilde{\Psi} = \int_{\Omega} f_{\boldsymbol{x}}(\boldsymbol{\omega}) \mathrm{d}\mathbb{E}[\tilde{\mu}] \, \middle| \, \boldsymbol{x} \in \Omega \right\}.$$

Corollary C.8(b) follows directly from Theorem 3.7 and the fact that for every $\boldsymbol{x} \in \Omega$,

$$\int_{\boldsymbol{\omega} \in \Omega} f_{\boldsymbol{x}}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \leq \frac{L^2}{4}.$$

E.7   Proof of Theorem C.10

This proof again involves the Talagrand's inequality, and therefore takes a similar shape to the proof of Theorem 3.4. We begin by defining an auxiliary function class.

Defining the auxiliary function class $\mathcal{G}$.   Recall that we choose the weight function as $f(\boldsymbol{\omega}) = \|\boldsymbol{\omega} - \partial\Omega\|_2^q$. Therefore, for any persistence measure $\mu \in \mathcal{Z}_{L,M}^q$, its corresponding persistence surface is characterized by

$$\rho_h(\mu)(\boldsymbol{u}) = \int_{\Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q \frac{1}{h^2} K\left(\frac{\boldsymbol{u} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\mu(\boldsymbol{\omega});$$

hence, by defining

$$g_{\boldsymbol{u}}(\mu) = \int_{\Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q \frac{1}{h^2} K\left(\frac{\boldsymbol{u} - \boldsymbol{\omega}}{h}\right) \mathrm{d}\left(\mu - \mathbb{E}[\mu]\right)(\boldsymbol{\omega})$$

and letting $\mathcal{G} = \{g_{\boldsymbol{u}}(\boldsymbol{\mu}) : \boldsymbol{u} \in \Omega\}$, we observe that $\mathbb{E}[g] = 0$ for all $g \in \mathcal{G}$ and

$$\|\rho_h(\boldsymbol{\mu}_n) - \mathbb{E}[\rho_h(\boldsymbol{\mu})]\|_{\infty} = \sup_{g \in \mathcal{G}} \left\| \frac{1}{n} \sum_{i=1}^{n} g(\mu_i) \right\|.$$

Bounding $\|g\|_\infty$ and $\mathbb{E}[g^2]$. Assumptions 3.2 and C.5 directly implies that for any $g \in \mathcal{G}$ and any $\boldsymbol{u} \in \Omega$,

$$|g_{\boldsymbol{u}}(\mu)| \leq \frac{\|K\|_\infty}{h^2} \max\left\{\int_\Omega \|\boldsymbol{\omega} - \partial\Omega\|_2^q \mathrm{d}\mu, \int_\Omega \|\boldsymbol{\omega} - \partial\Omega\|_2^q \mathrm{d}\mathbb{E}[\mu]\right\}$$

$$= \frac{\|K\|_\infty}{h^2} \max\left\{\mathsf{Pers}_q(\mu), \mathsf{Pers}_q(\mathbb{E}[\mu])\right\} \leq \frac{M\|K\|_\infty}{h^2}.$$

Regarding the variance of g, Assumption 3.3 implies that

$$\mathbb{E}[g_{\boldsymbol{u}}(\mu)^2] \leq \|g\|_\infty \cdot \int_\Omega \|\boldsymbol{\omega} - \partial\Omega\|_2^q \frac{1}{h^2}\left|K\left(\frac{\boldsymbol{u} - \boldsymbol{\omega}}{h}\right)\right| \mathrm{d}\mathbb{E}[\mu]$$

$$\leq \frac{M\|K\|_\infty}{h^2} \int_\Omega \frac{1}{h^2}\left|K\left(\frac{\boldsymbol{u} - \boldsymbol{\omega}}{h}\right)\right| \|\boldsymbol{\omega} - \partial\Omega\|_2^q p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}$$

$$\leq \frac{M\|K\|_\infty}{h^2} \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \, \mathrm{d}\boldsymbol{v} \cdot \sup_{\boldsymbol{\omega} \in \Omega} \|\boldsymbol{\omega} - \partial\Omega\|_2^q p(\boldsymbol{\omega})$$

$$\leq \frac{M\|K\|_1\|K\|_\infty\|\bar{p}\|_\infty}{h^2},$$

where in the third line we applied the change of variable $\boldsymbol{v} = (\boldsymbol{u} - \boldsymbol{\omega})/h$, and let

$$\|K\|_1 := \int_{\|\boldsymbol{v}\|_2 \leq 1} |K(\boldsymbol{v})| \, \mathrm{d}\boldsymbol{v}.$$

Covering number of $\mathcal{G}$. Similar to the proof of Theorem 3.4, we bound the covering number of $\mathcal{G}$ by the Lipchitz property of the kernel function K. For any two points $\boldsymbol{u}, \boldsymbol{u}' \in \Omega$, Assumption C.5 guarantees that

$$\left|K\left(\frac{\boldsymbol{u} - \boldsymbol{\omega}}{h}\right) - K\left(\frac{\boldsymbol{u}' - \boldsymbol{\omega}}{h}\right)\right| \leq \frac{L_K\|\boldsymbol{u} - \boldsymbol{u}'\|_2}{h}.$$

Therefore, it is easy to verify that

$$|g_{\boldsymbol{u}}(\mu) - g_{\boldsymbol{u}'}(\mu)| \leq \frac{ML_K\|\boldsymbol{u} - \boldsymbol{u}'\|_2}{h^3}.$$

A similar reasoning to the proof of Theorem 3.4 yields that the covering number of $\mathcal{G}$ is upper bounded by

$$\mathcal{N}(\mathcal{G}, L^2(Q), \eta) \leq \mathcal{N}\left(\Omega, \|\cdot\|_2, \frac{\eta h^3}{ML_K}\right) \leq 2\left(\frac{LML_K}{\eta h^3}\right)^2.$$

Completing the proof. Theorem C.10 is a direct application of Theorems D.6 and D.7 with the following choice of parameters:

$$\begin{cases} AB = \frac{2LML_k}{h^3}; \\ B = \frac{M\|K\|_\infty}{h^2}; \\ \sigma^2 = \frac{M\|K\|_1\|K\|_\infty\|\bar{p}\|_\infty}{h^2}; \\ \nu = 2. \end{cases}$$

E.8 Proof of Theorems C.1 and C.2

Observe that the persistence diagram of the Vietoris-Rips filtration of $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N)$ is decided purely by $\{\varphi[J](\boldsymbol{X})\}_{J \subset [N], |J|=2}$, in which

$$\varphi[J](\boldsymbol{X}) = \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2,$$

for $J = \{i, j\}$. In what follows, we firstly focus on the proof of Theorem C.1, and apply the techniques to that of Theorem C.2 in a similar manner.

Proof of Theorem C.1. Propositions D.4 and D.3 imply that for any Borel set $B \subseteq \Omega$,

$$\mathbb{E}[\mu](B) = \sum_{r=1}^{R}\sum_{i=1}^{N_r}\sum_{s\in S}\int_{V_r\cap W^s_{J^1_{ir},J^2_{ir}}\cap\Phi[J^1_{ir},J^2_{ir}]^{-1}(B)}\kappa(\boldsymbol{X})\mathrm{d}\boldsymbol{X}$$

$$= \sum_{r=1}^{R}\sum_{i=1}^{N_r}\sum_{s\in S}$$

$$\int_{\Psi^s_{J^1_{ir},J^2_{ir}}(V_r\cap W^s_{J^1_{ir},J^2_{ir}}\cap\Phi[J^1_{ir},J^2_{ir}]^{-1}(B))}\kappa((\Psi^s_{J^1_{ir},J^2_{ir}})^{-1}(u,y))J[\Psi^s_{J^1_{ir},J^2_{ir}}]^{-1}(\boldsymbol{u},\boldsymbol{Y})\mathrm{d}\boldsymbol{Y}\mathrm{d}u,$$

where in the second line we change the variable from $\boldsymbol{X}\in[0,1]^{d\times n}$ to $(\boldsymbol{Y},\boldsymbol{u})$ with $\boldsymbol{Y}\in[0,1]^{nd-2}$ and $\boldsymbol{u}\in\Omega$. Now, a change of order of summation gives

$$\mathbb{E}[\mu](B) = \sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}\sum_{r=1}^{R}\sum_{i=1}^{N_r}I(J^1_{ir}=J_1,J^2_{ir}=J_2)$$

$$\times\int_{\Psi^s_{J_1,J_2}(V_r\cap W^s_{J_1,J_2}\cap\Phi[J_1,J_2]^{-1}(B))}\kappa((\Psi^s_{J_1,J_2})^{-1}(\boldsymbol{u},\boldsymbol{Y}))J[\Psi^s_{J^1_{ir},J^2_{ir}}]^{-1}(u,y)\mathrm{d}\boldsymbol{Y}\mathrm{d}u$$

$$\leq \sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}\sum_{r=1}^{R}\sum_{i=1}^{N_r}I(J^1_{ir}=J_1,J^2_{ir}=J_2)$$

$$\times\int_{\Psi^s_{J_1,J_2}(V_r\cap W^s_{J_1,J_2}\cap\Phi[J_1,J_2]^{-1}(B))}d\sup\kappa\mathrm{d}\boldsymbol{Y}\mathrm{d}\boldsymbol{u}$$

$$\leq \sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}N(B)\int_{\Psi^s_{J_1,J_2}(W^s_{J_1,J_2}\cap\Phi[J_1,J_2]^{-1}(B))}d\sup\kappa\mathrm{d}\boldsymbol{Y}\mathrm{d}u, \tag{28}$$

where $N(B)$ is the number of persistent homology points in B, and in the second line we use the facts that $\{V_r\}_{r=1}^{R}$ are disjoint, $\kappa\leq\sup\kappa$ and $J[\Psi^s_{J^1_{ir},J^2_{ir}}]^{-1}\leq d$. Hence, bounding $\mathbb{E}[\mu](B)$ boils down to characterizing the domain of integration on the right hand side of (28). For this, notice that by definition,

$$(\boldsymbol{Y},\boldsymbol{u})\in\Psi^s_{J_1,J_2}(W^s_{J_1,J_2}\cap\Phi[J_1,J_2]^{-1}(B))$$

$$\leftrightarrow\exists\boldsymbol{X}\in W^s_{J_1,J_2},\text{ such that }\Phi[J_1,J_2](\boldsymbol{X})\in B,\Psi^s_{J_1,J_2}(\boldsymbol{X})=(\boldsymbol{Y},\boldsymbol{u})$$

$$\rightarrow\exists\boldsymbol{X}\in W^s_{J_1,J_2},\text{ such that }\Phi[J_1,J_2](\boldsymbol{X})\in B,\Phi[J_1,J_2](\boldsymbol{X})=\boldsymbol{u},\text{ and }\boldsymbol{Y}\in[0,1]^{Nd-2}$$

$$\rightarrow\boldsymbol{u}\in B,\text{ and }\boldsymbol{Y}\in[0,1]^{Nd-2}.$$

Hence, $\mathbb{E}[\mu](B)$ is upper bounded by

$$\mathbb{E}[\mu](B)\leq N(B)\sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}\int_{\boldsymbol{u}\in B,\boldsymbol{Y}\in[0,1]^{Nd-2}}d\sup\kappa\mathrm{d}\boldsymbol{Y}\mathrm{d}\boldsymbol{u}$$

$$= d\sup\kappa N(B)\sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}\int_{[0,1]^{Nd-2}}\mathrm{d}\boldsymbol{Y}\int_{B}\mathrm{d}u$$

$$= d\sup\kappa N(B)\sum_{s\in S}\sum_{\substack{J_1,J_2\subset[N]\\|J_1|=|J_2|=2\\J_1\neq J_2}}\int_{B}\mathrm{d}\boldsymbol{u}.$$

This effectively means that the intensity function $p(\boldsymbol{u})$ is upper bounded by

$$p(\boldsymbol{u}) \leq \mathbb{E}\left[N(\{\boldsymbol{u}\})\right] d \sup \kappa \sum_{s \in S} \sum_{\substack{J_1, J_2 \subset [N] \\ |J_1| = |J_2| = 2 \\ J_1 \neq J_2}} 1$$

$$< \mathbb{E}\left[N(\{\boldsymbol{u}\})\right] \operatorname{card}(S) |\{(J_1, J_2) : |J_1| = |J_2| = 2, J_1 \neq J_2, J_1 \subset [N], J_2 \subset [N]\}| d \sup \kappa.$$

Now, $N(\{\boldsymbol{u}\}) \leq N_\ell$, so Lemma D.5 implies $\mathbb{E}\left[N(\{\boldsymbol{u}\})\right] \leq CN$. And $\operatorname{card}(S) \leq 4d^2$ and $|\{(J_1, J_2) : |J_1| = |J_2| = 2, J_1 \neq J_2, J_1 \subset [N], J_2 \subset [N]\}| \leq \frac{N^4}{4}$, so

$$p(\boldsymbol{u}) \leq (CN) \cdot (4d^2) \cdot \left(\frac{N^4}{4}\right) \cdot d \sup \kappa$$

$$= C' N^5 d^3 \sup \kappa.$$

Theorem C.1 follows with the choice of

$$\operatorname{poly}(N, d) = N^5 d^3.$$

Proof of Theorem C.2. Propositions D.4 and D.3 imply that for any Borel set $B \subseteq \Omega$, the normalized persistence measure of B is expressed by

$$\mathbb{E}[\tilde{\mu}](B) = \sum_{r=1}^{R} \frac{1}{N_r} \sum_{i=1}^{N_r} \sum_{s \in S} \int_{V_r \cap W^s_{J^1_{ir}, J^2_{ir}} \cap \Phi[J^1_{ir}, J^2_{ir}]^{-1}(B)} \kappa(\boldsymbol{X}) \mathrm{d}\boldsymbol{X}$$

$$\leq \sum_{r=1}^{R} \max_{1 \leq i \leq N_r} \sum_{s \in S} \int_{V_r \cap W^s_{J^1_{ir}, J^2_{ir}} \cap \Phi[J^1_{ir}, J^2_{ir}]^{-1}(B)} \kappa(\boldsymbol{X}) \mathrm{d}\boldsymbol{X}.$$

Hence, same techniques can be applied to show that the persistence density function is upper bounded by

$$\tilde{p}(\boldsymbol{u}) \leq d \sup \kappa \mathbb{E}\left[\frac{N(\{\boldsymbol{u}\})}{N(\{\boldsymbol{u}\})}\right] \sum_{s \in S} \sum_{\substack{J_1, J_2 \subset [N] \\ |J_1| = |J_2| = 2 \\ J_1 \neq J_2}} 1$$

$$\leq d \sup \kappa \max_{1 \leq i \leq N(\boldsymbol{u})} \sum_{s \in S} \sum_{\substack{J_1, J_2 \subset [N] \\ |J_1| = |J_2| = 2 \\ J_1 \neq J_2}} 1$$

$$\leq \operatorname{card}(S) |\{(J_1, J_2) : |J_1| = |J_2| = 2, J_1 \neq J_2, J_1 \subset [N], J_2 \subset [N]\}| d \sup \kappa$$

$$\leq (4d^2) \cdot \left(\frac{N^4}{4}\right) \cdot d \sup \kappa.$$

Theorem C.2 follows from choosing

$$\operatorname{poly}(N, d) = N^4 d^3.$$

E.9 Proof of Theorem C.11

In this proof, we firstly define an auxiliary family of functions, and then verify the conditions in Theorem D.8.

Defining the auxiliary function class. For every $\boldsymbol{x} \in \Omega_\ell$ and $\mu \in \mathcal{Z}^q_{L,M}$, define

$$g_{\boldsymbol{x}}(\mu) = \mu(B_{\boldsymbol{x}}) - \mathbb{E}[\mu](B_{\boldsymbol{x}}), \tag{29}$$

and let $\mathcal{G} = \{g_{\boldsymbol{x}} : \boldsymbol{x} \in \Omega_\ell\}$. It is easy to verify that $\mathbb{E}[g_{\boldsymbol{x}}(\mu)] = 0$ for all $\boldsymbol{x} \in \Omega_\ell$, and that

$$\sup_{\boldsymbol{x} \in \Omega_\ell} \left| \hat{\beta}_{\boldsymbol{x}} - \mathbb{E}[\hat{\beta}_{\boldsymbol{x}}] \right| = \left| \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} g(\mu_i) \right|.$$

Bounding $\|g_x\|_\infty$ and $\mathbb{E}[g_x(\mu)^2]$. For any $\boldsymbol{x} \in \Omega_\ell$, the set $B_{\boldsymbol{x}}$ is contained in Ω_ℓ. Hence for any $\mu \in \mathcal{Z}_{L,M}^q$, $\mu(B_{\boldsymbol{x}})$ and $\mathbb{E}[\mu](B_{\boldsymbol{x}})$ can be bounded as

$$\mu(B_{\boldsymbol{x}}) \leq \mu(\Omega_\ell) \leq \ell^{-q}\mathsf{Pers}_q(\mu) \leq M\ell^{-q},$$
$$\mathbb{E}[\mu](B_{\boldsymbol{x}}) \leq \mathbb{E}[\mu](\Omega_\ell) \leq \ell^{-q}\mathsf{Pers}_q(\mathbb{E}[\mu]) \leq M\ell^{-q}. \tag{30}$$

Hence $\|g_{\boldsymbol{x}}\|_\infty$ can be bounded as

$$\|g_{\boldsymbol{x}}\|_\infty \leq \sup_{\mu \in \mathcal{Z}_{L,M}^q} \max\left\{\mu(B_{\boldsymbol{x}}), \mathbb{E}[\mu](B_{\boldsymbol{x}})\right\} \leq M\ell^{-q}. \tag{31}$$

As for the variance of $g_x(\mu)$, we firstly observe that

$$\mathbb{E}[g_{\boldsymbol{x}}(\mu)^2] \leq \|g_{\boldsymbol{x}}\|_\infty \mathbb{E}[\mu](B_{\boldsymbol{x}}) \tag{32}$$

Now, apart from using the bound $\mathbb{E}[\mu](B_{\boldsymbol{x}}) \leq M\ell^{-q}$ from (30), we can also have tighter bound with respect to ℓ when $q > 1$. To do this, we again take the coordinate transformation

$$\begin{cases} y_1 = \frac{x_2 - x_1}{\sqrt{2}} = \|\boldsymbol{x} - \partial\Omega\|_2, \\ y_2 = \frac{x_2 + x_1}{\sqrt{2}}. \end{cases}$$

It can be easily verified that the determinant of the Jacobian matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$ coordinates is 1, and that the Ω_ℓ can be represented using $\boldsymbol{y}$ coordinates by

$$\Omega_\ell = \left\{(y_1, y_2) : \ell < y_1 \leq \frac{L}{\sqrt{2}}, y_1 \leq y_2 \leq \sqrt{2}L - y_1\right\}.$$

Then, we have a tighter bound with respect to ℓ of $\mathbb{E}[\mu](B_{\boldsymbol{x}})$ when $q > 1$ as

$$\mathbb{E}[\mu](B_{\boldsymbol{x}}) \leq \mathbb{E}[\mu](\Omega_\ell) = \int_{\Omega_\ell} p(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$$
$$= \int_{\Omega_\ell} \|\boldsymbol{u} - \partial\Omega\|_2^{-q} \bar{p}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$$
$$\leq \|\bar{p}\|_\infty \int_\ell^{\frac{L}{\sqrt{2}}} \left(\int_{y_1}^{\sqrt{2}L - y_1} \mathrm{d}y_2\right) y_1^{-q}\mathrm{d}y_1$$
$$\leq \|\bar{p}\|_\infty \int_\ell^{\frac{L}{\sqrt{2}}} \sqrt{2}L y_1^{-q}\mathrm{d}y_1$$
$$\leq \frac{\sqrt{2}L\ell^{1-q}\|\bar{p}\|_\infty}{q - 1}.$$

Hence when we let $(q - 1)_+ = \max\{q - 1, 0\}$,

$$\mathbb{E}[\mu](B_{\boldsymbol{x}}) \leq \min\left\{M\ell^{-q}, \frac{\sqrt{2}L\ell^{1-q}\|\bar{p}\|_\infty}{(q - 1)_+}\right\}. \tag{33}$$

And hence by applying (33) to (32), the variance of $g_x(\mu)$ can be upper bounded as

$$\mathbb{E}[g_{\boldsymbol{x}}(\mu)^2] \leq \|g_{\boldsymbol{x}}\|_\infty \mathbb{E}[\mu](B_{\boldsymbol{x}})$$
$$\leq \min\left\{M^2\ell^{-2q}, \frac{\sqrt{2}ML\ell^{1-2q}\|\bar{p}\|_\infty}{(q - 1)_+}\right\} \tag{34}$$

Polynomial discrimination of $\mathcal{G}$. By definition, the empirical persistent measure μ_i can be represented as

$$\mu_i = \sum_j \delta_{\boldsymbol{r}_{ij}},$$

in which $\boldsymbol{r}_{ij} = (b_{ij}, d_{ij})$ represents the j-th point in the corresponding persistent diagram, with b_{ij} and d_{ij} being its birth and death weight respectively . Without loss of generality, we can sort the points in descending order of their distance to the diagonal $\partial\Omega$. Let $N_i = \mu_i(\Omega_\ell)$, then we have $N_i \leq M\ell^{-q}$. Hence, for every $\boldsymbol{x}$ with $||\boldsymbol{x} - \partial\Omega||_2 = \ell$, $\mu_i(B_{\boldsymbol{x}})$ can be represented as

$$\mu_i(B_{\boldsymbol{x}}) = \sum_{j=1}^{N_i} \mathbb{1}(b_{ij} < x_1)\mathbb{1}(d_{ij} > x_2). \tag{35}$$

With this expression, we are ready to bound the cardinality of $\mathcal{G}(\boldsymbol{\mu}_1^n)$. Notice that for any fixed $\boldsymbol{x}$, the value of the tuple $(g_x(\mu_1), ..., g_x(\mu_n))$ is completely decided by the Cartesian product of indicator functions

$$\{\mathbb{1}(b_{ij} < x_1)\}_{i\in[n], j\in[N_i]} \times \{\mathbb{1}(d_{ij} > x_2)\}_{i\in[n], j\in[N_i]} := S_b \times S_d.$$

It is easy to see that with the variation of $\boldsymbol{x} = (x_1, x_2)$, the number of different values taken by S_b and S_d can be bounded by

$$1 + \sum_{i=1}^{n} N_i \leq 1 + n \cdot M\ell^{-q}.$$

Hence, the cardinality of $\mathcal{G}(\boldsymbol{\mu}_1^n)$ is bounded by

$$\text{Card}(\mathcal{G}(\boldsymbol{\mu})) \leq \left(M\ell^{-q}n + 1\right)^2. \tag{36}$$

Completing the proof. The theorem is a direct result for applying Theorem D.8 with the following parameters:

$$\begin{cases} A = M\ell^{-q}; \\ B = M; \\ \sigma^2 = \min\left\{M^2\ell^{-2q}, \frac{\sqrt{2}ML\ell^{1-2q}\|\bar{p}\|_\infty}{(q-1)_+}\right\}; \\ \nu = 2. \end{cases}$$

F   EXPERIMENTAL DETAILS

Figure 3 shows two ORBIT5K simulations with different values of r (2.5 and 4) and the corresponding persistent diagrams. Figure 4 displays the kernel intensity functions for the ORBIT5K simulations set with $r = 2.5$ and $r = 4$ for varying sample sizes, while Figure 5 shows persistence density functions. Figures 6 and 7 show the Betti curves and estimated Betti curves using the kernel density function for the ORBIT5K simulations for $r = 2.5$ and $r = 4$.

Figure 8 displays the estimated persistence density functions computed over random draws of varying size of the digits "4" and "8" from the MNIST dataset.

Finally, Figure 9 shows the sample plots, sample persistent diagram, and the kernel-based estimators for the persistence intensity and density functions calculated from 1000 samples of 1000 data points drawn from the uniform distirbution and the power sphere distribution on the unit circle.
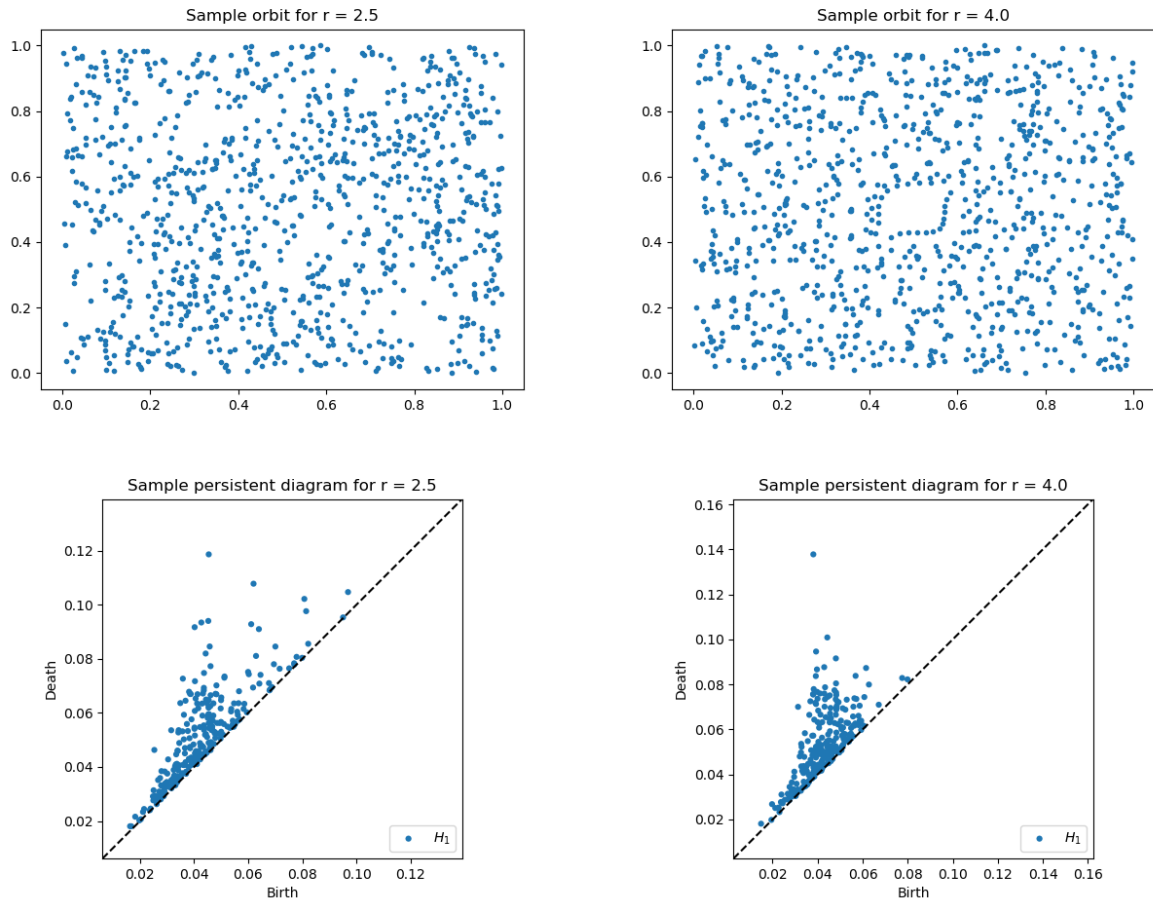
Figure 3: Top row: sample orbits from the ORBIT5K data set with $r = 2.5$ (left) and $r = 4.0$ (right). Bottom row: sample persistent diagrams.
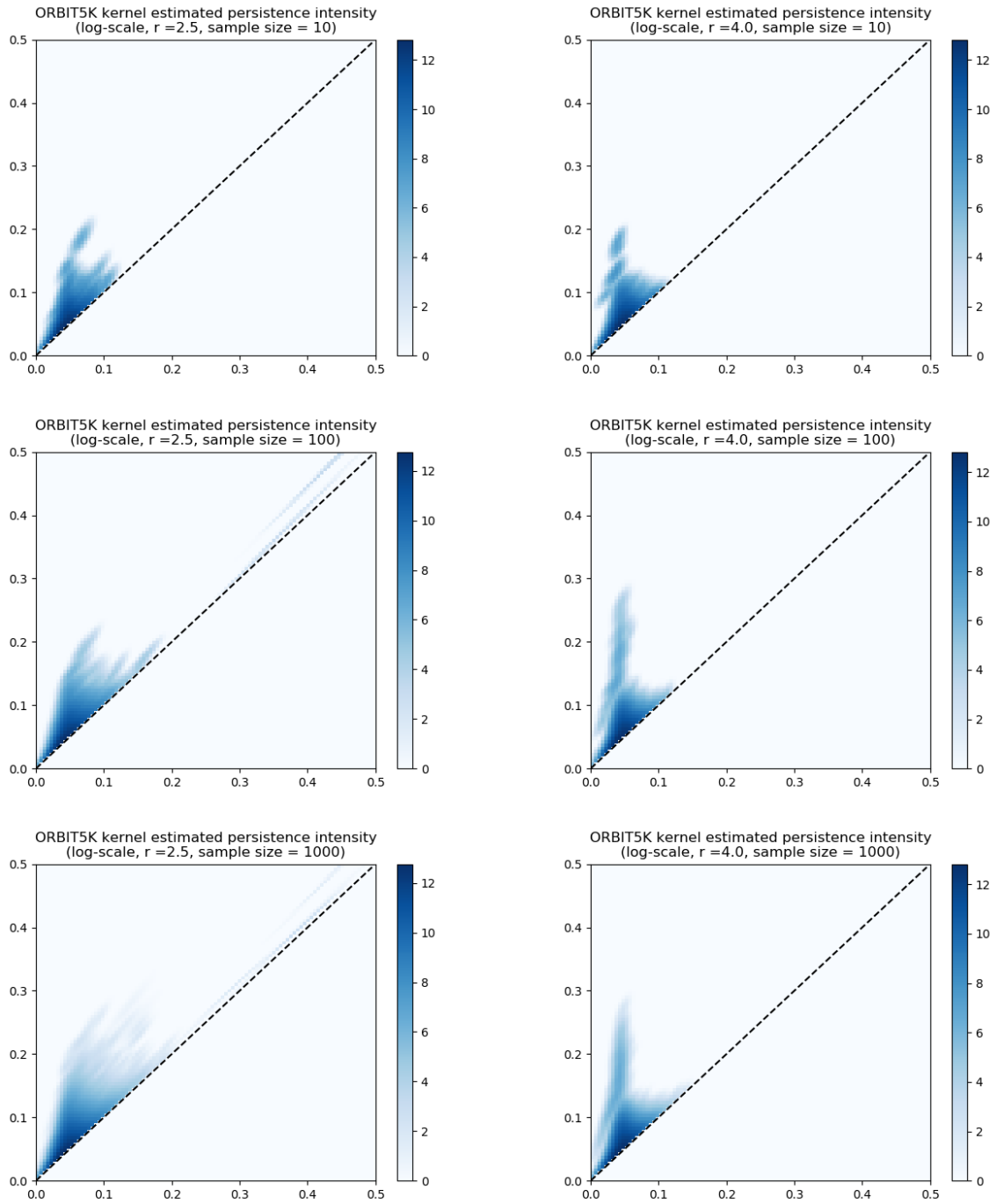
Figure 4: Kernel estimators for the persistence intensity function from the ORBIT5K data set with $r = 2.5$ (left) and $r = 4.0$ (right) and sample sizes 10, 100 and 1000 (top to bottom).
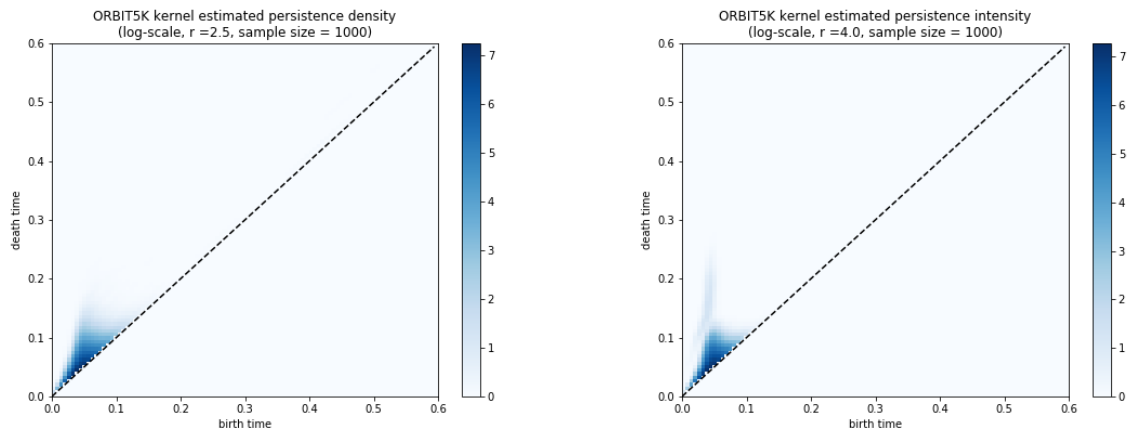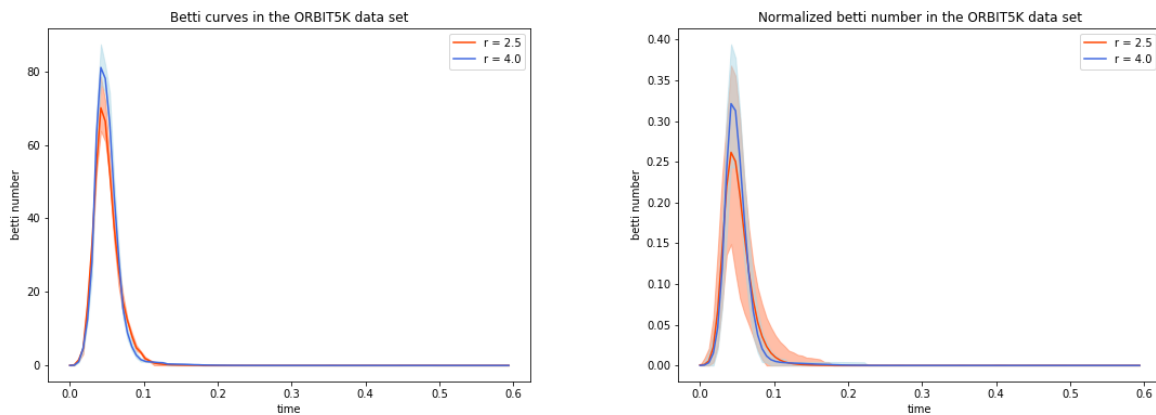
Figure 5: Kernel estimators for the persistence density function from the ORBIT5K data set with $r = 2.5$ (left) and $r = 4.0$ (right) and sample size $n = 1000$.

Figure 6: Empirical betti curves (left) and normalized betti curves (right) from the ORBIT5K data set with $r = 2.5$ and $r = 4.0$. Solid lines show sample average and the shades depict the lower and upper 2.5 percentiles.
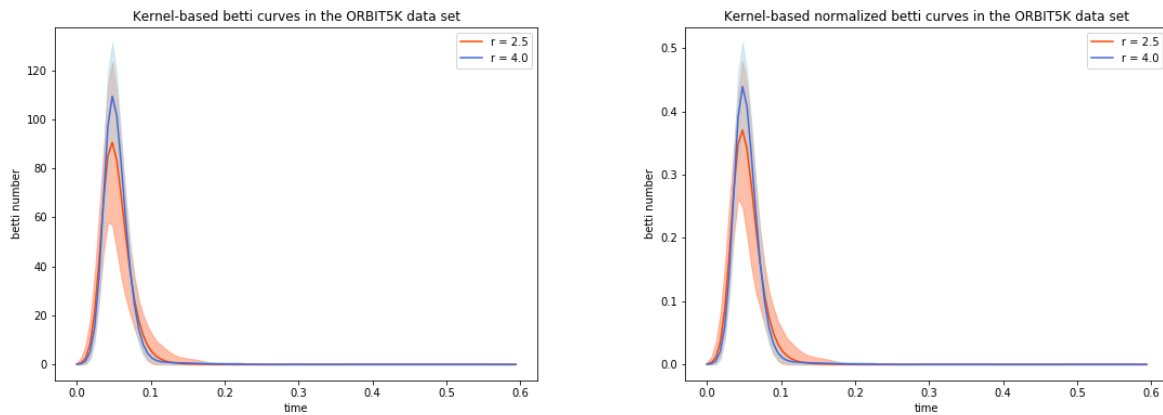

Figure 7: Kernel-based betti curves (left) and normalized betti curves (right) from the ORBIT5K data set with $r = 2.5$ and $r = 4.0$. Solid lines show sample average and the shades depict the lower and upper 2.5 percentiles.
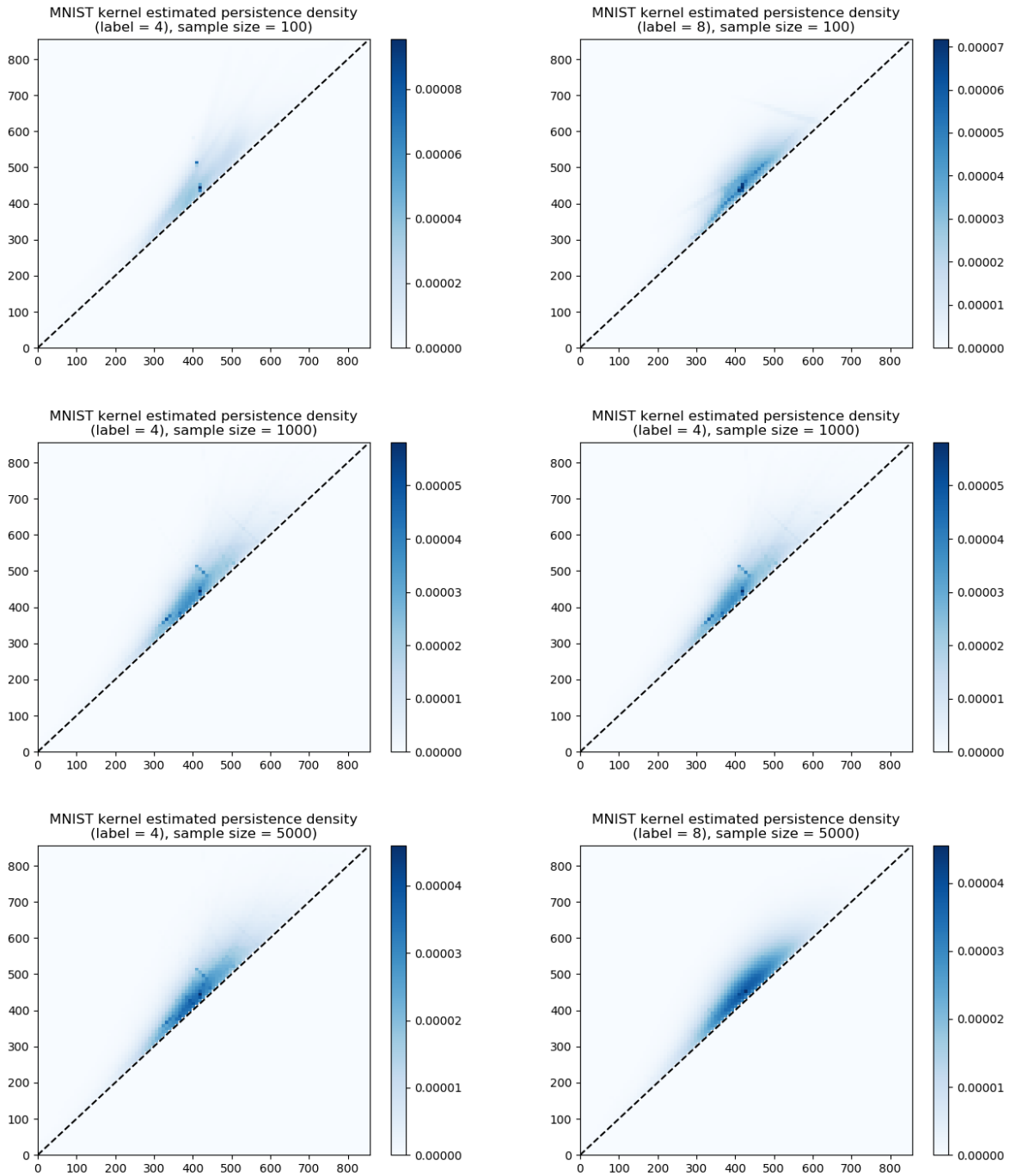
Figure 8: Kernel estimators for the persistence density function from the MNIST data set for the digits 4 (left column) and 8 (right column) based on random draws of sample sizes 100, 1000 and 5000 (top to bottom).
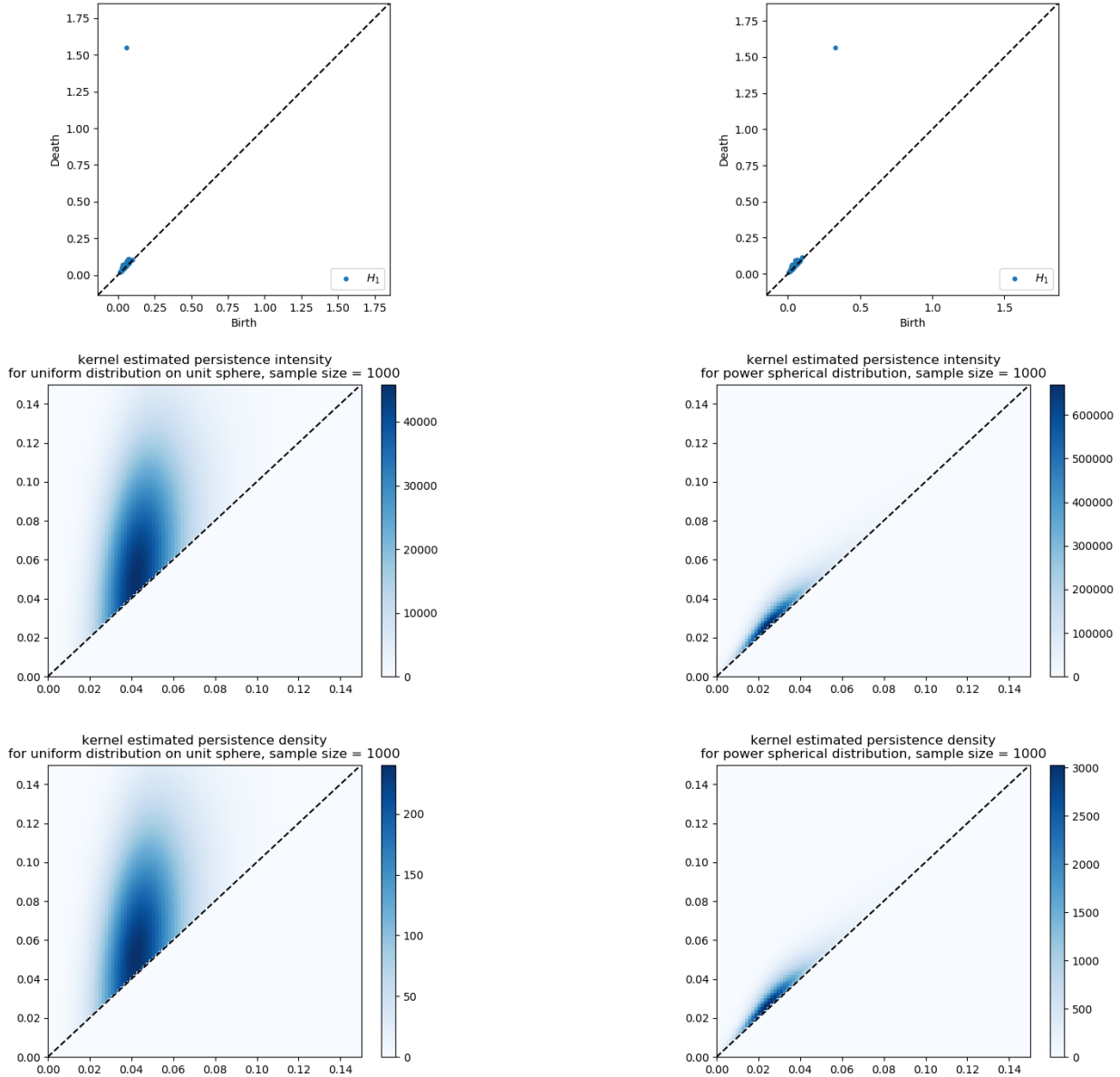
Figure 9: Sample persistence diagram, kernel-estimated persistence intensity and density functions (top to bottom) for the uniform distribution (left) and the power spherical distribution Cao and Aziz (2020) (right) on the unit circle $\mathbb{S}^1$. The parameters for the power spherical distribution are set to $\mu = \frac{\pi}{2}$ and $\kappa = 1$. Each sample contains 1000 points generated in an i.i.d. manner from the corresponding distributions on the unit circle and each perturbed by an additive noise term, sampled i.i.d. from the $N(0, 0.05^2 \boldsymbol{I}_2)$ distribution. The intensity and density functions are estimated through 1000 independent samples.