Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams

Mike, Joshua Lee Maroulas, Vasileios *†

University of Tennessee, Knoxville

Abstract

We introduce a nonparametric way to estimate the global probability density function for a random persistence diagram. Precisely, a kernel density function centered at a given persistence diagram and a given bandwidth is constructed. Our approach encapsulates the number of topological features and considers the appearance or disappearance of features near the diagonal in a stable fashion. In particular, the structure of our kernel individually tracks long persistence features, while considering features near the diagonal as a collective unit. The choice to describe short persistence features as a group reduces computation time while simultaneously retaining accuracy. Indeed, we prove that the associated kernel density estimate converges to the true distribution as the number of persistence diagrams increases and the bandwidth shrinks accordingly. We also establish the convergence of the mean absolute deviation estimate, defined according to the bottleneck metric. Lastly, examples of kernel density estimation are presented for typical underlying datasets.

1 Introduction

Topological data analysis (TDA) encapsulates a range of data analysis methods which investigate the topological structure of a dataset [Edelsbrunner and Harer 2010]. One such method, persistent homology, describes the geometric structure of a given dataset and summarizes this information as a persistence diagram. TDA, and in particular persistence diagrams, have been employed in several studies with topics ranging from classification and clustering [Venkataraman et al. 2016; Adcock et al. 2016; Pereira and de Mello 2015; Marchese and Maroulas 2017] to the analysis of dynamical systems [Perea and Harer 2015; Sgouralis et al. 2017; Guillemard and Iske 2011; Seversky et al. 2016] and complex systems such as sensor networks [De Silva and Ghrist 2007; Xia et al. 2015; Bendich et al. 2016]. In this work, we establish the probability density function (pdf) for a random persistence diagram.

Persistence diagrams offer a topological summary for a collection of d-dimensional data, say \( \{x_i\} \subset \mathbb{R}^d \), which focuses on their global geometric structure of the data. A persistence diagram is a multiset of homological features \( \{(b_i, d_i, k_i)\} \), each representing a \( k_i \)-dimensional hole which appears at scale \( b_i \in \mathbb{R}^+ \) and is filled at scale \( d_i \in (b_i, \infty) \). In general, the dataset arises from any metric space, though restricting to \( \{x_i\} \subset \mathbb{R}^d \) guarantees \( k_i \in \{0, \ldots, d-1\} \). For example, if the data form a time series trajectory \( x_i = f(t_i) \), the associated persistence diagram describes multistability through a corresponding number of persistent 0-dimensional features or periodicity through a single persistent 1-dimensional feature. In a typical persistence diagram, few features exhibit long persistence (range of scales \( d_i - b_i \)), and such features describe important topological characteristics of the underlying dataset. Moreover, persistent features are stable under perturbation of the underlying dataset [Cohen-Steiner et al. 2010].

Persistence diagrams have recently seen intense active research, including significant successful effort toward facilitating previously challenging computations with them; these efforts impact evaluation of Wasserstein distance in [Kerber et al. 2016] and the creation of persistence diagrams with packages such as Dionysus.

* Corresponding Author
† Research has been supported by the Army Research Office (ARO) Grant # W911NF-17-1-0313.
which take advantage of certain properties of simplicial complexes (Chen and Kerber, 2011). Recently, various approaches have defined specific summary statistics such as center and variance (Bobrowski et al., 2014; Mileyko et al., 2011; Turner et al., 2014; Marchese and Maroulas, 2017), birth and death estimates (Emmett et al., 2014), and confidence sets (Fasy et al., 2014).

Here we introduce a nonparametric method to construct density functions for a distribution of persistence diagrams. The development of these densities offers a consistent framework to understand the above summary statistic results through a single viewpoint.

We naturally think of a (random) persistence diagram as a random element which depends upon a stochastic procedure which is used to generate the underlying dataset that it summarizes. Given that geometric complexes are the typical paradigms for application of persistent homology to data analysis, see for example the partial list (De Silva and Ghrist, 2007; Emmett et al., 2014; Guillenard and Iske, 2011; Marchese and Maroulas, 2016; Perera and Harer, 2015; Seversky et al., 2016; Xia et al., 2015; Venkataraman et al., 2016; Edelsbrunner, 2013; Emrani et al., 2014), we consider persistence diagrams which arise from a dataset and its associated Čech filtration. Thus, sample datasets yield sample persistence diagrams without direct access to the distribution of persistence diagrams. In this sense, a distribution of persistence diagrams is defined by transforming the distribution of underlying data under the process used to create a persistence diagram, as discussed in (Mileyko et al., 2011). The persistence diagrams are created through a complex and nonlinear process which relies on the global arrangement of datapoints (see Section 2); thus, the structure of a persistence diagram distribution remains unclear even for underlying data with a well-understood distribution. Indeed, known results for the persistent homology of noise alone, such as (Adler et al., 2014), primarily concern the asymptotics of feature cardinality at coarse scale. With little previous knowledge, we study these distributions through nonparametric means. Kernel density estimation is a well known nonparametric technique for random vectors in \( \mathbb{R}^d \) (Scott, 2015); however, persistence diagrams lack a vector space structure and thus these techniques cannot be applied directly here.

There has been extensive work to devise various maps from persistence diagrams into Hilbert spaces, especially Reproducing Kernel Hilbert Spaces (RKHS). For example, (Chepushtanova et al., 2015) discretizes persistence diagrams via bins, yielding vectors in a high dimensional Euclidean space. The works (Reininghaus et al., 2014) and (Kusano et al., 2016) define kernels between persistence diagrams in a RKHS. By mapping into a Hilbert space, these studies allow the application of machine learning methods such as principal component analysis, random forest, support vector machine, and more. The universality of such a kernel is investigated in (Kwitt et al., 2015); this property induces a metric on distributions of persistence diagrams (by comparing means in the RKHS), as (Kwitt et al., 2015) demonstrates with a two-sample hypothesis test. In a similar vein, (Adler et al., 2017) utilizes Gibbs distributions in order to replicate similar persistence diagrams, e.g. for use in MCMC type sampling.

All previous approaches kernelize to map into a Hilbert space for typical statistical learning techniques. In a similar vein, the studies (Bobrowski et al., 2014) and (Fasy et al., 2014) work with kernel density estimation on the underlying data to estimate a target diagram as the number of underlying datapoints goes to infinity. In both cases, the target diagram is directly associated to the probability density function (pdf) of the underlying data (via the superlevel sets of the pdf). The first work constructs an estimator for the target diagram, while the second defines a confidence set. In either case, kernel density estimation is used to approximate the pdf of the underlying datapoints, assuming the data are independent and identically distributed (i.i.d.). In contrast, our work considers a different kind of kernel density which directly estimates probability densities for a random persistence diagram from a sample of persistence diagrams. This kernel density estimate converges to the true probability density as the number of persistence diagrams goes to infinity.

Instead of a transformed collection or a center diagram, the output of our method is an estimate of a probability density function (pdf) of a random persistence diagram. Access to a pdf facilitates definition and application of many statistical techniques, including hypothesis testing, utilization of Bayesian priors, or likelihood methods. The proposed kernel density is centered at a persistence diagram and describes each feature as having either short or long persistence; by treating each long-persistence point individually and short persistence points collectively, the kernel density strikes a careful balance between accuracy and
computation time. Our method also enables expedient sampling of new persistence diagrams from the kernel density estimate. In contrast to previous methodologies, our kernel density estimate has the potential to describe high probability features in a random persistence diagram, even if these features have brief persistence. Such features are typically indicative of the geometric structure, e.g., curvature, of the dataset rather than its topology.

The homological features \((b_i, d_i, k_i)\) in a persistence diagram come without an ordering and their cardinality is variable, being bounded but not defined by the cardinality of the underlying dataset. Thus, any notion of density must be (i) invariant to the ordering of features and (ii) account for variability in their cardinality. Indeed, the approach used to analyze a collection of persistence diagrams in \((\text{Bendich et al., 2010})\) is a good step toward understanding a random persistence diagram, but requires a choice of order and considers only a fixed number of features and is therefore unsuitable for creating probability densities. In this work, we offer a kernel density with the desirable properties (i) and (ii), which also calls attention to the persistence of each feature. A typical persistence diagram has many features with brief persistence and few with moderate or longer persistence; consequently, our kernel density groups features with short persistence together in order to combat the curse of dimensionality. Indeed, the kernel density still considers features of short persistence, but simplifies their treatment in order to facilitate computation. The kernel density is defined on a pertinent space of finite random sets which is equipped to describe pdfs for random persistence diagrams generated from associated data with bounded cardinality of topological features. In this sense, our kernel density provides estimation of the distribution of persistence diagrams which in turn describes the geometry of the random underlying dataset. The requirement of bounded feature cardinality is trivially satisfied for datasets with bounded cardinality, which is reasonable for application and theory. Indeed, the creation of a persistence diagram from an infinite collection of data is often nonsensical (e.g., for anything with unbounded noise), and a scaling limit should be considered instead.

We establish the kernel density estimation problem through the lens of finite set statistics and we consequently begin with relevant backgrounds in topological data analysis in Section 2 and finite set statistics in Section 3. For further details about these two subjects, the reader may refer respectively to \((\text{Edelsbrunner and Harer, 2010})\) and \((\text{Matheron, 1975})\). Our results are presented in Section 4. In Subsection 4.1 we construct the kernel density associated to a center persistence diagram and kernel bandwidth parameter. This consists of decomposing the center persistence diagram into lower and upper halves, finding pdfs associated to each half, and lastly determining the pdf for their union. After the kernel density is defined and an explicit pdf is delivered in Thm. [1] its convergence is presented in Theorem [2]. Next, Subsection 4.2 presents in detail a specific example of the kernel density. Additionally, an example of persistence diagram kernel density estimation and its convergence are demonstrated for persistence diagrams associated to underlying data with annular distribution. In Subsection 4.3 we define the mean absolute deviation (MAD) as a measure of dispersion, and present the convergence of its kernel density estimator (Thm. [3]). Finally, we end with conclusions and discussion in Section 5. Further examples of KDE convergence and the proofs of the main theorems, Thm. [2] and Thm. [3] are given in the supplementary materials.

## 2 Topological Data Analysis Background

The topological background discussed here builds toward the definition of persistence diagrams, the pertinent objects in this work. We begin by briefly discussing simplicial complexes and homology, an algebraic descriptor for coarse shape in topological spaces. In turn, persistent homology, and its summary, persistence diagrams, are techniques for bringing the power and convenience of homology to describe subspace filtrations of topological spaces. We first consider topological spaces of discernible dimension, called manifolds.

**Definition 2.1.** A topological space \(X\) is called a \(k\)-dimensional manifold if every point \(x \in X\) has a neighborhood which is homeomorphic to an open neighborhood in \(k\)-dimensional Euclidean space.

We generalize the fixed-dimension notion of a manifold in order to define simplicial homology for simplicial complexes. We then discuss the Čech construction which is used to associate simplicial complexes to datasets.
**Definition 2.2.** A $k$-simplex is a collection of $k+1$ linearly independent vertices along with all convex combinations of these vertices:

$$(v_0, ..., v_k) = \left\{ \sum_{i=0}^{k} \alpha_i v_i : \sum_{i=0}^{k} \alpha_i = 1 \text{ and } \alpha_i \geq 0 \forall i \right\}. \quad (2.1)$$

Topologically, a $k$-simplex is treated as a $k$-dimensional manifold (with boundary). An oriented simplex is typically described by a list of its vertices, such as $(v_0, v_1, v_2)$. The faces of a simplex consist of all the simplices built from a subset of its vertex set; for example, the edge $(v_1, v_2)$ and vertex $(v_2)$ are both faces of the triangle $(v_0, v_1, v_2)$.

**Definition 2.3.** A simplicial complex $\mathcal{K}$ is a collection of simplices wherein

(i) if $\sigma \in \mathcal{K}$, then all its faces are also in $\mathcal{K}$, and

(ii) the intersection of any pair of simplices in $\mathcal{K}$ is another simplex in $\mathcal{K}$.

We denote the collection of $k$-simplices within $\mathcal{K}$ by $\mathcal{K}^k$.

A simplicial complex is realized by the union of all its simplices; an example is shown in Fig. 1. Conditions (i) and (ii) in Defn. 2.3 establish a unique topology on the realization of a simplicial complex which restricts to the subspace topology on each open simplex. For finite simplicial complexes realized in $\mathbb{R}^d$, this topology is also consistent with the Euclidean subspace topology.

**Figure 1:** An example of a simplicial complex realized in $\mathbb{R}^3$. This particular complex has one connected component and two cycles, which generate the 0-homology and 1-homology groups, respectively. The other homology groups are trivial.

Here we define the homology groups for a simplicial complex through purely combinatorial means, which allows for automated computation.

**Definition 2.4.** The chain group (over $\mathbb{Z}$) on a simplicial complex $\mathcal{K}$ of dimension $k$ is denoted by $C_k(\mathcal{K})$ and is defined as formal sums of $k$-simplices in $\mathcal{K}$:

$$C_k(\mathcal{K}) = \left\{ \sum_{\sigma \in \mathcal{K}^k} n_\sigma \sigma : n_\sigma \in \mathbb{Z} \right\}. \quad (2.2)$$

**Definition 2.5.** The $k$-th boundary map is a homomorphism $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$ defined on each simplex as an alternating sum over the faces of one dimension less:

$$\partial_k(v_0, ..., v_k) = \sum_{n=0}^{k} (-1)^n (v_0, ..., v_{n-1}, v_{n+1}, ..., v_k). \quad (2.3)$$

**Remark 2.1.** Chain groups give an algebraic way to describe subsets of simplices as a formal sum. Toward this viewpoint, the chain group is often defined over $\mathbb{Z}_2 = \{0,1\}$ instead of $\mathbb{Z}$. In this case, the boundary maps can be understood classically; e.g., the boundary of a triangle yields (the sum of) its three edges and the boundary of an edge yields (the sum of) its endpoints. When viewed over $\mathbb{Z}$, the presence of sign specifies simplex orientation.
Putting chain groups of every dimension together along with the boundary maps successively defined between them, we obtain a chain complex:

\[
\begin{align*}
\{0\} & \xleftarrow{\partial_0} \mathbb{C}_0(K) & \xleftarrow{\partial_1} \mathbb{C}_1(K) & \xleftarrow{\partial_2} \mathbb{C}_2(K) & \xleftarrow{\partial_3} & \ldots
\end{align*}
\]

(2.4)

The composition of subsequent boundary maps yields the trivial map (Edelsbrunner and Harer, 2010); this property is typically rephrased as \( \text{im}(\partial_{k+1}) \subset \ker(\partial_k) \) which enables definition of the following modular groups.

**Definition 2.6.** The homology group of dimension \( k \) is given by

\[
H_k(K) = \ker(\partial_k)/\text{im}(\partial_{k+1}) = \{ [x] = x + \text{im}(\partial_{k+1}) : x \in \ker(\partial_k) \},
\]

(2.5)

where \([x] = \{ x + y : y \in \text{im}(\partial_{k+1}) \}\) defines the coset equivalence class of \( x \).

The generators of the homology group correspond to topological features of the complex \( K \); for example, generators for the 0-homology group correspond to connected components, generators of 1-homology group correspond to holes in \( K \), etc. The interpretation of these features is exemplified by taking the topological boundary of a \( k+1 \) ball (that is, a \( k \)-sphere); for example, the boundary of an interval is two (disconnected) points while the boundary of a disc is a loop.

We wish to extend the notion of homology for a discrete set of data \( x = \{x_i\}_{i=1}^N \) within a metric space \((X,d_X)\). Treating the set itself as a simplicial complex, its homology yields only the cardinality of the data points. So, we utilize the metric to obtain more information. Here we denote by \( B(x_0,r_0) \) a metric ball centered at \( x_0 \) of radius \( r_0 \). Fix a radius \( r > 0 \) and consider the collection of neighborhoods \( U = \{U_i\} = \{B(x_i,r)\} \) along with its union \( U_r = \bigcup_i B(x_i,r) \). The filtration of sets \( \{U_r\}_{r \in \mathbb{R}^+} \) naturally yields information about the arrangement within \( X \) of the dataset \( x \) at various scales. To make homology computations more tractable for \( U_r \), we instead consider the associated nerve complexes.

**Definition 2.7.** The nerve \( \mathcal{N}(U) \) of a collection of open sets \( U \) is the simplicial complex where a \( k \)-simplex \( (v_{i_0}, \ldots, v_{i_k}) \) is in \( \mathcal{N}(U) \) if and only if \( \bigcap_{j=0}^k U_{i_j} \neq \emptyset \). The nerve of the neighborhoods \( U = \{B(x_i,r)\} \) is called the Čech complex on the data \( \{x_i\} \) at radius \( r \) and is denoted by \( \check{C}ech(x,r) \).

Examples of the Čech complex for the same data at different radii are depicted in Fig. 2 where they are superimposed with the associated neighborhood space. Any nerve complex trivially satisfies the requirements for a simplicial complex (Edelsbrunner and Harer, 2010). Moreover, the nerve theorem states that the nerve and union of a collection of convex sets have similar topology (they are homotopy equivalent) (Hatcher, 2002); specifically, the Čech complex and neighborhood space \( U \) have identical homology for any given radius.

A priori, it is unclear which choice of scale (radius), best describes the data; and oftentimes different scales reveal different information. Thus, to investigate the topology of our data, we consider the appearance...
and disappearance of homological features at growing scale. This multiscale viewpoint, called persistent homology, is introduced in (Edelsbrunner et al., 2002) and yields a topological summary of the data called a persistence diagram. This is possible because we have a growing filtration of complexes, so each complex is included in the next (see Fig. 2). These inclusion maps induce inclusion maps at the chain group level and in turn induce maps (though not typically inclusions) at the level of homology groups. These induced maps $f_{r_1, r_2} : H_k(\tilde{\text{Cech}}(x, r_1)) \to H_k(\tilde{\text{Cech}}(x, r_2))$ are referred to here as the persistence maps, and take features to features (i.e., generators to generators) or to zero (Cohen-Steiner et al., 2007). Thus, each feature is tracked by how far the persistence maps preserve it. In turn, tracking features is boiled down to a very specific algorithm for obtaining the birth and death radii for each homological feature (e.g., see (Edelsbrunner and Harer, 2010)). Features which persist over a large range of scale are typically considered more important, and their presence is stable under small perturbations of the underlying data (Cohen-Steiner et al., 2010).

Persistent homology yields a multiset of homological features, each born at a scale $b_i$, lasting until its death scale $d_i$, with degree of homology $k_i$; in short, it yields a persistence diagram $\mathcal{D} = \{(\xi_i)_{i=1}^M = \{(b_i, d_i, k_i)\}_{i=1}^M$. We interpret the birth-death values as coordinate points with degree of homology as labels. For clarity and simplicity, we ignore any features with death value $d_i = \infty$, since these features are generally a characteristic of the ambient space. In particular, one homological feature with $(b, d, k) = (0, \infty, 0)$ is expected from any Čech filtration.

Specifically, for data in $\mathbb{R}^d$, we consider each feature as an element of

$$W_{d+1} = W \times \{0, ..., d - 1\},$$

where $W = \{(b, d) \in \mathbb{R}^2 : d > b \geq 0\}$ is the infinite wedge. As a topological space, the $d$-fold multiwedge $W_{d+1}$ is treated as $d$-disconnected copies of $W$, where $W$ has the Euclidean metric and topology.

It is desirable to define a metric between persistence diagrams with which to measure topological similarity. In TDA, Hausdorff distance is typically used to compare underlying datasets, while the bottleneck distance (Defn. 2.8) is used to compare their associated persistence diagrams (Fasy et al., 2014; Munch, 2017).

**Definition 2.8.** The bottleneck distance between two persistence diagrams $D_1$ and $D_2$ is given by

$$W_\infty(D_1, D_2) = \min_{\gamma} \max_{x \in D_1} \|x - \gamma(x)\|_\infty.$$  

where $\gamma$ ranges over all possible bijections between $D_1$ and $D_2$ which match in degree of homology. The diagonal $\{b = d\}$ is included in both persistence diagrams with infinite multiplicity so that any feature may be matched to the diagonal.

**Remark 2.2.** Due to the unstable presence of features near the diagonal, typical metrics on persistence diagrams such as the bottleneck distance treat the diagonal as part of every persistence diagram (Mileyko et al., 2011) in order to achieve stability with respect to Hausdorff perturbations of the underlying dataset (Cohen-Steiner et al., 2007). Morally, one considers the diagonal as representing vacuous features which are born and die simultaneously. For convenient computation, the definition of bottleneck distance can be applied to each degree of homology separately.

### 3 Random Persistence Diagrams

In this section we establish background to make the notion of probability density for a random persistence diagram explicit and well-defined. A persistence diagram changes its feature cardinality under small perturbation of the underlying dataset, and these features have no intrinsic order. Consequently, we cannot treat persistence diagrams as elements of a vector space. Instead, we consider a random persistence diagram $D$ as a random multiset of features $D = \{\xi_i\} \subset W_{d+1}$ in the multiwedge defined in Eq. (2.6). For underlying datasets sampled from $\mathbb{R}^d$ with bounded cardinality, the affiliated Čech persistence diagrams also have bounded feature cardinality and degree of homology. Thus, we assume that the cardinality of a
random persistence diagram is bounded above by some value $|D| \leq M \in \mathbb{N}$, and so consider the space $C_{\leq M}(W_{0\text{d-1}}) = \{ D \text{ multiset in } W_{0\text{d-1}} : |D| \leq M \}$. We view $C_{\leq M}(W_{0\text{d-1}})$ through a list of functions $h_N$ which each map the appropriate dimension of Euclidean space into its corresponding cardinality component, $C_N(W_{0\text{d-1}})$. This viewpoint facilitates the definition of probability densities.

**Definition 3.1.** For each $N \in \{0, \ldots, M\}$, consider the space of $N$ topological features, denoted $C_N(W_{0\text{d-1}}) = \{ D \text{ multiset in } W_{0\text{d-1}} : |D| = N \}$, and the associated map $h_N : W_{0\text{d-1}} \to C_N(W_{0\text{d-1}})$ defined by

$$h_N(\xi_1, \ldots, \xi_N) = \{\xi_1, \ldots, \xi_N\}.$$  

(3.1)

The map $h_N$ creates equivalence classes on $W_{0\text{d-1}}$ according to the action of the permutations $\Pi_N$; specifically, $[Z] = [(\xi_1, \ldots, \xi_N)]_{h_N} = \{(\xi_{\pi(1)}, \ldots, \xi_{\pi(N)}) : \pi \in \Pi_N\}$ for each $Z = (\xi_1, \ldots, \xi_N) \in W_{0\text{d-1}}$. These equivalence classes yield the space $W_{0\text{d-1}}/\Pi_N = \{[\xi]_{h_N} : \xi \in W_{0\text{d-1}}\}$.

(3.2)

We define a random persistence diagram $D$ to be a random element distributed according to some probability measure on $C_{\leq M}(W_{0\text{d-1}})$ for a fixed maximal cardinality $M \in \mathbb{N}$. We denote associated probabilities by $P[\cdot]$ and expected values by $E[\cdot]$. Since $W^N_{0\text{d-1}}/\Pi_N \cong C_N(W_{0\text{d-1}})$, we work toward defining probability densities on the collection of Euclidean spaces $\bigcup_{N=0}^M W^N_{0\text{d-1}}$.

**Definition 3.2.** For a given random persistence diagram $D$ and any Borel subset $A$ of $W_{0\text{d-1}}$, the belief function $\beta_D$ is defined as

$$\beta_D(A) = P[D \subset A].$$  

(3.3)

Since $A$ is a Borel subset of $W_{0\text{d-1}}$, the collection $O_A = \{ D \in C_{\leq M}(W_{0\text{d-1}}) : D \subset A \}$ is the quotient of $\bigcup_{N=0}^M A_N \subset \bigcup_{N=0}^M W^N_{0\text{d-1}}$ under $h_N$; moreover, $A_N$ is clearly Borel in the Euclidean topology of $\bigcup_{N=0}^M W^N_{0\text{d-1}}$. Therefore, since $h_N$ induces a homeomorphism (see Defn 3.1), $O_A$ is a Borel subset of $C_{\leq M}(W_{0\text{d-1}})$. The belief function of a random persistence diagram is similar to the joint cumulative distribution function for a random vector, in particular by yielding a probability density function through Radon-Nikodym type derivatives.

**Definition 3.3.** Fix $\phi$ defined on Borel subsets of $C_{\leq M}(W_{0\text{d-1}})$ into $\mathbb{R}$. For an element $\xi \in W_{0\text{d-1}}$ or a multiset $Z \subset W_{0\text{d-1}}$ with $Z = \{\xi_1, \ldots, \xi_N\}$, the set derivative (evaluated at $\emptyset$) is respectively given by

$$\frac{\delta \phi}{\delta \xi}(\emptyset) = \lim_{n \to \infty} \frac{\phi(B(\xi, 1/n))}{\lambda(B(\xi, 1/n))},$$

$$\frac{\delta \phi}{\delta Z}(\emptyset) = \frac{\delta^N \phi}{\delta \xi_1 \ldots \delta \xi_N} = \left[ \frac{\delta}{\delta \xi_1} \ldots \frac{\delta}{\delta \xi_N} \right] \phi(\emptyset),$$

(3.4)

where $B(\xi, 1/n)$ are Euclidean balls and $\lambda$ indicates Lebesgue measure on $W_{0\text{d-1}}$.

**Remark 3.1.** Defn. 3.3 for set derivatives at the empty set closely mirrors the Radon-Nikodym derivative with respect to Lebesgue measure. The definition of a set derivative evaluated on a nonempty set is more involved, and is found in [Matheron 1975]. Here we are primarily concerned with evaluation at $\emptyset$, since this suffices for the definition of a probability density function. Also, note that set derivatives satisfy the product rule.
Remark 3.2. Restricting to a particular cardinality \( N \), consider \( \phi_N = \phi \circ h_N \), a function on Euclidean space which is invariant under the action of \( \Pi_N \). The viewpoint of \( \phi_N \) elucidates the relationship between set derivatives and Radon-Nikodým derivatives with respect to Lebesgue measure. This viewpoint also shows that the iterated derivative given in Eq. (3.4) is independent of order and thus is well-defined for a multiset \( Z \).

As with typical derivatives, there is a complementary set integration operation for set derivatives. Set derivatives (at 0) are essentially Radon-Nikodým derivatives with order tied to cardinality, and so the corresponding set integral acts like Lebesgue integration summed over each cardinality.

Definition 3.4. Consider a Borel subset \( A \) of \( W_{0d-1} \) and a Borel subset \( O \) of \( C_{\leq M}(W_{0d-1}) \). For a set function \( f : C_{\leq M}(W_{0d-1}) \to \mathbb{R} \), its set integrals over \( A \) and \( O \) are respectively defined according to the following sums of Lebesgue integrals:

\[
\int_A f(Z) \delta Z = \sum_{N=0}^M \frac{1}{N!} \int_{A_N} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N , \tag{3.5a}
\]

\[
\int_O f(Z) \delta Z = \sum_{N=0}^M \frac{1}{N!} \int_{h^{-1}_N(O)} f(h_N(\xi_1, ..., \xi_N)) d\xi_1 ... d\xi_N , \tag{3.5b}
\]

where \( Z = \{\xi_1, ..., \xi_N\} \subset W_{0d-1} \) is a persistence diagram.

Dividing by \( N! \) in Eqs. (3.5a) and (3.5b) accounts for integrating over \( W_{0d-1}^N \) instead of \( W_{0d-1}^N / \Pi_N \cong C_N(W_{0d-1}) \). It has been shown that set derivatives and integrals are inverse operations [Matheron, 1975]; specifically, the set derivative of a belief function yields a probability density for a random diagram \( D \) such that

\[
\beta_D(A) = \int_A \delta \beta_D(\emptyset) \delta Z . \tag{3.6}
\]

Indeed, \( A_N = h^{-1}_N(\{D \subset A\}) \) so that Eq. (3.5a) also holds as an integral over \( O_A = \{D \in C_{\leq M} : D \subset A\} \) in the sense of Eq. (3.5b).

Definition 3.5. For a random persistence diagram \( D \), a global probability density function (global pdf) \( f_D : \bigcup_{N \in \mathbb{N}} W_{0d-1}^N \to \mathbb{R} \) must satisfy

\[
\sum_{\pi \in \Pi_N} f_D(\xi_{\pi(1)}, ..., \xi_{\pi(N)}) = \frac{\delta^N \beta_D}{\delta \xi_1 \cdots \delta \xi_N}(\emptyset) . \tag{3.7}
\]

and is described by its layered restrictions \( f_N = f_D |_{W_{0d-1}^N} : W_{0d-1}^N \to \mathbb{R} \) for each \( N \).

Remark 3.3. It is necessary to make a distinction between local and global densities because the global pdf is not defined on a single Euclidean space, and is instead expressed as a collection of densities over a range of dimensions. Specifically, while each local density \( f_N \) (for input cardinality \( N \)) is defined on \( W_{0d-1}^N \), the global pdf \( f_D \) is defined on \( \bigcup_{N=1}^\infty W_{0d-1}^N \) and restricts to a local density on each input dimension. Each local density \( f_N(Z) = f_D|_{W_{0d-1}^N}(Z) \) decomposes into the product of the conditional density \( f_D(Z | |Z| = N) \) and the cardinality probability \( \mathbb{P}[|Z| = N] \) (this follows from Prop. 3.1). Thus, each local density does not integrate to one, but instead to the associated probability \( \mathbb{P}[|Z| = N] \). Also, the global pdf is not a set function and does not require division by \( N! \), leading to the following relation:

\[
\int_{A_N} f_D(\xi_1, ..., \xi_N) d\xi_1 ... d\xi_N = \frac{1}{N!} \int_{A_N} \delta^N \beta_D(\emptyset) d\xi_1 ... d\xi_N .
\]

Remark 3.4. While the global pdf and its local constituents need not be symmetric with respect to \( \Pi_N \), there is a unique choice of global pdf (up to sets of Lebesgue measure 0) which satisfies Eq. (3.7) and is symmetric under the action of \( \Pi_N \). In this case, we safely abuse notation by denoting \( f_D(\{\xi_1, ..., \xi_N\}) := N! f_D(\xi_1, ..., \xi_N) \) and often write \( f_D(Z) \) and allow context to determine whether \( Z \) denotes a set or a vector.
The following proposition is critical to determine the global pdf for (i) the union of independent singleton diagrams (i.e., $|D_j| \leq 1$), (ii) a randomly chosen cardinality, $N$, followed by $N$ i.i.d. draws from a fixed distribution, and (iii) a random persistence diagram kernel density function. The proof of this proposition follows similar arguments to Mahler [1995] (Theorem 17, pp. 155–156).

**Proposition 3.1.** Let $D$ be a random persistence diagram with cardinality bounded by $M$ and let $\beta_D(S) = \mathbb{P}(D \subset S)$ be the belief function for $D$. Then $\beta_D$ expands as

$$\beta_D(S) = a_0 + \sum_{m=1}^{M} a_m q_m(S),$$

where $a_m = \mathbb{P}(|D| = m)$ and $q_m(S) = \mathbb{P}[D \subset S | |D| = m]$.

**Remark 3.5.** The decomposition in Prop. 3.1 is often applied as a first step toward finding the local density constituents of the global pdf. In particular, $f_N = f_D|_{W_{bd-1}} = 0$ for $N > M$.

Lastly, we encounter a computationally convenient summary for a random persistence diagram called the probability hypothesis density (PHD). The integral of the PHD over a subset $U$ in $W_{bd-1}$ gives the expected number of points in the region $U$; moreover, any other function on $W_{bd-1}$ with this property is a.e. equal to the PHD (Goodman et al. 2013).

**Definition 3.6.** (Matheron 1975) The probability hypothesis density (PHD) for a random persistence diagram $D$ is defined as the set function $F_D(a) = \delta \beta_D \delta Z(\{a\})$ and is expressed as a set integral as

$$F_D(a) = \int_{(Z:\{a\} \subset Z)} \delta \beta_D \delta Z.$$

(3.8)

In particular, $\mathbb{E}(|D \cap U|) = \int_U F_D(u) \, du$ for any region $U$.

### 4 Kernel Density Estimation

#### 4.1 Construction

To estimate distributions of persistence diagrams, our goal is the creation of a kernel density function about a center persistence diagram $\mathcal{D}$ with a kernel bandwidth parameter $\sigma > 0$, used for defining constituent Gaussians according to Definitions 4.1 and 4.2. Prop. 3.1 leads to the following lemma which is crucial for determining the kernel density. We refer to a random persistence diagram $D$ with $|D| \leq 1$ as a singleton diagram, and such singletons are indexed by superscripts.

**Lemma 4.1.** Consider a multiset of independent singleton random persistence diagrams $\{D_j\}_{j=1}^{M}$. If each singleton $D_j$ is described by the value $q^{(j)} = \mathbb{P}[D_j \neq \emptyset]$ and the subsequent conditional pdf, $p^{(j)}(\xi)$, given $|D_j| = 1$, then the global pdf for $D = \bigcup_{j=1}^{M} D_j$ is given by

$$f_D(\xi_1, \ldots, \xi_N) = \sum_{\gamma \in I(N,M)} Q(\gamma) \prod_{k=1}^{N} p^{(\gamma(k))}(\xi_k),$$

(4.1)

for each $N \in \{0, \ldots, M\}$ where

$$Q(\gamma) = Q^*(\gamma) \prod_{k=1}^{N} q^{(\gamma(k))},$$

(4.2)

$I(N,M)$ consists of all increasing injections $\gamma : \{1, \ldots, N\} \rightarrow \{1, \ldots, M\}$, and

$$Q^*(\gamma) = \frac{\prod_{j=1}^{M} (1 - q^{(j)})}{\prod_{k=1}^{N} (1 - q^{(\gamma(k))}).$$

(4.3)
Proof. Since the singleton events $D^j$ are independent, the belief function for $D = \bigcup_j D^j$ decomposes into $\beta_D(S) = \prod_{j=1}^M \beta_{D^j}(S)$. Next, we employ the product rule for the set derivative (see Defn. 3.3) to obtain the global pdf for $D$ in terms of the singleton belief functions and their first derivatives. Higher derivatives of $\beta_{D^j}$ are zero since $D^j$ are singletons (see Remark 3.5). Thus, the product rule yields first derivatives on all (ordered) subsets of the singleton belief functions:

$$\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{1 \leq j_1 < \ldots < j_N \leq M} \frac{\beta_{D^{j_1}}(\emptyset) \cdots \beta_{D^{j_N}}(\emptyset)}{\beta_D(\emptyset)} \left[ \frac{\delta \beta_{D^{j_1}}(\emptyset)}{\delta \xi_1}(\emptyset) \cdots \frac{\delta \beta_{D^{j_N}}(\emptyset)}{\delta \xi_N}(\emptyset) \right].$$

By Prop. 3.1 we have that $\beta_{D^{j_i}}(\emptyset) = (1 - q^{(j_i)})$ and $\frac{\delta \beta_{D^{j_i}}}{\delta \xi_i}(\emptyset) = q^{(j_i)}(\xi_i)$ and so

$$\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{1 \leq j_1 < \ldots < j_N \leq M} \left[ \prod_{j=1}^M (1 - q^{(j)}) \prod_{k=1}^N q^{(j_k)} \right] \prod_{k=1}^N p^{(j_k)}(\xi_k),$$

which nearly resembles Eq. (4.4). To bridge the gap, we describe the choice of indices $j_i$ by an injective function from $\{1, \ldots, N\}$ into $\{1, \ldots, M\}$. In turn, each such injective function is uniquely determined by the composition of an increasing injection $\gamma \in I(N, M)$ which decides the range of the function and permutations on the domain, $\Pi_N$. These permutations take into account the order of the range. The value of $\mathcal{Q}$ is independent of order, and thus is determined by $\gamma$ as in Eq. (4.2). We reorder the product in order to shift these permutations onto the input variables, obtaining

$$\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} \sum_{\gamma \in I(N, M)} \mathcal{Q}(\gamma) \prod_{k=1}^N p^{(\gamma(k))}(\xi_{\pi(k)}). \tag{4.4}$$

Finally, the global pdf in Eq. (4.1) follows directly from applying Eq. (3.7) to Eq. (4.4).

\begin{remark}
The global pdf in Eq. (4.1), and in particular the sum over $\gamma \in I(N, M)$, accounts for each possible combination of singleton presence. Moreover, summing over permutations as in Eq. (4.4) and dividing by $N!$ yields a symmetric pdf with terms for every possible assignment between singletons and inputs. The weights $\mathcal{Q}(\gamma)$ indicate the probability of each assignment occurring, and is the product of the appropriate probability for each singleton to be either present, $q^{(j)}$, or absent, $1 - q^{(j)}$, for each $j$.
\end{remark}

\begin{example}
Consider two 1-dimensional singleton diagrams, $D^1$ and $D^2$, with probabilities of being nonempty $q^{(1)} = 0.6$ and $q^{(2)} = 0.8$, respectively. The corresponding local densities when nonempty are given by $p^{(1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2}$ and $p^{(2)}(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2}$. Lemma 4.1 yields the global pdf for $D = D^1 \cup D^2$ through a set of local densities $\{f_0, f_1(x), f_2(x,y)\}$ such that $f_0 = \mathbb{P}[D = \emptyset] = (1 - q^{(1)})(1 - q^{(2)}) = 0.08$, $f_1 = f_D|_{\mathbb{R}}$, and $f_2 = f_D|_{\mathbb{R}^2}$. We sum over permutations and divide by $N!$ ($N = 1, 2$ is the input cardinality) to obtain a symmetric global pdf.

$$f_1(x) = (1 - q^{(2)})q^{(1)} p^{(1)}(x) + (1 - q^{(1)})q^{(2)} p^{(2)}(x) = 0.12 \frac{1}{\sqrt{2\pi}} e^{-(x+1)^2/2} + 0.32 \frac{1}{\sqrt{2\pi}} e^{-(x-1)^2/2}, \tag{4.5a}$$

$$f_2(x,y) = \frac{q^{(1)}q^{(2)}}{2} \left[ p^{(1)}(x)p^{(2)}(y) + p^{(1)}(y)p^{(2)}(x) \right] = 0.24 \left( e^{-(x+1)^2+(y+1)^2/2} + e^{-(x+1)^2+(y-1)^2/2} \right). \tag{4.5b}$$

Accounting for each cardinality and following Eq. (4.5a) and Eq. (4.5b), the total probability adds up to

$$\mathbb{P}[D = 0] + \mathbb{P}[D = 1] + \mathbb{P}[D = 2] = f_0 + \int_\mathbb{R} f_1(x) dx + \int_{\mathbb{R}^2} f_2(x,y) dxdy = (0.08) + (0.12 + 0.32) + (0.24 + 0.24) = 1,$$

which is less than 1, since the belief functions are not normalized. To correct for this we apply the normalization condition $\int \mathbb{P} dx = 1$ to obtain the final result.
\end{example}
as desired. The local densities in Eq. (4.5a) and Eq. (4.5b) are plotted in Fig. 3. Though $f_1(x)$ is the sum of two Gaussians, in Fig. 3 (Left) we see that the Gaussian centered at $x = 1$ dominates, while the Gaussian centered at $x = -1$ is only indicated by a heavy left tail. This behavior occurs because $q^{(2)} = 0.8$ is very close to 1.

Now we turn toward defining the kernel density. We first define a random persistence diagram as a union of simpler constituents, and then determine its global pdf by combination in a fashion similar to Lemma 4.1. Indeed, we define the desired kernel density as the global pdf for this composite random diagram. To start, we fix a degree of homology $k$ and consider a center diagram $D \subset \mathcal{W}_k = W \times \{k\}$ (see Eq. (2.6)). Since $k$ is fixed, we treat $D = \{(b_i, d_i)\}_{i=1}^M$ within $W = \{(b, d) \in \mathbb{R}^2 : d > b > 0\}$.

Long persistence points in a persistence diagram represent prominent topological features which are stable under perturbation of underlying data, and so it is important to track each independently. In contrast, we leverage the point of view that the small persistence features near the diagonal are considered together as a single geometric signature as opposed to individually important topological signatures. Toward this end, features with short persistence are grouped together and interpreted through i.i.d. draws near the diagonal. Since features cluster near the diagonal in a typical persistence diagram (see, e.g., Fig. 13 (Right) in the supplementary materials), treating short persistence features collectively simplifies our kernel density and thus speeds up its evaluation. It is imperative that these short persistence features are not ignored, because they still capture crucial geometric information for applications such as classification (Marchese and Maroulas 2016, 2017; De Silva and Ghrist 2007; Xia et al. 2015; Donato et al. 2016; Atienza et al. 2016). Thus, we split $D$ into upper and lower portions according to a bandwidth $\sigma$ as

$$D^u = \{(b_i, d_i, k) \in D : d_i - b_i \geq \sigma\} \text{ and } D^l = \{(b_i, d_i, k) \in D : d_i - b_i < \sigma\}. \quad (4.6)$$

Now define random diagrams $D^u$ centered at $D^u$ and $D^l$ centered at $D^l$ such that $D = D^u \cup D^l$. Ultimately, the global pdf of $D$ centered at $D$ is our kernel density.

**Definition 4.1.** Each feature $\xi_j = (b_j, d_j) \in D^u$ yields an independent random singleton diagram $D^j$ defined by its chance to be nonempty $q^{(j)}$ (via Eq. (4.8)) along with its potential position $(b, d)$ sampled according to a modified Gaussian distribution, denoted by $N^*((b_j, d_j), \sigma I)$. The global pdf for $D^u$ is then determined by Lemma 4.1 where each $p^{(j)}$ is given by the pdf associated with $N^*((b_j, d_j), \sigma I)$, which is given by

$$p^{(j)}(b, d) = \frac{\varphi_j(b, d)}{\mathcal{W} \varphi_j(u, v) du dv} \mathbb{1}_W(b, d), \quad (4.7)$$

where $\varphi_j$ is the pdf of the (unmodified) normal $N((b_j, d_j), \sigma I)$, and $\mathbb{1}_W(\cdot)$ is the indicator function for the wedge.
The global pdf for each $D^j$ is readily obtained by a pair of restrictions. First, we restrict the usual Gaussian distribution to the halfspace $T = \{(b, d) \in \mathbb{R}^2 : b < d \}$. Features sampled below the diagonal are considered to disappear from the diagram and thus we define the chance to be nonempty by

$$q^{(j)} = P(D^j \neq \emptyset) = \int_{\{v > u \}} \varphi_j(u, v) \, du \, dv. \quad (4.8)$$

Afterward, the Gaussian restricted to $T$ is further restricted to $W$ and renormalized to obtain a probability measure as in Eq. (4.7). This double restriction to both $T$ and $W$ is necessary for proper restriction of the Gaussian pdf and definition of $q^{(j)} = P(D^j \neq \emptyset)$. Indeed, restriction to $W$ alone causes points with small birth time to have an artificially high chance to disappear; while restriction to $T$ alone yields nonsensical features with negative radius (with $b < 0$). In kernel density estimation, the effects of this distinction become negligible as the bandwidth goes to zero. In practice, this distinction is important for features with small birth time relative to the bandwidth.

**Remark 4.2.** In the Čech construction of a persistence diagram, a feature lies on the line $b = 0$ if and only if it has degree of homology $k = 0$. Consequently, for a feature $(0, d)$ with $k = 0$, we instead take

$$p^{(j)}(d) = \frac{\phi_j(d)}{\int_{\mathbb{R}^+} \phi_j(u) \, du} \mathbb{1}_{\mathbb{R}^+}(d) \quad \text{and} \quad q^{(j)} = \int_{\mathbb{R}^+} \phi_j(u) \, du$$

where $\phi_j$ is the 1-dimensional Gaussian centered at $d_j$ with standard deviation $\sigma$.

Whereas the large persistence features in $D^u$ have small chance to fall below the diagonal and disappear, the existence of the small persistence features in $D^l$ is volatile: these features disappear and appear fluidly under small changes in the underlying data. The distribution of $D^l$ is described by a probability mass function (pmf) $\nu$ and lower density $p^l$.

**Definition 4.2.** The lower random diagram $D^l$ is defined by choosing a cardinality $N$ according to a pmf $\nu$ followed by $N$ i.i.d. draws according to a fixed density $p^l$. First, take $N_\ell = |\mathcal{D}^\ell|$ and define $\nu(\cdot)$ with mean $N_\ell$ and so that $\nu(n) = 0$ for $n > mN_\ell$ for some $m > 0$ independent of $N_\ell$. The subsequent density $p^l(b, d)$ is given by projecting the lower features $\mathcal{D}^\ell$ of the center diagram $\mathcal{D}$ onto the diagonal $b = d$, then creating a restricted Gaussian kernel density estimation for these features; specifically,

$$p^l(b, d) = \frac{1}{N_\ell} \sum_{(b, d) \in \mathcal{D}^\ell} \frac{1}{\pi \sigma^2} e^{-\left(\frac{(b-d)^2}{2 \sigma^2} + \frac{d^2}{2 \sigma^2}\right)}.$$

(4.9)

Projecting the lower features $\mathcal{D}^\ell$ of the center diagram $\mathcal{D}$ onto the diagonal simplifies later analysis and evaluation of $p^l$; without projecting, a unique normalization factor, similar to $q^{(j)}$ in Defn. 4.1, would be required for each Gaussian summand in Eq. (4.9). By Prop. 3.1 and Eq. (3.7), global pdfs of random persistence diagrams are described by a random vector pdf for each cardinality layer, resulting in the following global pdf for $D^l$:

$$f_{D^l}(\xi_1, \ldots, \xi_N) = \nu(N) \prod_{j=1}^N p^l(\xi_j). \quad (4.10)$$

Combining the expressions for $D^l$ and $D^u$, we arrive at the following proposition.

**Theorem 1.** Fix a center persistence diagram $\mathcal{D}$ and bandwidth $\sigma > 0$. Split $\mathcal{D}$ into $\mathcal{D}^l$ and $\mathcal{D}^u$ according to Eq. (4.6). Define $D^l$ with global pdf from Eq. (4.10), and $D^u$ with global pdf from Eq. (4.11). Treating the random persistence diagrams $D^u$ and $D^l$ as independent, the kernel density centered at $\mathcal{D}$ with bandwidth $\sigma$ is given by

$$K_\sigma(Z, \mathcal{D}) = \sum_{j=0}^{N_u} \nu(N - j) \sum_{\gamma \in \ell(j, N_u)} Q(\gamma) \prod_{k=1}^j p^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^N p^l(\xi_k), \quad (4.11)$$
where \( Z = (\xi_1, \ldots, \xi_N) \) is the input, \( \xi_i = (b_i, d_i) \) for \( i = 1, \ldots, N \) are the features, and \( N_u = |\mathcal{D}^u| \) depends on both \( \mathcal{D} \) and \( \sigma \). Here \( Q(\gamma) \) is given by Eq. (4.2), each \( p^{(j)} \) refers to the modified Gaussian pdf as shown in Eq. (4.7) for its matching feature \( \xi_j \) in \( D^u \), and \( p^\ell \) is given by Eq. (4.9).

Proof: Since \( D^u \) and \( D^v \) are independent random persistence diagrams, the belief function decomposes into
\[
\beta_D(S) = \beta_{D^u}(S) \beta_{D^v}(S).
\]
Moreover, since derivatives above order \( N_u \) vanish for \( \beta_{D^u} \) (see Remark 3.5), the product rule and binomial-type counting yield
\[
\frac{\delta^N \beta_{D^u}}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{j=0}^{N_u} \sum_{1 \leq i_1 \neq \cdots \neq i_j \leq N} \frac{\delta^j \beta_{D^u}}{\delta \xi_{i_1} \ldots \delta \xi_{i_j}}(\emptyset) \frac{\delta^{N-j} \beta_{D^v}}{\delta \xi_{1} \ldots \delta \xi_{i_1} \ldots \delta \xi_{i_j} \ldots \delta \xi_N}(\emptyset)
\]
(4.12)
where \( \delta \xi_i \) indicates that the given index is skipped in the set derivative (having been allocated to the other factor). Similar to the proof of Lemma 4.1, the choice of indices \( i_j \) is replaced with a permutation \( \pi \in \Pi_N \); however, the ordering within each derivative is unrelated to the choice of \( i_j \), leading to \( j! \)-fold and \( (N-j)! \)-fold redundancy within each term.

Taking Eq. (4.10) together with Eq. (3.7) yields
\[
\frac{\delta \beta_{D^v}}{\delta \xi_{\pi(j+1)} \ldots \delta \xi_{\pi(N)}}(\emptyset) = (N-j)! \nu(N-j) \prod_{j=1}^{N-j} p^\ell(\xi_j).
\]
Also, Eq. (4.1) and Eq. (3.7) yield
\[
\frac{\delta \beta_{D^v}}{\delta \xi_{\pi(1)} \ldots \delta \xi_{\pi(j)}}(\emptyset) = \sum_{\pi^* \in \Pi_j} \sum_{\gamma \in I(j, N_u)} Q(\gamma) \prod_{k=1}^{j(\gamma(k))} p^\ell(\xi_{\pi^*(k)}).
\]
We substitute these relations into the final expression of Eq. (4.12). The first of these substitutions is straightforward, while the second has \( j! \)-fold redundant permutations overtop the existing permutations in \( \Pi_N \). These substitutions yield that
\[
\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} K_\sigma(Z, \mathcal{D})
\]
as described in Eq. (4.11) and shows that the kernel \( K_\sigma(Z, \mathcal{D}) \) satisfies the definition of a global pdf for \( D \) (Defn. 3.5). Finally, the sum over permutations is removed according to Eq. (3.7) to obtain the expression for \( f_\mu(Z) = K_\sigma(Z, \mathcal{D}) \).

Remark 4.3. A specific example of the component distributions provided for the kernel in Thm. 7 is presented in Fig. 3. Since the kernel density \( K_\sigma \) of Eq. (4.11) is a probability density according to Defn. 3.5, it is a function on \( J^N_{N_u} \mathcal{W}_{0:1} \) and so the sum of several such kernels is defined by adding each local pdf layer separately.

Remark 4.4. Each feature in the upper random persistence diagram is described independently, while all the features in the lower random persistence diagram are described by a single density \( p^\ell \). Evaluation of the kernel density in Eq. (4.11) is made rapid by factoring the repeated (product) evaluation of \( p^\ell \), a typical 2D Gaussian KDE, despite the kernel’s definition in a very high-dimensional space. Indeed, while evaluation computation increases exponentially when upper features are added (of which there should be few), it only increases linearly for additional lower features. Furthermore, in datasets with too many points, one typically subsamples, e.g. by min-max sampling, to reduce the computational burden of calculating the persistence diagram itself (e.g., see Chazal et al., 2015), yielding a persistence diagram with fewer features.

Remark 4.5. In the definition of our kernel, a single parameter \( \sigma \) has been chosen for both the split of center diagrams, as well as the standard deviation used in the Gaussians which build our kernel. Without loss of generality, this choice simplifies the presentation of the kernel density and the proof of kernel density estimate (KDE) convergence (Theorem 2). In general, the bandwidth parameter \( \sigma_2 \) which refers to the standard deviation used to define the Gaussians (as \( \sigma \) appears in Defs. 4.1 and 4.2) need not be equal to the
splitting parameter \( \sigma_1 \) which determines which points are in \( \mathcal{D}^u \) or \( \mathcal{D}^l \) (as \( \sigma \) appears in Eq. (4.6)). Still, it is certainly desirable that \( \sigma_1 = C\sigma_2 \) when taking a limit of KDEs as the number of persistence diagrams grows to infinity (Theorem 3). For a fixed kernel bandwidth \( \sigma_2 \), increasing \( C \) (and thus \( \sigma_1 \)) moves more features into the lower portion of the diagram. This choice may be useful in practice when underlying data are known to be noisy and more noise-related features are expected near the diagonal. By the same token, for \( \sigma_1 >> \sigma_2 \), projecting the lower features onto the diagonal may lead to significant error in the approximation. On the other hand, taking \( \sigma_1 << \sigma_2 \) eliminates the computational benefit of splitting the diagram and is probably not useful in practice. For most cases, taking \( \sigma_1 = \sigma_2 \), is a reasonable balance between KDE accuracy and evaluation computation.

**Remark 4.6.** Since the associations dictated by \( \gamma \in I(j, N_j) \) in Thm. 4 are known a priori, the calculation is embarrassingly parallelizable, and computation can be made rapid even for the evaluation of the global density function associated with a diagram \( D \) with many features (see, e.g., Fig. 13 (Right) in the supplementary materials). Nevertheless, the density described in Eq. (4.11) is well organized for approximate evaluation. While Eq. (4.11) is sufficient for set integration, it is not symmetric under permutations of the inputs \( \xi_i \), and consequently does not represent the density at a set \( \{\xi_1, ..., \xi_N\} \). A symmetric version is desirable for methods such as maximum likelihood or mode estimation (Gelman et al., 2014). Indeed, a symmetric pdf is available by summing over \( \Pi_N \) as per Eq. (4.7) to obtain the set derivative of the belief function. At no loss of accuracy, the large sum over \( \Pi_N \) need not range over all permutations and one may instead sort over compositions \( \beta \circ \pi \) for \( \beta \in I(j, N) \) and \( \pi \in \Pi_j \) for each \( j \in \{0, ..., |D^u|\} \). Since one expects that \( |D^u| = N_u << N \), this reorganization significantly diminishes the number of computations.

Figure 4: Left: A persistence diagram split according to Eq. (4.6). The dashed black line, \( d = b + \sigma \), separates the diagram into the red upper points of \( \mathcal{D}^u \) and the yellow lower points of \( \mathcal{D}^l \). Right: The red and blue gradients represent the upper singleton densities \( p^1 \) and \( p^2 \) given by Eq. (4.7). The green gradient represents the lower density \( p^l \) defined in Eq. (4.9). While each of these densities is defined on the wedge \( W \subset \mathbb{R}^2 \), the global kernel in Eq. (4.11) is defined on \( \bigcup_N W^N \) for each input-cardinality \( N \).

Since the kernel density is a probability density function for a random persistence diagram, it has an associated probability hypothesis density (See Defn. 3.6).

**Corollary 4.2.** Fix a center persistence diagram \( \mathcal{D} \) and bandwidth \( \sigma > 0 \). Split \( \mathcal{D} \) into \( \mathcal{D}^u \) and \( \mathcal{D}^l \) according to Eq. (4.6). Define \( \mathcal{D}^l \) with global pdf from Eq. (4.10), and \( \mathcal{D}^u \) with global pdf from Eq. (4.11). Treating the random persistence diagrams \( \mathcal{D}^u \) and \( \mathcal{D}^l \) as independent, the probability hypothesis density (PHD) associated with the kernel density centered at \( \mathcal{D} \) with bandwidth \( \sigma \) of Thm. 4 is given by

\[
K_{\sigma, PHD}(\xi, \mathcal{D}) = N_{\xi} p^l(\xi) + \sum_{j=1}^{N_u} q^{(j)} p^{(j)}(\xi),
\]

(4.13)

where the feature \( \xi \) is the input and \( N_u = |\mathcal{D}^u| \) and \( N_l = |\mathcal{D}^l| \) depend on both \( \mathcal{D} \) and \( \sigma \). Here each \( p^{(j)} \) refers to the modified Gaussian pdf as shown in Eq. (4.8) for its matching singleton feature \( \xi_j \) in \( \mathcal{D}^u \), \( q^{(j)} \) given by Eq. (4.9) is the probability each singleton is present, and the lower density \( p^l \) is given by Eq. (4.7).
Proof. The PHD is uniquely defined by its integral over a region $U$, which yields the expected number of points in the region. Consequently, the independent upper and lower random draws which build the kernel contribute additively to the PHD. Within the sum, each singleton density $p^{(j)}$ is weighted by the chance for $D^j$ to be present, $q^{(j)}$ and the lower density $p^f$ is weighted according to the mean draw cardinality, which was chosen to be $|D^f|$. \qed

Notice that in Cor. 4.2, the input for the PHD is a single feature $\xi$ as opposed to a list of features $Z = \{\xi_1, ..., \xi_N\}$ for the global kernel in Thm. \[.\] Furthermore, Thm. \[.\] extends to the analogue result for a center persistence diagram with features of varied degree of homology.

**Corollary 4.3.** Consider a persistence diagram $\mathcal{D} = \bigcup_{k=0}^{d-1} \mathcal{D}_k \times \{k\}$ split according to the degrees of homology with associated random persistence diagrams $D_k$ defined according to Eq. (4.11) for each center diagram $\mathcal{D}_k$.

Treating each $D_k$ as independent, the full global pdf for $D = \bigcup D_k$ centered at $\mathcal{D}$ with bandwidth $\sigma$ is given by

$$K_\sigma(Z, \mathcal{D}) = \Lambda(N) \prod_{k=0}^{d-1} K_\sigma(Z_k, \mathcal{D}_k),$$

where $Z = \bigcup_{k=0}^{d-1} Z_k \times \{k\} \subset W_{0:d-1}$ with each $Z_k \subset W$ of cardinality $|Z_k| = N_k$ within the multi-index $N = (N_0, ..., N_{d-1})$ and

$$\Lambda(N) = \frac{N!}{|N|!} := \prod N_k! \left(\sum N_k\right)!.$$

Proof. The result follows immediately from taking set derivatives of the full belief function $\beta_D(S) = \prod_k \beta_{D_k}(S)$. In particular, the set derivatives $\frac{\delta \beta_{D_k}}{\delta S}(\emptyset)$ are zero unless $Z \subset W_k$. Thus, the product rule leaves only the single term $\frac{\delta \beta_{D_k}}{\delta Z}(\emptyset) = \prod_{k=0}^{d-1} \frac{\delta \beta_{D_k}}{\delta S_{Z,k}}(\emptyset)$. In turn, each kernel global pdf $K_\sigma(Z_k, \mathcal{D}_k)$ is related to the associated belief function derivative by a sum over permutations $\Pi_{N_k}$ (see Eq. (3.7)). Compositions of these permutations are $N_k!$-fold redundant against the $|N|!$ permutations in $\Pi_{|N|}$, yielding the coefficient $\Lambda(N)$. \qed

Next, to prove the convergence (to the target distribution) of the kernel density estimate defined via the kernel established in Thm. \[.] we consider persistence diagrams $\{\mathcal{D}_i\}_{i=1}^u$ which are i.i.d. sampled from a target distribution with global pdf $f$. Toward this end, we require the following assumptions on $f$:

(A1) $f(Z) = 0$ for $|Z| > M \in \mathbb{N}$ (bounded cardinality).

(A2) The local density $f_N : \mathcal{W}_k^N \to \mathbb{R}$ is bounded for each $N \in \{1, ..., M\}$.

(A3) There exists $C_N > 0$ so that $f(\xi_1, ..., \xi_N) \leq C_N \|\xi_1, ..., \xi_N\|^{-2N}$ for each $N \in \{1, ..., M\}$.

The assumptions (A1), (A2), and (A3) describe conditions on the target random persistence diagram pdf. It is important that these assumptions also hold for a random persistence diagram associated with typical (random) underlying datasets. For example, (A1) trivially holds for underlying data in $\mathbb{R}^d$ of bounded cardinality. The conditions (A2) and (A3) hold for underlying data sampled from a compact set $E \subset \mathbb{R}^d$ perturbed by Gaussian noise. The work \[Adler et al., 2014\] (see Cor. 2.3 and Thm. 2.6 therein) describes the persistent homology of noise, and describes a ‘core’ neighborhood. Specifically for Gaussian noise, features are retained in the ‘core’, but then extreme decay occurs for features of arbitrary degree outside the ‘core’. Intuitively, by bounding death values by the diameter of the underlying dataset, one expects that the decay will be at worst a polynomial times Gaussian decay, which is sufficient for (A3).

The following theorem shows that the kernel density estimate converges to the true global pdf of a random persistence diagram as the number of persistence diagrams increases. The pdf tracks not only the birth and death of features, but also their prevalence. In particular, the persistence diagram pdf tied to a random dataset can determine which geometric features are stable regardless of their persistence. The proof of this theorem is delegated to the supplementary materials.
Theorem 2. Consider a random persistence diagram global pdf $f$ satisfying assumptions (A1)-(A3). Define the kernel $K_\sigma(Z,\mathcal{D})$ according to Thm. 2 and consider the kernel density estimate $\hat{f}(Z) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(Z,\mathcal{D}_i)$, with centers $\mathcal{D}_i$ sampled i.i.d. according to global pdf $f$ and bandwidth $\sigma = O(n^{-\alpha})$ chosen with $0 < \alpha < \alpha_{2M}$. Then, as $n \to \infty$, $\hat{f} \to f$ uniformly on compact subsets of $W$.

Remark 4.7. The value of $\alpha_{2M}$ is inherited from bandwidth selection for $2M$-dimensional kernel density estimates (Scott, 2015). While the scaling of the bandwidth in the limit is determined by the maximum cardinality $M$ (and thus, the largest dimension of the local pdfs), choosing a bandwidth for a specific sample is an important step in applying kernel density estimation. If the bandwidth is too narrow, the estimate is overfitted and potentially biased; if the bandwidth is too large, the estimate will be oversmoothed, resulting in accuracy loss. Several methods for bandwidth selection in multivariate kernel estimation are discussed in Silverman (1986). As a general rule of thumb, Silverman (1986) recommends choosing the bandwidth as $\sigma_{opt} = A(K)n^{-1/(2d+4)}$, where $n$ is the sample size (i.e., the number of persistence diagrams), $d$ is the dimension, and $A(K)$ is a constant depending on the kernel, $K$. In particular, one may choose $\sigma \equiv 1/(2M + 4)$ as an unbiased estimator for all local pdfs with cardinalities $m \leq M$ (Scott, 2015). Silverman’s rule of thumb works best for distributions which are nearly Gaussian; for more general distributions, the bandwidth may be chosen empirically.

### 4.2 Examples

Here we provide detailed examples of the kernel density and kernel density estimation of an unknown pdf. For simplicity, we restrict to a single degree of homology, say $k = 1$. Due to the intrinsic high dimension of the kernel, we present contour plots for slices of the kernel density. Specifically, for inputs $((b_1, d_1), \ldots, (b_N, d_N))$, we consider the kernel density evaluated at $(b_i, d_i) \in W$ with $(b_i, d_i)$ fixed for $i \geq 2$. For clarity, the unique symmetric pdf $f_{sym}(\xi_1, \ldots, \xi_N) = \frac{1}{N!} \sum_{\pi \in \Pi_N} f(\xi_{\pi(1)}, \ldots, \xi_{\pi(N)})$ is used in the contour plots (see Remark 3.4). For explicit computation, we choose the probability mass function

$$\nu(N) = \max \left\{ \frac{N_t + 1 - |N_t - N|}{(N_t + 1)^2}, 0 \right\}$$

(4.15)

when evaluating the lower density in Eq. (4.10), where $N_t = |\mathcal{D}|$ is the lower cardinality of the center diagram. This probability mass function is chosen to satisfy the requirements of Defn. 4.2 and specifically has the property that $\nu(N) > 0$ for $0 \leq N \leq 2|\mathcal{D}|$.

Example 2. Consider the center persistence diagram $\mathcal{D} = \{(1,3),(2,4),(1,1.3),(3,3.2)\} \subset W$ and bandwidth $\sigma = 1/2$. We construct the associated kernel density $K_\sigma(Z,\mathcal{D})$ according to Thm. 1 and follow with some plots and analysis of the kernel density. The random persistence diagram $D$ associated with the kernel density $K_\sigma(Z,\mathcal{D})$ has a variable number of features $N = |D|$, consequently, the input diagram $Z = \{\xi_1, \ldots, \xi_N\}$ must have variable length and therefore the kernel density has local definitions (see Rmk. 3.3) on $W$ for each possible input cardinality $N$.

Since each modified Gaussian $p^{(j)}$ (Defn. 4.1) and the lower density $p^\ell$ (Defn. 4.2) integrate to 1 over the wedge $W$, an expression for the probability mass function (pmf) $P[|D| = N]$ can be expressed solely in terms of $\nu$ and $q^{(j)}$:

$$P[|D| = N] = \left[ q^{(1)}q^{(2)} \right] \nu(N - 2) + \left[ q^{(1)}(1 - q^{(2)}) + q^{(2)}(1 - q^{(1)}) \right] \nu(N - 1) + \left[ (1 - q^{(1)})(1 - q^{(2)}) \right] \nu(N)$$

(4.16)

The plot of this pmf is shown in Fig. 3. Recall that $D = D^u \cup D^\ell$, so that $|D| = |D^u| + |D^\ell|$. Since $q^{(j)} \approx 1$ for $j = 1, 2$, $|D^u| = 2$ with high probability and the pmf $P[|D| = N]$ is nearly the pmf for $|D^\ell|$, $\nu$, shifted up by
yields the following expression:

\[ W \text{ restricted to the wedge pdfs } p \]

In order: \((2, 4), (1, 3), \) or none.). Since the values of \( q^{(3)} \) are very close to 1, terms which include the upper pdfs \( p^{(j)} \) have much larger total mass.

2 units. Fig. 5 suggests that understanding the kernel density requires investigation into higher cardinality inputs. In general, it is important to consider input diagrams \( Z \) with \(| Z | \geq | P^u | \).

First, we describe the random diagram associated to the lower features \( D^f \) is described in Defn. 4.2 according to a probability mass function (pmf) \( \nu \) for the cardinality of \( D^f \) and a single probability density \( p^f(b, d) \) for the subsequent features’ locations in the wedge \( W \). The pmf \( \nu \) is defined according to Eq. \((4.15)\) with \( N_f = 2; \) that is, \( \nu((0, 1, 2, 3, 4)) = 1/9, 2/9, 3/9, 2/9, 1/9 \) respectively, and zero otherwise. Following Defn. 4.2 we project the features of \( D^f \) onto the diagonal to obtain \((1 1.15, 3.1, 3.1)\). Relying on Eq. \((4.9)\), the resulting lower density is given by

\[ p^f(b, d) = \frac{2}{\pi} \left[ e^{-\frac{1}{2}(b-1.15)^2+(d-1.15)^2} + e^{-\frac{1}{2}(b-3.1)^2+(d-3.1)^2} \right]. \]

restricted to the wedge \( W \). The coefficient \( \frac{2}{\pi} \) is obtained by a direct substitution into Eq. \((4.9)\).

Due to the flexible input cardinality, the kernel will be expressed and plotted separately for different input cardinalities. For brevity, we present the local kernels on \( W^N \subset \mathbb{R}^N \) for cardinalities \( N = 1, 2, 3 \). First, we consider the probability hypothesis density (or PHD, as defined in Eq. \((3.8)\)) along with the kernel density evaluated at a single input feature in Fig. 6. Recall that the integral of the PHD over a region \( U \) yields the expected number of features in \( U \) (see Defn 3.6). The kernel’s corresponding PHD is a sum of Gaussians as described in Cor. 1.2

\[ K_{\sigma, \text{PHD}}((b, d), \mathcal{D}) = 2p^f(b, d) + q^{(1)}p^{(1)}(b, d) + q^{(2)}p^{(2)}(b, d) \]

\[ = 1.273 \left( e^{-2((b-3.1)^2+|d-3.1|^2)} + e^{-2((b-1.15)^2+|d-1.15|^2)} \right) \]

\[ + 0.635e^{-2((b-2)^2+|d-4|^2)} + 0.635e^{-2((b-1)^2+|d-3|^2)} \]. \]

Next, for input of cardinality \(| Z | = 1\), we obtain an easily viewable 2-dimensional distribution. Thm. 1 yields the following expression:

\[ K_{\sigma}((b_1, d_1), \mathcal{D}) = \nu(0) \left[ (1 - q^{(2)})q^{(1)}p^{(1)}(b_1, d_1) + (1 - q^{(1)})q^{(2)}p^{(2)}(b_1, d_1) \right] \]

\[ + \nu(1) \left[ (1 - q^{(1)})(1 - q^{(2)})p^f(b_1, d_1) \right]. \]

\[ = 7.74 \times 10^{-2} \left( e^{-2((b_1-2)^2+(d_1-4)^2)} + e^{-2((b_1-1)^2+(d_1-3)^2)} \right), \]

\[ + 1.65 \times 10^{-4} p^f(b_1, d_1). \]

The kernel is treated as a global pdf as in Prop. 3.5 and Rmk. 3.3; thus, this 2-D density is only a local density for the whole kernel. Each term is a weighted product of the combination of upper features considered (In order: \((2, 4), (1, 3), \) or none.).
Contour plots of the densities expressed in Eqs. (4.18) and (4.19) (restricted to $W$) are respectively shown in Figs. (a) and (b). In Fig. (a), the PHD indicates that in general, as many features will appear near the diagonal as will appear near the upper features. According to the local kernel shown in Fig. (b), if only a single feature is present, this feature is far more likely to have long persistence. Indeed, the kernel density is defined (see Eq. (4.11)) so that the number of points near the diagonal is fluid (by our choice of $\nu$), whereas the probability of each feature in the upper diagram is nearly 1. In essence, this demonstrates that the kernel density naturally considers features with long persistence to be stable or prominent in density estimation.

Figure 6: Contour maps for (a) the probability hypothesis density associated to the kernel density (Eq. (4.18)) and (b) the kernel density restricted to a single input feature (Eq. (4.19)). The center diagram is indicated by red (upper) and green (lower) points. Scale bars at the right of each plot indicate the range of probability density in each shaded region.

Taking $Z = (\xi_1, \xi_2) = ((b_1, d_2), (b_2, d_2))$, we arrive at a more complex expression for the kernel density when considering 2 input features. From Eq. (4.11), we obtain:

$$K_\sigma((\xi_1, \xi_2); \mathcal{D}) = \nu(0)q^{(1)}q^{(2)}p^{(1)}(b_1, d_1)p^{(2)}(b_2, d_2)$$
$$+ \nu(1) \left[ (1 - q^{(2)})q^{(1)}p^{(1)}(b_1, d_1) + (1 - q^{(1)})q^{(2)}p^{(2)}(b_1, d_1) \right] p^f(b_2, d_2)$$
$$+ \nu(2) (1 - q^{(1)})(1 - q^{(2)})p^f(b_1, d_1)p^f(b_2, d_2)$$
$$= 4.5 \times 10^{-2} e^{-2((b_1 - 2)^2 + (d_1 - 4)^2)} e^{-2((b_1 - 1)^2 + (d_1 - 3)^2)}$$
$$+ 2.11 \times 10^{-4} \left[ e^{-2((b_1 - 2)^2 + (d_1 - 4)^2)} + e^{-2((b_1 - 1)^2 + (d_1 - 3)^2)} \right] p^f(b_2, d_2)$$
$$+ 7.39 \times 10^{-7} p^f(b_1, d_1)p^f(b_2, d_2).$$  

(4.20)

Notice that this local kernel also decomposes into terms which describe presence of upper features: one term for both, one term for each of the two upper features, and the last term has no upper features. Contour plots of slices of this local kernel are shown in Fig. 7. A general description of slicing is given in Rmk. 4.8.

**Remark 4.8.** Slices are used to view local pdfs defined on a high dimensional space $W^N \subset \mathbb{R}^{2N}$ for $N > 1$. To obtain these slices, one fixes features $(b_j, d_j) = (b_j', d_j')$ for $j = 2, \ldots, N$, and views the density on the corresponding hyperplane $W \times \{(b_1', d_1')\} \times \cdots \times \{(b_N', d_N')\} \subset W^N$. In practice, the fixed features are chosen as modes of earlier (smaller $N$) slices in order to view important parts of the distribution. We also sum over possible permutations in order to view a slice of the symmetric pdf, as was done for Ex. 2.

If we consider the density evaluated along slices as $K_\sigma(((b, d), (1, 3)); \mathcal{D})$ or $K_\sigma(((b, d), (2, 4)); \mathcal{D})$ (Fig. 7(a) or (b), respectively), the restricted plot is a Gaussian centered at the other upper feature. If the fixed
From Eq. (4.11), we obtain:

$$D_\xi \text{feature } \xi K$$

Figure 7: Contour maps for slices of the kernel density $p$ two upper Gaussians feature is instead close to the diagonal, as in Fig. 7 (c), the density slice is close to a mixture between the two upper Gaussians $p^{(1)}$ and $p^{(2)}$.

In a similar fashion, we also express the kernel density with input cardinality $|Z|=3$. Since there are only 2 upper features in $\mathcal{D}$, this and further expressions are not markedly more complicated than Eq. (4.20). From Eq. (4.11), we obtain:

$$K_\sigma((\xi_1, \xi_2, \xi_3), \mathcal{D}) = \nu(1) \left[ q^{(1)} q^{(2)} p^{(1)}(b_1, d_1) p^{(2)}(b_2, d_2) \right] p^\ell(b_3, d_3) + \nu(2)(1 - q^{(2)}) q^{(1)} p^{(1)}(b_1, d_1) p^\ell(b_2, d_2) p^\ell(b_3, d_3) + \nu(2)(1 - q^{(1)}) q^{(2)} p^{(2)}(b_1, d_1) p^\ell(b_2, d_2) p^\ell(b_3, d_3) + \nu(3)(1 - q^{(1)})(1 - q^{(2)}) p^\ell(b_1, d_1) p^\ell(b_2, d_2) p^\ell(b_3, d_3).$$

(4.21)

One may notice that Eq. (4.21) has the same 4 terms as Eq. (4.20), but with another factor of $p^\ell$ in each term. Indeed, the local kernels for input cardinality $N=4,5,6$ appear very similar as well, and with progressively more factors of $p^\ell$. Contour plot slices of this local kernel are shown in Fig. 8 following Rmk. 4.8. In this case, since the local pdf is defined in $W^3$, we must fix a pair of features in order to view a slice in $W \times \{(b'_2, d'_2)\} \times \{(b'_3, d'_3)\}$. In Eq. (4.21), the heaviest weighted term consists of both upper features' densities as well as the lower density $p^\ell(b_3, d_3)$. Indeed, Fig. 8(a) shows the slice $K_\sigma(((b, d), (1, 3), (2, 4)), \mathcal{D})$, which leaves both upper features fixed, and the resulting slice is nearly proportional to the lower density $p^\ell$. Fig. 8(b) shows the slice $K_\sigma(((b, d), (1, 3), (2, 5, 3, 5)), \mathcal{D})$, which fixes one of the upper features of $\mathcal{D}$ as well as a feature of moderate persistence. This slice does not go through a mode of the local kernel, and so the geometry of the dataspace $W^3/\Pi$ makes the slice look multi-modal, depending on whether $(2, 5, 3, 5)$ is assigned to $p^{(2)}$ or $p^\ell$. Other assignments have negligible mass. Thus, Fig. 8(b) resembles a mixture of these two densities.

The terms $(1 - q^{(k)})$ within the $Q^*$ expression (see Eq. (4.3)) are very small and appear in terms for which the corresponding upper feature is unassigned. These terms are so small because both upper features
Figure 8: Contour maps for slices of the kernel density $K_\sigma((\xi, \xi', \mathcal{Z}), \mathcal{Z})$ with input cardinality 3. A pair of features $\xi'_2$ and $\xi'_3$, indicated by white crosshairs, are fixed to restrict to a 2D subspace as follows: (a) $(\xi'_2, \xi'_3) = ((1, 3), (2, 4))$ and (b) $(\xi'_2, \xi'_3) = ((1, 3), (2, 5, 3, 5))$.

Since the symmetric version of the density is used, the order of these features is irrelevant. The center diagram is indicated by red (upper) and green (lower) points. Scale bars at the right of each plot indicate the range of probability density in each shaded region.

| KDE | (1) | (2) | (3) | (4) |
|-----|-----|-----|-----|-----|
| n   | 100 | 300 | 1000| 5000|
| $\sigma$ | 0.03 | 0.025 | 0.020 | 0.015 |

Table 1: Choices of sample size $n$ (number of persistence diagrams) and bandwidth $\sigma$ for each kernel density estimate $\hat{f}_{n,\sigma}(Z)$ shown in Fig. 10.

have very long persistence in this example (four times the bandwidth), and so the terms in Eqs. (4.19), (4.20), and (4.21) which do not include one or both upper Gaussians $p^{(1)}$ and $p^{(2)}$ have progressively smaller contribution to the overall local kernel. Consequently, the kernel places much higher probability density near input diagrams with features nearby each upper feature in the center diagram. This behavior is seen in Fig. 6, 7, 8, and their respective analyses, and is directly correlated to the ratio of persistence to bandwidth for each feature.

Example 3. Here we consider the random persistence diagram generated from a specific random dataset in $\mathbb{R}^2$. Our goal in this example is to build and demonstrate convergence of the kernel density estimate for the pdf of the associated random persistence diagram. Specifically, we generate sample datasets which each consist of 10 points sampled uniformly from the unit circle with additive Gaussian noise, $N((0, 0), (\frac{1}{50})^2 I_2)$. This toy dataset is prototypical for signal analysis (corresponding to the circular dynamics of a noisy sine curve), wherein the high dimensional point cloud is obtained through delay-embedding of the signal. An in-depth analysis of using delay embedding alongside persistent homology is found in (Perea and Harer, 2015).

These datasets each yield a Čech persistence diagram as described in Section 2 for degree of homology $k = 1$. A sample dataset and its associated $k = 1$ persistence diagram are shown in Fig. 9. Since these datasets are sampled from the unit circle perturbed by relatively small noise, one expects the associated 1-homology to have a single persistent feature with $d \approx 1$ with possible brief features caused by noise.

We consider several KDEs as we simultaneously increase the number of persistence diagrams ($n$) and narrow the bandwidth ($\sigma$) as shown in Table 1. The bandwidth was chosen to scale according to Silverman’s rule of thumb (Silverman, 1986) (see Rmk 4.7). Since the KDEs $\hat{f}_{n,\sigma}(Z)$ are defined on $\bigcup_N W^N$ for several input cardinalities $N$, we present them in multiple slices by fixing a cardinality and then fixing all but one input feature as described in Rmk 4.8.
Figure 9: An example underlying dataset and its associated persistence diagram. The persistence diagrams are used as the centers for the kernel density estimate. For this example, persistence diagrams with more than one feature are relatively rare.

For example, \( g(\xi) = \hat{f}_{n,\sigma}(\xi, \xi'_2, \ldots, \xi'_N) \) for fixed \( \xi'_j \ (j = 2, \ldots, N) \) is a function on \( W \) and represents a slice of the local KDE on \( W^N \). The progression of KDE slices can be seen in Fig. 10 wherein the same slices (i.e., the same features are fixed) are viewed for each choice of \((n, \sigma)\). These plots demonstrate in practice the convergence of the kernel density estimator shown in Theorem 1. Because the sample points for the underlying dataset lie so close to the unit circle, one expects the topological feature to die near scale \( d = 1 \), as is reflected in the KDEs shown in Fig. 10 (left); however, the distribution of points along the circle allows its birth scale to vary quite a lot. Additional features with brief persistence are concentrated very close to the diagonal due to small noise. These features tend to be either spurious holes near the edge (smaller \( b \) and \( d \)) or a short split of the main topological loop in two (larger \( b \) and \( d \)); this behavior is reflected in the two peaks for slices of the KDEs shown in Fig. 10 (right). Indeed, the persistence diagram shown in Fig. 9 is typical for this example. Overall, by scanning from top to bottom, Fig. 10 demonstrates the convergence of the KDEs as \( n \) increases and \( \sigma \) decreases. The location and mass of each mode is as expected from underlying data sampled from the unit circle. Moreover, very small spread in the limiting density arises from the small noise in the underlying data. The shape and spread of each mode converges, and the densities for \( n = 1000 \) and \( n = 5000 \) are nearly the same.

Two more examples of persistence diagram KDEs at increasing \( n \) and decreasing \( \sigma \) are given in the supplementary materials, but which involve more complex underlying data.

### 4.3 A Measure of Dispersion

Theorem 2 has established the convergence of a kernel density estimator. Along with density function estimation, one would like to verify the convergence of properties such as spread. In the absence of vector space structure on the space of persistence diagrams, we turn to the bottleneck metric (Defn. 2.8) to define a notion of spread.

Specifically, we measure dispersion with respect to a distribution of persistence diagrams through its mean absolute deviation in this metric.

**Definition 4.3.** The mean absolute bottleneck deviation (MAD) from origin diagram \( \mathcal{D} \) with respect to a global pdf \( f \) is given by

\[
MAD_f(\mathcal{D}) = \int_{W_{0,d-1}} W_\infty(\mathcal{D}, Z)f(Z)\delta Z
\]

The following proposition aids in proving convergence of MAD kernel estimates. Proofs for this section are delegated to the supplementary materials.
Figure 10: Plots of persistence diagram KDEs for Ex. 3. Each plot is presented as a heat map where color indicates the probability density. White regions above the diagonal indicate portions of very low probability density. Each column is a particular slice, while each row is a particular global KDE with $n$ and $\sigma$ as indicated in Table 1. The left column are the local KDEs $\hat{f}_{n,\sigma}((b,d))$ evaluated at a diagram with only one feature. The mode of the converged density is approximately $(b_2, d_2^*) = (0.77, 0.98)$. The right column are the local KDEs $f_{n,\sigma}((b, d), (0.77, 0.98))$ evaluated at a diagram with two features and one feature fixed. These slices have two modes which are very close to the diagonal at $(0, 0)$ and $(1, 1)$. Overall, this figure demonstrates KDE convergence.
Proposition 4.4. Consider $D$ distributed according to the kernel density $K_\sigma(\cdot, D)$ with center diagram $D$ and bandwidth $\sigma$. Fix $\delta \geq 1$. Then,

$$\mathbb{P}[W_\infty(D, D) < \delta \sigma] \geq \left( \int_{B(0, \delta)} \frac{1}{\sqrt{2\pi}} e^{-(x^2+y^2)/2} \, dx \, dy \right)^M$$

(4.23)

where $M$ is the maximal cardinality of $D$ (a multiple of $|D|$). Here $B(x, r)$ refers to a ball with respect to the infinity metric (as is used in bottleneck distance).

Next, we relax assumption (A2) by considering the entire multi-wedge $W_{0\text{-}1}$ and tighten the decay control from assumption (A3). Formally,

(A2)* The local density $f_N : W_{0\text{-}1} \to \mathbb{R}$ is bounded for each $N \in \{1, \ldots, M\}$.

(A3)* There exists $C > 0$ so that $f(\xi_1, \ldots, \xi_N) \leq C \|\sum_{i=1}^N |\xi_i\|^{-2N-2}$ for $N \in \{1, \ldots, M\}$.

These assumptions (and (A1)) are required for the subsequent lemma, which ensures that the mean absolute bottleneck deviation (MAD) is finite.

Lemma 4.5. Consider a random persistence diagram $D$ distributed according to a global pdf $f$ satisfying assumptions (A1), (A2)*, and (A3)*. Then $D$ has finite MAD for any choice of origin diagram $D$.

Similar to assumption (A3) (given prior to Thm. 2), (A3)* holds for a random persistence diagram associated with underlying data sampled from a compact set perturbed by Gaussian noise. One may also replace Lemma 4.5 and its assumptions by directly assuming that the maximal persistence moment is bounded; with this, the results of the lemma follow immediately from Eq. (4.3) in the supplementary. This direct assumption is weaker (implied by (A1), (A2)*, and (A3)*), but may be difficult to show directly in practice.

Theorem 3. Consider a distribution of persistence diagrams with bounded global pdf, $f$, satisfying assumptions (A1), (A2)*, and (A3)*. Let $f(Z) = \frac{1}{n} \sum_{i=1}^n K_\sigma(Z, D_i)$ be a kernel density estimate with centers $D_i$ sampled i.i.d. according to global pdf $f$ and bandwidth $\sigma = O(n^{-\alpha})$ chosen with $0 < \alpha < \alpha_{2M}$. Then, the mean absolute bottleneck deviation estimate converges; in other words,

$$\int_{W_{0\text{-}1}} W_\infty(D_0, Z) f(Z) \delta Z \to \int_{W_{0\text{-}1}} W_\infty(D_0, Z) f(Z) \delta Z$$

(4.24)

as $n \to \infty$ for any origin diagram $D_0$.

5 Discussion and Conclusions

A nonparametric approach to approximating density functions of finite random persistence diagrams has been presented. This includes the introduction of a kernel density function, as well as proof that the kernel density itself and its mean absolute deviation converge to those of the target distribution. Future work will investigate the convergence of powers of the absolute deviation (e.g., bottleneck variance) and deviations involving the Wasserstein metric (an $L^p$ generalization of bottleneck metric, see Edelsbrunner and Harer [2010]). Our framework is presented through the lens of geometric simplicial complexes, and in particular Čech complexes. The resulting persistence diagrams are based on underlying datasets in a metric space. In general, one may define persistent homology for a function $f$ defined on a topological space (Edelsbrunner and Harer [2010] and Adler et al. [2010]) and therefore random functions may also give rise to random persistence diagrams, see Adler et al. [2010] for an example. A similar kernel density estimate approach can be formulated in this case, but perhaps different assumptions may be needed on the target pdf.

Our approach is fully data-driven, a necessary step since distributions of persistence diagrams were previously poorly understood. The assumptions (A1)-(A3), (A2)*, and (A3)* are typical for kernel density estimators Scott [2015]. Similar assumptions on the underlying data are inherited by the random persistence
diagram, because variation in Čech persistent homology is controlled by interpoint distances. In particular, probability density decay follows the same trends as noise in the underlying data; this is seen in Fig. 10 (a) for Gaussian noise. Thus, the kernel density estimates defined here can be reliably used for data analysis, adding a detailed tool to the methods used in topological data analysis. In particular, this is the first result yielding probability density functions which directly analyze the full distribution information of a random persistence diagram. For applications in machine learning such as classification, the kernel density estimates carry information for generating more sophisticated features than previously available; e.g., the value of the global pdf at a specific input or list of inputs or the integral of the global pdf over a specified region. Access to a pdf also provides a tool with which one can check for classification robustness in terms of likelihood or Bayes factors, providing a measure of the confidence in a particular outcome.

Lending credence to applicability in data analysis, an example of kernel density estimation is presented in Subsection 4.2. In this example, underlying datasets are generated to lie on the unit circle with additive noise, a prototypical example for topological data analysis. Our analysis yields detailed information about the distribution of diagrams, even though only two 2-dimensional slices of the kernel density estimate are shown. This example demonstrates the convergence of the kernel density estimator in practice for large enough sample size (number of persistence diagrams). This example along with the supplementary examples also demonstrate the detailed information contained in a persistence diagram KDE.

In the context of Fig. 4, it is clear that sampling from the kernel density is straightforward, and in fact computation time scales linearly in the number of features in the center diagram $D$. In contrast, precise evaluation of the kernel global pdf at a diagram requires the more thorough computations shown in Eq. (4.11). This evaluation is made tractable due to the separation of the center diagram into upper and lower portions: $D = D^u \cup D^l$ as described in Eq. (4.6). In practice, diagrams should split so that $|D^u|$ is small while $|D^l|$ is large. Evaluation of individual feature pdfs on the multi-wedge $W_{0: d-1}$ only scales quadratically on the cardinality $|D|$ and higher degree calculations are required only for combinatorics on the large persistence features in the upper diagram $D^u$. Consequently, these calculations are tractable so long as $D^u$ does not grow too much in cardinality, while an increased cardinality for $D^l$ has a lesser effect on computation time.

The kernel density presented here treats the small persistent features in $D^l$ as a single group. Since convergence (Thm. 2) requires very little structure in the lower random diagram, it may be helpful in practice to cluster the lower portion of the center diagram, followed by defining a random diagram centered at each cluster. This approach somewhat complicates the expression and evaluation of the kernel density, but does not complicate sampling from the kernel density. The goal of this approach is to more carefully capture the geometric features of the underlying random dataset, since such geometric features often correspond to briefly persistent homological features. For example, geometric features are of paramount importance for classifying periodic signals through their delay embeddings, wherein the large persistent feature indicates periodicity and thus is expected to appear in every class.

References

Adcock, A., Carlsson, E., and Carlsson, G. (2016). The ring of algebraic functions on persistence bar codes. Homology, Homotopy and Applications, 18(1):381–402.

Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E., and Weinberger, S. (2010). Persistent homology for random fields and complexes. In Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown (pp. 124–143). Institute of Mathematical Sciences.

Adler, R. J., Bobrowski, O., and Weinberger, S. (2014). Crackle: The homology of noise. Discrete & Computational Geometry, 52(4):680–704.

Adler, R. J., Agami, S., and Pranav, P. (2017). Modeling and replicating statistical topology, and evidence for CMB non-homogeneity. arXiv:1704.08248.
Atienza, N., Gonzalez-Diaz, R., and Rucco, M. (2016). Separating topological noise from features using persistent entropy. In Federation of International Conferences on Software Technologies: Applications and Foundations (pp. 3–12). Springer International Publishing.

Bauer, U. (2015). Ripser. https://github.com/Ripser/ripser.

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. The Annals of Applied Statistics, 10(1):198–218.

Bobrowski, O., Mukherjee, S., and Taylor, J. E. (2014). Topological consistency via kernel estimation. arXiv:1407.5272.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015). Persistence images: An alternative persistent homology representation arXiv:1507.06217.

Chen, C. and Kerber, M. (2011). Persistent homology computation with a twist. In Proceedings 27th European Workshop on Computational Geometry, volume 11.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. Discrete Comput. Geom, 37:103–120.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010). Lipschitz functions have 1 p-stable persistence. Foundations of computational mathematics, 10(2):127–139.

De Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. Algebraic & Geometric Topology, 7(1):339–358.

Donato, I., Gori, M., Pettini, M., Petri, G., De Nigris, S., Franzosi, R., and Vaccarino, F. (2016). Persistent homology analysis of phase transitions. Physical Review E, 93(5), 052138.

Edelsbrunner, H. and Harer, J. (2010). Computational topology: an introduction. American Mathematical Society.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. Discrete and Computational Geometry, 28(4):511–533.

Edelsbrunner, H. (2013). Persistent homology in image processing. In International Workshop on Graph-Based Representations in Pattern Recognition, pages 182–183. Springer, Berlin, Heidelberg.

Emmett, K., Rosenbloom, D., Camara, P., and Rabadan, R. (2014). Parametric inference using persistence diagrams: A case study in population genetics. arXiv:1406.4582.

Emrani, S., Gentimis, T., and Krim, H. (2014). Persistent homology of delay embeddings and its application to wheeze detection. IEEE Signal Processing Letters, 21(4):459–463.

Fasy, B. T., Kim, J., Lecci, F., Maria, C., Rouvreau., V., The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, P. b. U. B. M. K., and Reininghaus., J. (2015). Tda: Statistical tools for topological data analysis r package version 1.4.1.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence sets for persistence diagrams. The Annals of Statistics, 42(6):2301–2339.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). Bayesian data analysis, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.
Goodman, I. R., Mahler, R. P., and Nguyen, H. T. (2013). Mathematics of data fusion, volume 37. Springer Science & Business Media.

Guillemand, M. and Iske, A. (2011). Signal filtering and persistent homology: an illustrative example. Proc. Sampling Theory and Applications (SampTA’11).

Hatcher, A. (2002). Algebraic topology. 2002. Cambridge UP, Cambridge, 606(9).

Kerber, M., Morozov, D., and Nigmetov, A. (2016). Geometry helps to compare persistence diagrams. Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments, pages 103–112.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016). Persistence weighted Gaussian kernel for topological data analysis. In Proceedings of the 33rd International Conference on Machine Learning.

Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. Statistical topological data analysis-a kernel perspective. In Advances in neural information processing systems, pages 3070–3078.

Mahler, R. P. (1995). Unified nonparametric data fusion. In SPIE’s 1995 Symposium on OE/Aerospace Sensing and Dual Use Photonics, pages 66–74. International Society for Optics and Photonics.

Marchese, A. and Maroulas, V. (2016). Topological learning for acoustic signal identification. In Information Fusion (FUSION), 2016 19th International Conference on, pages 1377–1381. ISIF.

Marchese, A. and Maroulas, V. (2017). Signal classification with a point process distance on the space of persistence diagrams. Advances in Data Analysis and Classification, Springer Berlin Heidelberg, https://doi.org/10.1007/s11634-017-0294-x.

Marchese, A., Maroulas, V., and Mike, J. (2017). K-means clustering on the space of persistence diagrams. In Wavelets and Sparsity XVII (Vol. 10394, p. 103940W). International Society for Optics and Photonics.

Matheron, G. (1975). Random Sets and Integral Geometry. John Wiley & Sons.

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). Probability measures on the space of persistence diagrams. Inverse Problems, 27(12).

Munch, E. (2017). A user’s guide to topological data analysis Journal of Learning Analytics, 4(2):47–61.

Perea, J. A. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics, 15(3):799–838.

Perea, C. M. and de Mello, R. F. (2015). Persistent homology for time series and spatial data clustering. Expert Systems with Applications, 42(15):6026–6038.

Reininghaus, J., Huber, S., Bauer, S., and Kwitt, R. (2014). A stable multi-scale kernel for topological machine learning. arXiv:1412.6821

Scott, D. W. (2015). Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Seversky, L. M., Davis, S., and Berger, M. (2016). On time-series topological data analysis: New data and opportunities. The IEEE Conference on Computer Vision and Pattern Recognition, pages 59–67.

Sgouralis, I., Nebenführ, A., and Maroulas, V. (2017). A Bayesian topological framework for the identification and reconstruction of subcellular motion. SIAM Journal on Imaging Sciences, 10(2):871–899.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press, New York.

26
TURNER, K., MILEYKO, Y., MUKHERJEE, S., AND HARER, J. (2014). Fréchet means for distribution of persistence diagrams. *Discrete & Computational Geometry*, 52:44–70.

VENKATARAMAN, V., RAMAMURTHY, K. N., AND TURAGA, P. (2016). Persistent homology of attractors for action recognition. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4150–4154.

XIA, K., FENG, X., TONG, Y., AND WEI, G. W. (2015). Persistent homology for the quantitative prediction of fullerene stability. *Journal of computational chemistry*, 36(6):408–422.
A Proof of Theorem 2

The proof presented in this section describes the case for degree of homology \( k > 0 \). The case for \( k = 0 \) is obtained by a slight modification and the full result follows by an application of Corollary 4.3.

Recall Thm. 1 which defines the pertinent kernel density \( K_\alpha(Z, \mathcal{D}) \) evaluated at \( Z = (\xi_1, ..., \xi_N) \) according to center diagram \( \mathcal{D} \) and bandwidth \( \sigma \) by

\[
K_\alpha(Z, \mathcal{D}) = \sum_{\gamma \in \mathcal{D}} \nu(N-j) \sum_{j \in (J, N_u)} Q(\gamma) \prod_{k=1}^{j} p^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p^{f}(\xi_k)
\]

where \( Q(\gamma) \) is given by Eq. (4.2), each \( p^{(j)} \) refers to the modified Gaussian pdf shown in Eq. (4.7) for its matching feature \( \xi_j \) in \( D^u \), \( N_u = |\mathcal{D}^u| \), and \( p^{f} \) is given by Eq. (4.9). Also recall that \( \mathcal{D} \) is split into \( \mathcal{D}^{f} \) and \( \mathcal{D}^{u} \) according to Eq. (4.6). \( D^{f} \) is defined with global pdf from Eq. (4.10), and \( D^{u} \) is defined with global pdf from Eq. (4.11).

Throughout the proof we use \( \xi_i \) to denote input features and \( Z = \{\xi_1, ..., \xi_N\} \) or \( Z = (\xi_1, ..., \xi_N) \) to denote an input persistence diagram as a set or vector of features. Several preliminary lemmas are presented before the main body of the proof. We begin with a critical lemma which controls the number of features sampled in the band diagonal \( \Delta^u_{\alpha} = \{(b, d) \in \mathbb{W} : \alpha < d - b < \beta\} \).

**Lemma A.1.** Consider a random persistence diagram \( D \) distributed according to \( f \) satisfying assumptions (A1)-(A3). Then there exists \( C > 0 \) so that \( \mathbb{E}^{f}(|\Delta^u_{\alpha} \cap D|) \leq C\sigma \).

**Proof.** Consider a region \( A \subset \mathbb{W} \) and a counting function \( \kappa_A(Z) = |Z \cap A| \) such that \( \kappa_A(\{\xi_1, ..., \xi_N\}) = \sum_{i=1}^{N} 1_A(\xi_i) \). It is clear that this set function is well defined and measurable if \( A \) is measurable. Using set integration (Defn. 3.4),

\[
\mathbb{E}(|\Delta^u_{\alpha} \cap D|) = \int_{\mathbb{W}} \kappa_{\Delta^u_{\alpha}}(Z) f(Z) \delta Z = \sum_{N=0}^{M} \frac{N!}{N} \int_{W} \mathbb{1}_{\Delta^u_{\alpha}}(\xi_1) \left[ \int f(\xi_1, ..., \xi_N) d\xi_2 ... d\xi_N \right] d\xi_1 \quad (A.1)
\]

The expressions in Eq. (A.1) can be phrased in terms of the probability hypothesis density from Eq. (4.8), and for any choice of \( L > 0 \) are bounded by

\[
\int_{\Delta^u_{\alpha}} F_D(\xi) d\xi \leq \int_{0}^{L} \int_{y-\sigma}^{y} F_D(x, y) dx dy + \int_{L}^{\infty} \int_{y-\sigma}^{\infty} C_3 y^{-2} dx dy \leq LC_2 \sigma + 3C_3 \sigma / L = (LC_2 + C_3 / L) \sigma
\]

where assumptions (A2) and (A3) respectively yield the bounds \( C_2 \) and \( C_3 y^{-2} \) on the probability hypothesis density, \( F_D \).

Lemma A.1 yields control over the counting measure \( \nu_i \) defined in Defn. 4.2 and the coefficients \( Q^*_\alpha(\cdot) \) of Eq. (4.3) which respectively determine the distribution of lower and upper cardinalities for a persistence diagram sampled according to the kernel density \( K_\alpha(Z, \mathcal{D}_i) \).

**Corollary A.2.** Consider a random persistence diagram \( D \) distributed according to \( f \) satisfying assumptions (A1)-(A3). Take \( \nu \) to be the lower cardinality probability mass function for the kernel density \( K_\alpha(Z, D) \) shown in Eq. (4.11). Then, there exists \( C > 0 \) so that \( \mathbb{E}^{f}(\nu(j_0)) \leq C\sigma \) whenever \( j_0 \neq 0 \).

**Proof.** Since \( D \) is random with respect to \( f \), \( \nu \) is random with respect to \( f \) as well. Recall that \( \nu \) is defined so that \( \mathbb{E}^{f}(\nu) = |D^{f}| \) for distributed according to \( \nu \) and thus \( \mathbb{E}^{f}(\mathbb{E}^{f}(\nu)) \leq C\sigma \) for some \( C > 0 \) by Lemma A.1. Subsequently, the value \( \mathbb{E}^{f}(\nu(j_0)) \) is controlled by this double expectation so long as \( j_0 \neq 0 \). Indeed,

\[
\mathbb{E}(a) = \sum_{j=0}^{\infty} j \nu(j) = \sum_{j=1}^{\infty} j \nu(j) + \sum_{j=1}^{\infty} j \nu(j) \geq \sum_{j=1}^{\infty} \nu(j) \geq \nu(j_0)
\]

for any \( j_0 > 0 \) and \( \nu_i(j_0) = 0 \) for \( j_0 < 0 \) since it represents a cardinality distribution. \( \square \)
In the following lemma, the result of Lemma \textbf{A.1} is used to control the expressions $Q(\gamma)$ or $Q^*(\gamma)$, of Eq. (4.2) and Eq. (4.3) respectively, in the kernel density estimate.

\textbf{Lemma A.3.} Consider a random persistence diagram $D$ distributed according to $f$ satisfying assumptions (A1)-(A3). Take $Q$ of Eq. (4.2) and $Q^*$ of Eq. (4.3) to be the upper singleton probabilities for the kernel density $K_\sigma(Z, D)$ shown in Eq. (4.11). Then, there exists $C > 0$ so that $\mathbb{E}^f[Q(\gamma)] \leq \mathbb{E}^f[Q^*(\gamma)] \leq C \sigma$ for any $\gamma \in I(j, N)$ with $j < N$.

\textit{Proof.} Since every $q^{(k)} \in (0, 1)$, we have that $Q(\gamma) \leq Q^*(\gamma)$; and furthermore, since $\gamma \in I(j, N)$ are not onto when $j < N$, each product $Q^*$ is bounded by one of the terms of the $(1 - q^{(k)}_i)$ type. By construction, these terms depend monotonically upon a feature’s persistence, and the maximum (over all indices $j < N$ and functions $\gamma$) is tied to the least persistent feature of $\mathcal{D}^n_i$.

For a feature $(b, d)$ of persistence $p = d - b$, we define $q(p) := \int_{-p/(\sqrt{2}\sigma)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ in concordance with Eq. (4.8); or in terms of the error function $\Phi$, $q(p) = \frac{1}{2} (1 + \Phi \left( \frac{p}{\sqrt{2}\sigma} \right))$. Define the minimal persistence as $p_{\min}(Z) = \sup \{ p : |\Delta^{n}_0 \cap Z | = 0 \}$ which satisfies $p_{\min}(Z) \geq p$ if and only if $|\Delta^{n}_0 \cap Z | = 0$. In turn, we may bound $Q^*(\gamma) \leq (1 - q(p_{\min}(D)))$ independently of $\gamma$. By Lemma \textbf{A.1}, there is $C > 0$ such that $\mathbb{P}^f[|\Delta^{n}_0 \cap D | \neq 0] \leq \mathbb{E}^f[|\Delta^{n}_0 \cap D |] \leq C \sigma$, which controls the distribution of the minimal persistence.

In particular, $q(p) = \frac{1}{2\sqrt{\pi}} \sigma^{-2} e^{-p^2/4\sigma^2}$ by the fundamental theorem of calculus. The control of Lemma \textbf{A.1} and the fact that $p_{\min}(Z) \geq 0$ also allows us to utilize integration via the probability of sublevel sets. Take $g(p) = 1 - q(p)$ so that $\lim_{p \to \infty} q(p) = 0$. Specifically, since $Q^*(\gamma) \leq (1 - q(p_{\min}(D)))$, and using the fundamental theorem of calculus then Fubini’s theorem, we have:

$$
\mathbb{E}^f[Q^*(\gamma)] \leq \int_{W_{n-1}} g(p_{\min}(Z)) f(Z) \delta Z = \int_{\text{W}_{n-1}} \left( \int_{p_{\min}(Z)}^{\infty} g'(p)dp \right) f(Z) \delta Z \tag{A.2}
$$

We now further bound the expectation in Eq. (A.2). Replacing terms with their definitions and using the bound control from Lemma \textbf{A.1} we obtain:

$$
\mathbb{E}^f[Q^*(\gamma)] \leq \int_{0}^{\infty} \mathbb{P}^f(|\Delta^{n}_0 \cap D | \neq 0) \frac{1}{2\sigma \sqrt{\pi}} e^{-p^2/4\sigma^2} dp \leq \frac{C}{2\sigma \sqrt{\pi}} \int_{0}^{\infty} p e^{-(p/2\sigma)^2} dp = \frac{C}{2\sigma \sqrt{\pi}} \left[ -2\sigma^2 e^{-(p/2\sigma)^2} \right]_{p=0}^{\infty} = \frac{C}{\sqrt{\pi}} \sigma.
$$

\textbf{Proof of Theorem 2.}

For convenience, we denote the upper cardinalities by $N_i = |\mathcal{D}^n_i|$ and total cardinalities by $M_i = |\mathcal{D}_i|$ for the sample persistence diagrams. Denote the set of strictly increasing functions from $\{1, ..., j\}$ to $\{1, ..., N_i\}$ by $I(j, N_i)$. Here we use ‘id’ to denote the identity map, where $I(N_i, N_i) = \{\text{id}\}$. The proof is organized by splitting the kernel densities into several pieces and then controlling each piece separately.

First, we separate the kernel $K_\sigma(Z, \mathcal{D}_i)$, defined in Eq. (4.11), into three portions, $A_i$, $B_i$, and $C_i$,
according to the upper cardinality $j$:

$$K_\sigma(Z, \mathcal{D}_i) = \sum_{j=0}^{N_i} \nu_i(N - j) \sum_{\gamma \in I(j, N_i)} Q_\delta(\gamma) \prod_{k=1}^{j} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{(\xi_k)}$$

$$= \nu_i(N - N_i) \sum_{k=1}^{N_i} Q_\delta(\gamma) \prod_{k=1}^{j} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{(\xi_k)}$$

$$+ \sum_{j=0, j \neq N}^{N_i-1} \nu_i(N - j) \sum_{\gamma \in I(j, N_i)} Q_\delta(\gamma) \prod_{k=1}^{j} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{(\xi_k)}$$

$$+ \mathbb{1}_{\{n \not\in \gamma: n < N_i\}}(N) \nu_i(0) \sum_{\gamma \in I(N, N_i)} Q_\delta(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k)$$

$$= A_i + B_i + C_i,$$

where $A_i$ follows from $j = N_i$, $C_i$ follows from $j = N$ ($C_i = 0$ if $N_i \leq N$), and $B_i$ consists of all remaining terms.

The terms $B_i$ in Eq. (A.3) are controlled by the lower product $\left[\prod_{k=j+1}^{N} p_i^{(\xi_k)}\right]$. Since $(1 - q_i^{(j)}) \leq 1$ and $\nu_i(N - j) \leq 1$ for any choice of $\gamma$ and $j$, we have that $B_i$ is bounded above by

$$B_i \leq \sum_{j=0, j \neq N}^{N_i-1} \sum_{\gamma \in I(j, N_i)} \left[\prod_{k=1}^{j} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{(\xi_k)}\right].$$

(A.4)

The bounding sum of Eq. (A.4) consists of restricted $2N$-dimensional Gaussians, with the weights $q_i^{(j)}$ dominating the restriction rescaling in Eq. (4.7). Fix $\pi \in \Pi_N$ and $j \in \{0, ..., M - 1\} \setminus \{N\}$. Without loss of generality, we treat the case when the permutation $\pi$ is the identity. Since our ultimate goal is to control the kernel density estimate $\hat{f}$, consider the portion of $\sum_{i=1}^{n} \frac{1}{n} B_i$ for which the cardinalities $M_i = |\mathcal{D}_i|$ are fixed at level $M_i = m \in \{0, ..., M\}$. Now, $m = |\mathcal{D}_i| \geq N_i > j$, so there is some extension for every $\gamma$ within the sum, $\gamma^* \in \Pi_m$. Recall that this collection is random because each $\mathcal{D}_i$ is randomly distributed according to $\pi$, therefore we consider the expectation with respect to this randomness:

$$E^f \left[ \sum_{\{i: M_i = m\}} \prod_{k=1}^{M_i} q_i^{(\gamma^*(k))} p_i^{(\gamma^*(k))}(\xi_k) \right] \rightarrow f(\xi_1, ..., \xi_m),$$

for any point $(\xi_1, ..., \xi_m)$ as a $2m$-dimensional Gaussian kernel density estimate with a proper choice of $\sigma = O(n^{-\alpha})$ appropriate for $2M$ (and hence $2m$) dimensions [Scott 2015]. Integrating both sides against the extra coordinates, Assumptions (A2) and (A3) along with the dominated convergence theorem yield

$$E^f \left[ \sum_{\{i: M_i = m\}} \prod_{k=1}^{M_i} q_i^{(\gamma^*(k))} p_i^{(\gamma^*(k))}(\xi_k) \right] \rightarrow \int_{W_{m-j}} f(\xi_1, ..., \xi_m) d\xi_{j+1}...d\xi_m,$$

(A.5)

which is again bounded via (A2) and (A3). Of course, $|\{i: M_i = m\}| \leq n$, so taking Eq. (A.5) into account for every $m$ bounds the averaging sum of the upper product: $\frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{j} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k)$. Since the points $\xi_i$ are fixed, we focus on their minimal persistence $p_{\min} = \min_i (d_i - b_i)$. Thus,

$$p_i^{(\xi_i)} \leq \frac{1}{2\pi\sigma^2} e^{-(b-d)^2/4\sigma^2} \leq \frac{1}{2\pi\sigma^2} e^{-p_{\min}^2/4\sigma^2},$$

30
and subsequently,
\[
\left[ \prod_{k=j+1}^{N} p_i^f(\xi_k) \right] \leq \frac{1}{(2\pi \sigma^2)^N} e^{-N\bar{f}^2/4\sigma^2} \to 0, \tag{A.6}
\]
as \sigma \to 0, uniformly on any compact subset of \( W \) (or \( W_{0,\text{id}−1} \)). Altogether, Eqs. (A.5) and (A.6) guarantee that the term \( \sum_{i=1}^{n} \frac{1}{n}B_i \to 0 \) as \( n \to \infty \) in the kernel density estimation.

Next we focus on the terms \( A_i \) in Eq. (A.3). We split the sum \( \frac{1}{n} \sum_{i=1}^{n} A_i \) according to the cardinality of \( \mathcal{D}_i \). Specifically, separate \( A_i \) into the cases where \( M_i \neq N_i \) or \( M_i = N_i \). First consider the associated set of indices \( \{i : M_i \neq N_i\} \) and define the mismatch number \( \text{MM}(n) \) to be its cardinality. Critical to our argument, the mismatch number is random with respect to \( f \) because it is defined according to the features in \( \mathcal{D}_i \). We obtain the following mismatched term:
\[
\frac{1}{n} \sum_{\{i : N_i \neq M_i\}} A_i \leq \left( \frac{\text{MM}(n)}{n} \right) \sum_{\{i : N_i \neq M_i\}} \left[ Q_i(\text{id}) \prod_{k=1}^{N_i} p_i^{(k)}(\xi_k) \prod_{k=N_i+1}^{N} p_i^f(\xi_k) \right]. \tag{A.7}
\]
The bounding sum in Eq. (A.7) is split into pieces where \( M_i = m \) for each \( m \) between 0 and \( M \). Using the same strategy yielding Eq. (A.5), with \( \text{MM}(n) \) in place of \( n \), the sum of the upper product converges to layered integrals of \( f \) for each level \( m \) and each \( N_i < m \) by extending \( \gamma = \text{id} \). Using the same approach leading to Eq. (A.6), the lower product vanishes in the limit if \( N_i \neq N \), or is an empty product if \( N_i = N \); in either case, this factor is bounded. Now, according to Lemma A.1, \( \mathbb{P}^f(M_i \neq N_i) = \mathbb{P}^f(\mathcal{D}_i \cap \Delta^0_\sigma \neq \emptyset) \leq C_5 \sigma \); consequently, \( \mathbb{E}^f[\text{MM}(n)/n] \to 0 \) and the mismatch terms on left hand side of Eq. (A.7) follow.

Now consider the indices for which \( N_i = M_i \). In this case, \( \mathcal{D}_i^f \) are empty, \( \nu_i = \delta_0 \), and the only values which contribute to the sum are for \( N_i = N \). The remaining portion of the sum is bounded by
\[
\frac{1}{n} \mathbb{E}^f \sum_{\{i : N_i = M_i\}} A_i = \frac{1}{n} \mathbb{E}^f \left[ \sum_{\{i : N_i = M_i\}} Q_i(\text{id}) \prod_{k=1}^{N_i} p_i^{(k)}(\xi_k) \right]. \tag{A.8}
\]
As shown, the terms in Eq. (A.8) are restricted \( 2N \) dimensional Gaussians. It is known (Scott, 2015) that restricted Gaussian kernel density estimates like \( \prod_{k=1}^{N_i} q_i^{(k)}(\xi_k) \) converge (uniformly on compactly contained sets) to the true value of the chosen draws \( \mathcal{D}_i \) for a suitable choice of \( a \) in \( \sigma = O(n^{-a}) \) as restricted by \( N \leq M \). After correcting for the samples with \( N_i < M_i = N \), the samples \( \mathcal{D}_i \) are treated as random draws from \( f(D|D = N) \). Consequently, we may conclude that the target distribution associated with \( \prod_{k=1}^{N} q_i^{(k)}(\xi_k) \) is the rescaled \( f(\xi_1, ..., \xi_N) \), where \( f(N) := \mathbb{P}^f(|D| = N) \). This rescaling for the conditional pdf \( f(D|D = N) \) is necessary to reweight according to Prop. 3.1.

Application of classical kernel density estimate results require division by the cardinality of the draw, when in context \( n \) is generally larger than this cardinality. Thus, we must again consider the cases wherein \( N_i \neq M_i \). Consequently, we find that the expectation for the ratio between the true draw cardinality and \( n \) is given by \( \mathbb{P}^f(|D| = N) + O(\sigma) \) according to Lemma A.1. Indeed, this ratio converges to \( f(N) := \mathbb{P}^f(|D| = N) \). After this final correction, we have shown that \( \frac{1}{n} \sum_{i=1}^{n} A_i \) approach the true pdf \( f(\xi_1, ..., \xi_N) \).

Lastly, we need only to control the terms \( C_i \) from Eq. (A.3). We begin by bounding the probability mass functions \( \nu_i \) by 1 and considering only terms for which the characteristic function is nonzero:
\[
\frac{1}{n} \sum_{i=1}^{n} C_i = \frac{1}{n} \sum_{\{i : N_i < N\}} \nu_i(0) \sum_{\gamma \in \mathcal{I}(N,N_i)} Q_i(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k) \leq \frac{1}{n} \sum_{\{i : N_i < N\}} \sum_{\gamma \in \mathcal{I}(N,N_i)} Q_i(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k). \tag{A.9}
\]
Next, we restrict the lower bounding integral for each term of $\mathcal{Q}(\gamma)$ by $P$ mapping to the diagonal yields a bounding correspondence and the associated probability is bounded below $\mathcal{P}$.

Consider the sets $W$ in which the maximal cardinality is drawn. Given a fixed cardinality, the draws of $D_\delta\sigma$ bounds the bottleneck distance by $\mathcal{P}(D, \mathcal{D}) \leq 1$. For clarity, features in $D$ are denoted using $\zeta$ while features in $\mathcal{D}$ are denoted using $\xi$.

Putting together the limits of each portion built from $K_{\sigma}(Z, \mathcal{D}) = A_i + B_i + C_i$, the theorem follows.

**B Proofs from Section 4.3**

**B.1 Proof of Proposition 4.4**

Note that the lower bound integral is the probability for a pair $z = (x, y)$ of independent standard normal variables to lie in $B((0,0), \delta)$. In order to bound the bottleneck distance $W_\infty(D, \mathcal{D}) < \delta\sigma$, it is sufficient that each constituent feature does not stray too far from either its corresponding center or the diagonal (see Fig. 4 for reference). Specifically, we follow Defn. 2.8 to build a correspondence between $D$ and $\mathcal{D}$ so that the maximal distance undercuts $\delta\sigma$, and thus the (potentially smaller) bottleneck distance is also bounded by $\delta\sigma$. For clarity, features in $D$ are denoted using $\zeta$ while features in $\mathcal{D}$ are denoted using $\xi$.

Consider each feature $\xi^j \in \mathcal{D} = \mathcal{D} \cap \{d - b > \sigma\}$ and its associated random singleton diagram $D^j = \{\zeta^j\} + \emptyset$ as in Defn. 4.1. Assuming the disc neighborhood $B_j = B(\xi^j, \delta\sigma)$ is contained in the wedge $W = \{(b, d) \in \mathbb{R}^2 : d > b \geq 0\}$, the density of $z^j = \frac{\zeta^j - \xi^j}{\sigma}$ is a multiple ($> 1$) of the density of the Gaussian random variable $z \sim N((0,0), I_2)$ in the region where $z^j \in B_j$ (or equivalently $z^j \in B((0,0), \delta)$). Thus, we obtain $\mathbb{P}(z^j \in B(\xi^j, \delta\sigma)) \geq \mathbb{P}(|z| \leq \delta)$ for the probability that $\zeta^j$ can be mapped to $\xi^j$ in a bounding correspondence. If $B_j \not\subseteq W$, this probability is even higher because $\zeta^j$ can be mapped to the diagonal and thus the case $D^j = \emptyset$ is included.

Now take into account the features in $D^\ell$ and the associated random features $D^\ell$ as in Defn. 4.2. Although the features in $D^\ell$ are not necessarily independent, we may assume without loss of generality the worst case, in which the maximal cardinality is drawn. Given a fixed cardinality, the draws of $D^\ell$ are independent. Since any feature may be mapped to the diagonal in the bottleneck distance, a bounding correspondence can be obtained whenever the draws in $D^\ell$ and features in $\mathcal{D}^\ell$ are close enough to the diagonal (within $\delta\sigma$). Indeed, the features in $\mathcal{D}^\ell$ are by definition distance $\sigma \leq \delta\sigma$ from the diagonal. Restricting to $W$, the pdf for the draws of $D^\ell = \{(b_j, d_j)\}_{j=1}^{N_\ell}$ is given by $p^\ell(b, d) = \frac{1}{2\pi \delta\sigma^2} \sum_{j=1}^{N_\ell} \exp{\left(-\frac{1}{2\delta\sigma^2} \left((x - \frac{b_j + d_j}{2})^2 + (y - \frac{b_j + d_j}{2})^2\right)\right)}$.

Consider the sets $U_j = B(\left(\frac{b_j + d_j}{2}, \frac{b_j + d_j}{2}\right), \delta\sigma)$ and $U = \bigcup_{j=1}^{N_\ell} U_j$. For each lower feature $(b, d) \in D^\ell$, mapping to the diagonal yields a bounding correspondence and the associated probability is bounded below by $\mathbb{P}(d - b \leq \delta\sigma) = \int_{\Delta_0^\delta} p^\ell(x, y) \, dx \, dy \geq \int_{W \cap U} p^\ell(x, y) \, dx \, dy$ since $W \cap U \subseteq \Delta_0^\delta = \{(b, d) \in W : d - b \leq \delta\sigma\}$. Next, we restrict the lower bounding integral for each term of $p^\ell$ to its matching subset $U_j$ and change variables to attain the desired form:

\[
\int_{W \cap U} p^\ell(x, y) \, dx \, dy \geq \sum_{j=1}^{N_\ell} \int_{U_j} \frac{1}{2\pi \delta\sigma^2} \exp{\left(-\frac{1}{2\delta\sigma^2} \left((x - \frac{b_j + d_j}{2})^2 + (y - \frac{b_j + d_j}{2})^2\right)\right)} \, dx \, dy
\]

\[
= \int_{B((0,0),\delta)} \frac{1}{2\pi} \exp{\left(-\frac{(x^2 + y^2)}{2}\right)} \, dx \, dy.
\]

Overall, this argument shows that with probability at least $\mathbb{P}(|z| \leq \delta)^M$ there is a correspondence which bounds the bottleneck distance by $\delta\sigma$ and the result follows.
B.2 Proof of Lemma 4.5

Choose an arbitrary persistence diagram $\mathcal{D}$. Since bottleneck distance is defined according to the sup-norm (see Eq. (2.7)), the bottleneck distance to the null persistence diagram (i.e., without any features) is precisely half the maximal persistence. Thus, we begin by showing that the maximal persistence moment is finite. Taking $Z = \{\xi_1, \ldots, \xi_N\}$ with $\xi_i = (b_i, d_i, k_i)$, we have:

$$\int_{W_{\text{bd}}-1} \max(d_i - b_i) \delta Z \leq \int_{W_{\text{bd}}-1} \|Z\| f(Z) \delta Z$$

(B.1)

since $\max(d_i - b_i) \leq \max(\|(b_i, d_i)\|) \leq \|Z\|$. Consider a compact set $K \subset W_{\text{bd}}-1$ which contains a neighborhood of the origin. Given assumptions $(A2)^*$ and $(A3)^*$, Eq. (B.1) is bounded by the following finite expression.

$$\int_{W_{\text{bd}}-1} \|Z\| f(Z) \delta Z \leq \int_K C_2 \|Z\| \delta Z + \sum_{N=1}^{M} \int_{h_N^{-1}(h_N(K)^c)} C_3 \|Z\|^{-2N-1} d\xi_1 \ldots d\xi_N.$$  

(B.2)

Lastly, we take advantage of the Minkowski inequality, which holds trivially for set integration since it is a linear combination of Lebesgue integrals. Indeed, the MAD centered at $\mathcal{D}_0$ is bounded as follows.

$$\int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z \leq \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, \emptyset) f(Z) \delta Z + \int_{W_{\text{bd}}-1} W_\infty(\emptyset, Z) f(Z) \delta Z$$  

(B.3)

where $\emptyset$ represents the null persistence diagram and the distance to the null persistence diagram is precisely half the maximal persistence. Since $f$ integrates to 1, the first integral simplifies to the finite distance $W_\infty(\mathcal{D}_0, \emptyset)$, while the second integral is finite according to Eq. (B.2).

B.3 Proof of Theorem 3

The MAD of $f$ with origin $\mathcal{D}_0$ is finite by Lemma 4.5. To show convergence of the estimate, we begin by adding and subtracting the integral of the sample estimator for the MAD. Then, we split the sum into $n + 1$ terms via the triangle inequality to obtain

$$\left| \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z - \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) \hat{f}(Z) \delta Z \right|$$

$$\leq \left| \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z - \frac{1}{n} \sum_{i=1}^{n} \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, \mathcal{D}_i) K_\sigma(Z, \mathcal{D}_i) \delta Z \right|$$

(B.4)

$$+ \frac{1}{n} \sum_{i=1}^{n} \left| \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) K_\sigma(Z, \mathcal{D}_i) \delta Z - \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, \mathcal{D}_i) K_\sigma(Z, \mathcal{D}_i) \delta Z \right| .$$

The term of the upper bound in Eq. (B.4) trivially simplifies to obtain the sample estimator for the MAD:

$$\left| \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z - \frac{1}{n} \sum_{i=1}^{n} \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, \mathcal{D}_i) K_\sigma(Z, \mathcal{D}_i) \delta Z \right|$$

(B.5)

$$= \left| \int_{W_{\text{bd}}-1} W_\infty(\mathcal{D}_0, Z) f(Z) \delta Z - \frac{1}{n} \sum_{i=1}^{n} W_\infty(\mathcal{D}_0, \mathcal{D}_i) \right|. $$

The MAD sample estimator converges since the MAD is finite, and thus this term vanishes as $n \to \infty$. The remaining term of the upper bound in Eq. (B.4) is further bounded via the reverse triangle inequality;
specifically,
\[
\sum_{i=1}^{n} \frac{1}{n} \left| \int_{W_{0,d-1}} W_{\infty}(\mathcal{D}_0, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z - \int_{W_{0,d-1}} W_{\infty}(\mathcal{D}_0, \mathcal{D}_i) K_{\sigma}(Z, \mathcal{D}_i) \delta Z \right| \\
\leq \sum_{i=1}^{n} \frac{1}{n} \left| \int_{W_{0,d-1}} W_{\infty}(\mathcal{D}_0, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z \right|.
\]

(B.6)

Toward bounding Eq. (B.6), choose a threshold parameter \( a = O(\sigma^\beta) \) for some \( \beta \in (0, 1) \), so that \( a \rightarrow 0 \) but \( a/\sigma \rightarrow \infty \) in the sample size (and bandwidth) limit. Next, take \( A_i = \{ Z \in W : W_{\infty}(Z, \mathcal{D}_i) \leq a \} \) and split the integral between \( A_i \) and its complement as

\[
\int_{W_{0,d-1}} W_{\infty}(\mathcal{D}_i, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z = \int_{A_i} W_{\infty}(\mathcal{D}_i, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z + \int_{A_i^c} W_{\infty}(\mathcal{D}_i, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z.
\]

The integral over \( A_i \) is trivially bounded by \( a \). Integration over the complementary events is controlled via layered integration along with Prop. 4.4. For \( a/\sigma > 1 \), which occurs when \( n \) is large enough, we obtain

\[
\int_{A_i^c} W_{\infty}(\mathcal{D}_i, Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z = a \mathbb{P}[W_{\infty}(\mathcal{D}_i, Z) > a] + \int_{a}^{\infty} \mathbb{P}[W_{\infty}(\mathcal{D}_i, Z) > b] \, db \\
\leq a \left( \mathbb{P}[|z| > a/\sigma]^M \right) + \int_{a}^{\infty} \mathbb{P}[|z| > b/\sigma]^M \, db,
\]

(B.7)

where \( z = (x, y) \) is distributed as a pair of independent standard normals. We chose \( a/\sigma = O(\sigma^{\beta-1}) \rightarrow \infty \) and so \( \mathbb{P}(|z| < a/\sigma) \rightarrow 0 \) exponentially fast and the last term vanishes quickly as \( \sigma \rightarrow 0 \).

Indeed, let \( g(Z) = W_{\infty}(\mathcal{D}_i, Z) \), then by the fundamental theorem of calculus and Fubini’s theorem:

\[
\int_{A_i^c} g(Z) K_{\sigma}(Z, \mathcal{D}_i) \delta Z = \int_{\{Z : g(Z) > a\}} \left( \int_{0}^{g(Z)} \, db \right) K_{\sigma}(Z, \mathcal{D}_i) \delta Z \\
= \int_{0}^{\infty} \int_{\{Z : g(Z) > a \text{ and } g(Z) > b\}} \, db \\
= \int_{0}^{\infty} \mathbb{P}[g(Z) > \max \{a, b\}] \, db \\
= a \mathbb{P}[g(Z) > a] + \int_{a}^{\infty} \mathbb{P}[g(Z) > b] \, db.
\]

Applying Prop. 4.4 changes the probabilities on \( g(Z) \) to normal tail probabilities. Thus, both bounding terms in Eq. (B.4) converge to zero and thus the kernel estimate converges to the true mean absolute deviation.

C Extra Examples

Here we present two more examples of constructing a kernel density estimator (KDE) according to the kernel given in Eq. (4.11). In these examples, we view slices of the KDE at various sample sizes and bandwidths. In the first example, the underlying dataset consists of points sampled from a circle with relatively large noise, in contrast to Ex. 3 in Subsection 4.2. This example demonstrates how, despite the symmetry of the unit circle and Gaussian noise of the underlying data, the resulting persistence diagram KDE and eventually its limiting behavior lacks Gaussian structure. In the second example, the underlying dataset consists of points sampled from a pinched circle. The underlying dataset has only one loop, but the persistence diagrams typically have a feature of long persistence and another feature of moderate persistence. Both features are
Figure 11: (a) An example of the underlying datasets generated for Ex. 4. Each dataset consists of 25 points sampled uniformly on the unit circle which are then perturbed by i.i.d. Gaussian noise with variance \((1/6)^2 I_2\). (b) The persistence diagram associated to the Čech filtration of the dataset captured by the KDE, and are clearly separable into distinct features despite their adjacency. To keep the presentation relatively simple to interpret, the same slices will be presented for each KDE (see Rmk. 4.8). This allows one to track the convergence of the KDE as the sample size of persistence diagrams, \(n\), increases and the bandwidth, \(\sigma\), decreases.

**Example 4.** Consider random underlying datasets each consisting of 25 points sampled uniformly from the unit circle, which are then perturbed by Gaussian noise with variance \((1/6)^2 I_2\), and their associated Čech persistence diagrams for degree of homology \(k = 1\). An example dataset and its associated Čech persistence diagram for \(k = 1\) are shown in Fig. 11.

Since the underlying datasets are sampled from a perfect circle perturbed by large noise, one expects the associated 1-homology to have a single persistent feature with several smaller features caused by noise. We consider several KDEs as we simultaneously increase the number of persistence diagrams and narrow the bandwidth. The bandwidth was chosen to vary according to Silverman’s rule of thumb (Silverman, 1986). Since the KDEs are defined on \(\bigcup_N W^N\) for several input cardinalities \(N\), we present \(\hat{f}_{n,\sigma}(Z)\) in multiple slices by fixing a cardinality and then fixing all but one input feature, as explained in Rmk. 4.8. For example, \(g(\xi) = \hat{f}_{n,\sigma}(\xi, \xi_2, \ldots, \xi_N)\) for fixed \(\xi_j\) is a function on \(W\) and represents a slice of the local KDE on \(W^N\). This progression of KDEs can be seen in Fig. 12, wherein the same slices are viewed for each choice of \(n\) and \(\sigma\). Modes of each slice are used as fixed features in the slices of higher cardinality inputs; consequently, the presented slices capture portions of the KDE with high probability density.

Fig. 12 demonstrates slower convergence of the KDEs than in Ex. 3, which is expected due to larger noise. Though the tail behavior of the KDEs remains Gaussian in nature, the limiting density is not Gaussian. In fact, the KDEs \(\hat{f}(n, \sigma)\) are neither symmetric nor unimodal, even for a single input. Much like the kernel densities themselves, each KDE separates into upper and lower densities on \(W\); however, the lower density varies depending on which upper mode is fixed in \(\hat{f}(\xi, \xi_j)\).

While the underlying dataspace is the unit circle in both Ex. 3 and Ex. 4, the precise presentation of the underlying data affects the pdf of the associated random persistence diagram. Precisely, two primary parameters for the underlying dataset are involved: (i) the scale of Gaussian noise and (ii) the sample size of the underlying dataset. The persistence diagram (for degree of homology \(k = 1\)) associated with the ’true’
Figure 12: Key slices of persistence diagram KDEs for Ex. 4. Each column is a particular slice, while each row is a particular KDE: (1) \( n = 20 \) and \( \sigma = 0.05 \), (2) \( n = 100 \) and \( \sigma = 0.03 \), and (3) \( n = 300 \) and \( \sigma = 0.02 \). The first column are the local KDEs \( \hat{f}_{n,\sigma}((b,d)) \) evaluated at a diagram with only one feature. The second column are the local KDEs \( \hat{f}_{n,\sigma}((b,d), (0.4, 0.8)) \) evaluated at a diagram with two features but one feature fixed. The third column are the local KDEs \( \hat{f}_{n,\sigma}((b,d), (0.56, 0.8)) \) evaluated at a diagram with two features but with a different feature fixed. Overall, this figure demonstrates convergence of the KDE as the number of persistence diagrams increases and the bandwidth decreases. Indeed, the two modes on the left already stabilize after \( n = 300 \), and the spread is no longer determined by the kernel bandwidth.
Figure 13: (a) An example of the underlying datasets generated for Ex. 4. Each dataset consists of 100 points sampled uniformly (according to angle) on the two-lobed polar curve which are then perturbed by i.i.d. Gaussian noise with variance \((1/30)^2 I_2\). (b) The persistence diagram associated to the Čech filtration of the underlying dataset.

The unit circle is not random and has a single feature at \((b, d) = (0, 1)\). The random, discrete nature of these examples creates persistence diagrams which deviate from this ‘truth.’

As described for Ex. 3 with very little noise all the sample points lie close to the unit circle, and so the Čech complex becomes contractible at a radius \(r \approx 1\). Consequently, the death value of the main topological feature is near the ‘true’ value (e.g., the mode in Fig. 10 is \(d \approx 0.98 \approx 1\)). However, since we are working with discrete points, this feature does not appear immediately: the gaps in the circle need to be filled in (this is even true without noise). In Ex. 3, the sample size is only 10, so the birth value is typically much larger than the ‘true’ value (e.g., the mode in Fig. 10 is \(b < 0.77\)).

In comparison to Ex. 3, Ex. 4 has relatively more noise; this results in a random persistence diagram with smaller death values for the main feature (e.g., the mode in Fig. 12 is \(d = 0.8 < 0.98\)). It is evident from Fig. 12 that while the noise is additive on the underlying data, its precise effect on the random persistence diagram is nonlinear. Moreover, Ex. 4 has a larger sample size (25 as opposed to 10), resulting in more consistent and smaller birth times for the main feature (e.g., the mode in Fig. 12 is \(b = 0.4 < 0.77\)). In addition, larger noise and sample size both result in more features near the diagonal in Ex. 4 as compared to Ex. 3.

**Example 5.** While Ex. 4 demonstrates the effect of noise on a persistence diagram pdf, this example will look into the effect of geometry. Consider random underlying datasets each consisting of 100 points sampled from a two-lobed polar curve, which are then perturbed by Gaussian noise with variance \((1/30)^2 I_2\), and their associated Čech persistence diagrams for degree of homology \(k = 1\). An example dataset and its associated Čech persistence diagram for \(k = 1\) are shown in Fig. 13.

We consider several KDEs as we simultaneously increase the number of persistence diagrams and narrow the bandwidth. The bandwidth was chosen to vary according to Silverman’s rule of thumb (Silverman, 1986). Since the KDEs are defined on \(\bigcup_N W^N\) for several input cardinalities \(N\), we present \(f_{n, \sigma}(Z)\) in multiple slices by fixing a cardinality and then fixing all but one input feature, as explained in Rmk. 4.8. For example, \(g(\xi) = f_{n, \sigma}(\xi, \xi_2, ..., \xi_N)\) for fixed \(\xi_j’\) is a function on \(W\) and represents a slice of the local KDE on \(W^N\). This progression of KDEs can be seen in Fig. 14, wherein the same slices are viewed for each choice of \(n\) and \(\sigma\) for several input cardinalities \(N\).
Modes of each slice are used as fixed features in the slices of higher cardinality inputs; consequently, the presented slices capture portions of the KDE with high probability density. Moreover, Fig. 14 demonstrates that these slices tend to capture specific topological or geometric features of the underlying dataspace.

The two-lobed curve in this example has a Čech persistence diagram consisting of two features, a topological feature of very long persistence and a geometric feature of moderate persistence. The moderate persistence feature describes the pinching of the curve. These two features are captured as separate points by the KDEs, and are thus viewed in completely separate slices of the KDE. By observing the KDE in the last row of Fig. 14, the geometric feature with moderate persistence has considerably less variance. Indeed, while the birth time of the topological feature relies on bridging gaps around the entire shape, which can all vary, the larger birth time of the geometric feature has less variance since it relies solely only on the short circuit between the lobes. As a result of this small variance, the geometric feature is emphasized for the local KDEs with a single input feature; also, the density takes longer to converge near this feature.

The lower portion of the KDE shows three separate modes. Features which build the largest mode consists of small loops, caused by local noise and gaps along the curve. The two modes which appear at larger scale indicate short circuiting of the pinch (smaller) or one of the lobes (larger, like the second mode in the circle example); These two lower modes are separate from noise-based features and are indicative of geometry in the underlying data.
Figure 14: Key slices of persistence diagram KDEs for Ex. Each column is a particular slice, while each row is a particular KDE (1) $n = 20$ and $\sigma = 0.03$, (2) $n = 100$ and $\sigma = 0.02$, (3) $n = 300$ and $\sigma = 0.015$, and (4) $n = 1000$ and $\sigma = 0.01$. The first column are the local KDEs $\hat{f}_{n,\sigma}((b,d))$ evaluated at a diagram with only one feature; the mode is $\xi'_1 = (0.2, 0.4)$. The second column are the local KDEs $\hat{f}_{n,\sigma}((b,d), (0.2, 0.4))$ evaluated at a diagram with two features, but with one feature fixed; the mode is $\xi'_2 = (0.14, 0.42)$ The third column are the local KDEs $\hat{f}_{n,\sigma}((b,d), (0.2, 0.4), (0.14, 0.42))$ evaluated at a diagram with three features, but with two features fixed. The fourth column shows the same slices as the third, but with the colormap shifted down to show the smaller modes. The variance of certain features effects the rate of convergence nearby, similar to Gaussian KDE in Euclidean space for a distribution with modes of different variance.