Journal of Machine Learning Research 20 (2019) 1-49 Submitted 9/18; Revised 4/19; Published 7/19

Nonparametric Estimation of Probability Density Functions of Random Persistence Diagrams

Vasileios Maroulas VMAROULA@UTK.EDU
Department of Mathematics
University of Tennessee
Knoxville, TN 37996, USA

Joshua L Mike MIKEJOSH@MSU.EDU
Computational Mathematics, Science, and Engineering Department
Michigan State University
East Lansing, MI 48823, USA

Christopher Oballe COBALLE@VOLS.UTK.EDU
Department of Mathematics
University of Tennessee
Knoxville, TN 37996, USA

Editor: Boaz Nadler

Abstract

Topological data analysis refers to a broad set of techniques that are used to make inferences about the shape of data. A popular topological summary is the persistence diagram. Through the language of random sets, we describe a notion of global probability density function for persistence diagrams that fully characterizes their behavior and in part provides a noise likelihood model. Our approach encapsulates the number of topological features and considers the appearance or disappearance of those near the diagonal in a stable fashion. In particular, the structure of our kernel individually tracks long persistence features, while considering those 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 as well as for virtual electroencephalographic data related to cognition.

Keywords: Topological Data Analysis; Persistence Homology; Finite Set Statistics; Global Distribution of Persistence Diagrams; Kernel Density Estimation; EEG Signals

1. Introduction

Topological data analysis (TDA) encapsulates a range of data analysis methods that 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, 2018) 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 $\{x_i\} \subset \mathbb{R}^{\mathrm{d}}$, which focuses on the 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 $\{x_i\} \subset \mathbb{R}^{\mathrm{d}}$ guarantees $k_i \in \{0, ..., \mathrm{d}-1\}$. 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 efforts 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 (Fasy et al., 2015) and Ripser (Bauer, 2015) 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, 2018), birth and death estimates (Emmett et al., 2014), and confidence sets (Fasy et al., 2014). While the aforementioned studies focus on valuable specific summaries, here we view a distribution of persistence diagrams in a global sense via a nonparametric method to estimate its density function.

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; Guillemard and Iske, 2011; Marchese and Maroulas, 2016; Perea 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 diagrams are created through a highly 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, results for the persistent homology of noise alone, such as (Adler et al., 2014), primarily concern the asymptotics of feature cardinality at coarse scale. More recently, it was proved that under mild conditions there exists a density with respect to Lebesgue measure for the intensity of persistence diagrams induced by filter functions defined on random variables having support on manifolds; in addition it was shown that kernel density estimation of this intensity is possible through persistence surfaces (Chazal and Divol, 2018). Another approach is to study these distributions through non-

parametric means. Kernel density estimation is a well known nonparametric technique for random vectors in $\mathbb{R}^d$ (Scott, 2015) which relies on convolution with a smooth kernel.

There has been extensive work to devise various maps from persistence diagrams into Hilbert spaces, especially Reproducing Kernel Hilbert Spaces (RKHS). For example, (Adams et al., 2017) discretizes persistence diagrams via bins, yielding vectors in a high dimensional Euclidean space; the authors in (Chazal and Divol, 2018) introduce a novel bandwidth selection procedure for this particular vectorization based on cross-validation. The persistence landscapes of (Bubenik, 2015) reinterpret the space of persistence diagrams within a function-based vector 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) uses Gibbs distributions in order to replicate similar persistence diagrams, e.g. for use in MCMC type sampling.

By mapping into a Hilbert space, the preceding approaches kernelize persistence diagrams for express application in typical machine learning methodology. 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 complementary fashion, our work considers a new kernel density defined on a completely different space: the space of persistence diagrams.

Since persistence diagrams lack a vector space structure, we must smooth without convolution. Instead, we treat persistence diagrams as random multisets and provide a noise model for persistence diagrams which ascribes a great level of uncertainty near the diagonal. The theory of random sets is an interpretation of point processes which provides the tools necessary to describe the global pdf of the noise model as a kernel density centered at a particular persistence diagram. 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 persistence diagram pdf facilitates definition and application of new statistical techniques in this context such as hypothesis testing, utilization of Bayesian priors, or likelihood based methods, e.g. see (Maroulas et al., 2019). 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 (Bendich et al., 2016) 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; For example, this problem can be approached via the studies (Bobrowski et al., 2014) and (Fasy et al., 2014).

The overall *contribution* of this article is listed below.

1. A probabilistic framework for defining distributions of random persistence diagrams.

2. A novel kernel density estimator centered at persistence diagrams, which in part provides a noise likelihood model, and its convergence to the true distribution as the number of persistence diagrams goes to infinity.

3. A new dispersion statistic for persistence diagrams, the mean absolute bottleneck deviation (MAD) and the convergence of the sample MAD computed with our kernel density estimator.

4. Applications of the kernel density estimator to analyze data data arising from a neuroscience problem related to cognition.

Organization: We establish the kernel density estimation problem through the lens of finite set statistics and we consequently begin with relevant background in topological data analysis in Section 2. The reader may refer to (Edelsbrunner and Harer, 2010) for a more rigorous treatment of the subject. Section 3 contains our framework for random persistence diagrams and their probability distributions. Our main theoretical results are 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. Convergence of this estimator to the true distribution of random persistence diagrams is proved in Subsection 4.2. In Subsection 4.3, we introduce a new measure of dispersion for persistence diagrams, the MAD, and show that the sample MAD computed with our kernel converges (with mild assumptions). Next, Section 5 contains practical examples of our kernel density. Namely, an example of persistence diagram kernel density estimation and its convergence are demonstrated for persistence diagrams associated to underlying data with annular distribution and we apply our kernel to a neuroscience

problem (Example 4). Finally, we end with conclusions and discussion in Section 6. Further examples of KDE convergence and the proofs of auxiliary propositions and lemmas 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. The reader should refer to (Edelsbrunner and Harer, 2010), for example, for a rigorous treatment of persistent homology. We first consider topological spaces of discernible dimension, called manifolds.

Definition 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 *A k-simplex is a collection of $k+1$ linearly independent vertices along with all convex combinations of these vertices: $(v_0, ..., v_k) = \left\{\sum_{i=0}^{k} \alpha_i v_i : \sum_{i=0}^{k} \alpha_i = 1 \text{ and } \alpha_i \geq 0 \, \forall i\right\}$. 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 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 $\mathcal{K}$.*

A simplicial complex is realized by the union of all its simplices; some examples are shown in Fig. 1. Conditions (i) and (ii) in Def. 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 $\mathbb{R}^{\mathrm{d}}$, this topology is also consistent with the Euclidean subspace topology.

Given a simplicial complex, we are interested in describing its global topology and local geometry through homological features. For our purposes, it suffices to define $k-$dimensional homological features of simplicial complexes as $k-$dimensional holes, so that, for example, $0-$dimensional homological features are connected components, $1-$dimensional homological features are loops, and $2-$dimensional homological features are voids.

We wish to extend the notion of homology for a discrete set of data $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$ 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 use 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 $\mathcal{U}_r = \cup_i B(x_i, r)$. The filtration of sets $\{\mathcal{U}_r\}_{r \in \mathbb{R}^+}$ naturally yields information about the arrangement within X of the dataset $\boldsymbol{x}$ at various

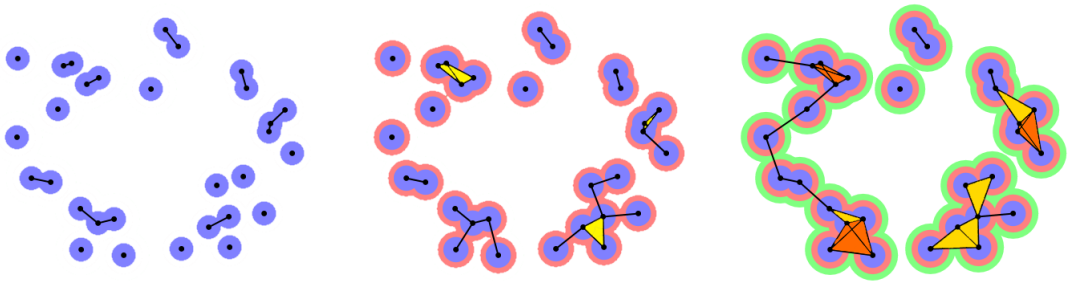


Figure 1: The neighborhood space and Čech complex of matching radius plotted at three different radii. Yellow indicates a triangle while orange indicates a tetrahedron. This family of simplicial complexes is the filtration used to compute and define persistent homology.

scales. To make homology computations more tractable for $\mathcal{U}_r$, we instead consider the associated nerve complexes.

Definition 4 *The nerve $\mathcal{N}(U)$ of a collection of open sets U is the simplicial complex where a k-simplex $(v_{i_0}, ..., v_{i_k})$ is in $\mathcal{N}(U)$ if and only if $\cap_{j=0}^{k} U_{i_j} \neq \emptyset$. 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 Čech$(\boldsymbol{x}, r)$.*

Examples of the Čech complex for the same data at different radii are depicted in Fig. 1, 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 $\mathcal{U}$ 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. 1). These inclusion maps induce maps at the level of homology groups. These induced maps are referred to here as the persistence maps, and take features to features 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 $\mathscr{D} = \{\xi_i\}_{i=1}^{M} = \{(b_i, d_i, k_i)\}_{i=1}^{M}$. 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 $\mathbb{R}^{\mathrm{d}}$, we consider each feature as an element of

$$\mathcal{W}_{0:\mathrm{d}-1} = W \times \{0, ..., \mathrm{d}-1\}, \tag{2.1}$$

where $W = \left\{(b,d) \in \mathbb{R}^2 : d > b \geq 0\right\}$ is the infinite wedge. As a topological space, the d-fold multiwedge $\mathcal{W}_{0:\mathrm{d}-1}$ 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 (Def. 5) is used to compare their associated persistence diagrams (Fasy et al., 2014; Munch, 2017). A distance that accounts for cardinality differences between persistence diagrams was introduced in (Marchese and Maroulas, 2018) and its stability with respect to perturbations in the underlying point cloud was proved in (Maroulas et al., 2018).

Definition 5 *The bottleneck distance between two persistence diagrams D_1 and D_2 is given by*

$$W_\infty(D_1, D_2) = \min_{\gamma} \max_{x \in D_1} \|x - \gamma(x)\|_\infty . \tag{2.2}$$

where γ 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 6 *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 $D = \{\xi_i\} \subset \mathcal{W}_{0:\mathrm{d}-1}$ in the multiwedge defined in Eq. (2.1). For underlying datasets sampled from $\mathbb{R}^{\mathrm{d}}$ 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 $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathrm{d}-1}) = \{D \text{ multiset in } \mathcal{W}_{0:\mathrm{d}-1} : |D| \leq M\}$. We view $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathrm{d}-1})$ through a list of functions h_N which each map the appropriate dimension of Euclidean space into its corresponding cardinality component, $\mathcal{C}_N(\mathcal{W}_{0:\mathrm{d}-1})$. This viewpoint facilitates the definition of probability densities.

Definition 7 *For each $N \in \{0, ..., M\}$, consider the space of N topological features, denoted $\mathcal{C}_N(\mathcal{W}_{0:\mathrm{d}-1}) = \{D \text{ multiset in } \mathcal{W}_{0:\mathrm{d}-1} : |D| = N\}$, and the associated map $h_N : \mathcal{W}_{0:\mathrm{d}-1}^N \to \mathcal{C}_N(\mathcal{W}_{0:\mathrm{d}-1})$ defined by*

$$h_N(\xi_1, ..., \xi_N) = \{\xi_1, ..., \xi_N\} . \tag{3.1}$$

The map h_N creates equivalence classes on $\mathcal{W}^N_{0:\mathbb{d}-1}$ according to the action of the permutations Π_N; specifically, $[Z] = [(\xi_1, ..., \xi_N)]_{h_N} = \left\{\left(\xi_{\pi(1)}, ..., \xi_{\pi(N)}\right) : \pi \in \Pi_N\right\}$ for each $Z = (\xi_1, ..., \xi_N) \in \mathcal{W}^N_{0:\mathbb{d}-1}$. These equivalence classes yield the space

$$\mathcal{W}^N_{0:\mathbb{d}-1}/\Pi_N = \left\{[\boldsymbol{\xi}]_{h_N} : \boldsymbol{\xi} \in \mathcal{W}^N_{0:\mathbb{d}-1}\right\}, \tag{3.2}$$

equipped with the quotient topology. The topology on $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1})$ is defined so that each h_N lifts to a homeomorphism between $\mathcal{W}^N_{0:\mathbb{d}-1}/\Pi_N$ and $\mathcal{C}_N(\mathcal{W}_{0:\mathbb{d}-1})$, and we write $\mathcal{W}^N_{0:\mathbb{d}-1}/\Pi_N \cong \mathcal{C}_N(\mathcal{W}_{0:\mathbb{d}-1})$.

With a topology in hand, one can define probability measures on the associated Borel σ-algebra. Thus, we define a random persistence diagram D to be a random element distributed according to some probability measure on $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1})$ for a fixed maximal cardinality $M \in \mathbb{N}$. We denote associated probabilities by $\mathbb{P}[\cdot]$ and expected values by $\mathbb{E}[\cdot]$. Since $\mathcal{W}^N_{0:\mathbb{d}-1}/\Pi_N \cong \mathcal{C}_N(\mathcal{W}_{0:\mathbb{d}-1})$, we work toward defining probability densities on the collection of Euclidean spaces $\cup_{N=0}^M \mathcal{W}^N_{0:\mathbb{d}-1}$.

Definition 8 *For a given random persistence diagram D and any Borel subset A of $\mathcal{W}_{0:\mathbb{d}-1}$, the belief function β_D is defined as*

$$\beta_D(A) = \mathbb{P}[D \subset A]. \tag{3.3}$$

Since A is a Borel subset of $\mathcal{W}_{0:\mathbb{d}-1}$, the collection $O_A = \{D \in \mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1}) : D \subset A\}$ is the quotient of $\cup_{N=0}^M A^N \subset \cup_{N=0}^M \mathcal{W}^N_{0:\mathbb{d}-1}$ under h_N; moreover, A^N is clearly Borel in the Euclidean topology of $\cup_{N=0}^M \mathcal{W}^N_{0:\mathbb{d}-1}$. Therefore, since h_N induces a homeomorphism (see definition 7), O_A is a Borel subset of $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1})$. 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-Nikodým type derivatives.

Definition 9 *Fix ϕ defined on Borel subsets of $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1})$ into $\mathbb{R}$. For an element $\xi \in \mathcal{W}_{0:\mathbb{d}-1}$ or a multiset $Z \subset \mathcal{W}_{0:\mathbb{d}-1}$ with $Z = \{\xi_1, ..., \xi_N\}$, the set derivative (evaluated at the empty set $\emptyset$) is respectively given by*

$$\begin{aligned} \frac{\delta\phi}{\delta\xi}(\emptyset) &= \lim_{n\to\infty} \frac{\phi(B(\xi, 1/n))}{\lambda(B(\xi, 1/n))}, \\ \frac{\delta\phi}{\delta Z}(\emptyset) &= \frac{\delta^N\phi}{\delta\xi_1 ... \delta\xi_N} = \left[\frac{\delta}{\delta\xi_1} \cdots \frac{\delta}{\delta\xi_N}\phi\right](\emptyset), \end{aligned} \tag{3.4}$$

where $B(\xi, 1/n)$ are Euclidean balls and λ indicates Lebesgue measure on $\mathcal{W}_{0:\mathbb{d}-1}$.

Remark 10 *Def. 9 for set derivatives at the empty set closely mirrors the Radon-Nikodým 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 11 *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 Π_N. The viewpoint of ϕ_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 $\emptyset$) 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 12 *Consider a Borel subset A of $\mathcal{W}_{0:\mathbb{d}-1}$ and a Borel subset O of $\mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1})$. For a set function $f : \mathcal{C}_{\leq M}(\mathcal{W}_{0:\mathbb{d}-1}) \to \mathbb{R}$, its set integrals over A and O are respectively defined according to the following sums of Lebesgue integrals:*

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

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

where $Z = \{\xi_1, ..., \xi_N\} \subset \mathcal{W}_{0:\mathbb{d}-1}$ is a persistence diagram.

Dividing by $N!$ in Eqs. (3.5a) and (3.5b) accounts for integrating over $\mathcal{W}_{0:\mathbb{d}-1}^N$ instead of $\mathcal{W}_{0:\mathbb{d}-1}^N / \Pi_N \cong \mathcal{C}_N(\mathcal{W}_{0:\mathbb{d}-1})$. 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

$$\beta_D(A) = \int_A \frac{\delta \beta_D}{\delta Z}(\emptyset)\delta Z. \tag{3.6}$$

Indeed, $A^N = h_N^{-1}(\{D \subset A\})$ so that Eq. (3.5a) also holds as an integral over the set $O_A = \{D \in \mathcal{C}_{\leq M} : D \subset A\}$ in the sense of Eq. (3.5b).

Definition 13 *For a random persistence diagram D, a global probability density function (global pdf) $f_D : \cup_{N \in \mathbb{N}} \mathcal{W}_{0:\mathbb{d}-1}^N \to \mathbb{R}$ must satisfy*

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

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

Remark 14 *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 $\mathcal{W}_{0:\mathbb{d}-1}^N$, the global pdf f_D is defined on $\cup_{N=1}^M \mathcal{W}_{0:\mathbb{d}-1}^N$ and restricts to a local density on each input dimension. Each local density $f_N(Z) = f_D\big|_{\mathcal{W}_{0:\mathbb{d}-1}^N}(Z)$ decomposes into the product of the conditional density $f_D(Z\big||Z| = N)$ and the cardinality probability $\mathbb{P}[|Z| = N]$ (this follows from Proposition 16). 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: $\int_{A^N} f_D(\xi_1, ..., \xi_N) d\xi_1 ... d\xi_N = \frac{1}{N!} \int_{A^N} \frac{\delta^N \beta_D}{\delta^N Z}(\emptyset) d\xi_1 ... d\xi_N$.*

Remark 15 *While the global pdf and its local constituents need not be symmetric with respect to Π_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 Π_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., $\left|D^j\right| \leq 1$), (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 16 *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 β_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 $q_m(S) = \mathbb{P}[D \subset S \big| \, |D| = m]$.*

Remark 17 *The decomposition in Proposition 16 is often applied as a first step toward finding the local density constituents of the global pdf. In particular, $f_N = f_D\big|_{\mathcal{W}^N_{0:\mathrm{d}-1}} = 0$ 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 $\mathcal{W}_{0:\mathrm{d}-1}$ gives the expected number of points in the region U; moreover, any other function on $\mathcal{W}_{0:\mathrm{d}-1}$ with this property is a.e. equal to the PHD (Goodman et al., 2013).

Definition 18 *(Matheron, 1975) The probability hypothesis density (PHD) for a random persistence diagram D is defined as the set function $F_D(a) = \frac{\delta \beta_D}{\delta Z}(\{a\})$ and is expressed as a set integral as*

$$F_D(a) = \int_{\{Z : \{a\} \subset Z\}} \frac{\delta \beta}{\delta Z}(\emptyset) \delta Z. \tag{3.8}$$

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

Remark 19 *Def. 18 is equivalent to an intensity function of a point process. In general, the intensity function induced by a given global pdf may be undefinied, but under mild conditions Eq. (3.8) is finite (we discuss this further in Section 4.2). Since D is a random persistence diagram, the PHD is always defined as a distribution and can always be integrated to obtain the identity $\mathbb{E}(|D \cap U|) = \int_U F_D(u)\, du$ for any region U.*

Proposition 16 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 20 *Consider a multiset of independent singleton random persistence diagrams $\left\{D^j\right\}_{j=1}^{M}$. 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 $\left|D^j\right| = 1$, then the global pdf for $D = \cup_{j=1}^{M} D^j$ is given by*

$$f_D(\xi_1, ..., \xi_N) = \sum_{\gamma \in I(N,M)} \mathcal{Q}(\gamma) \prod_{k=1}^{N} p^{(\gamma(k))}(\xi_k), \tag{3.9}$$

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

$$\mathcal{Q}(\gamma) = \mathcal{Q}^*(\gamma) \prod_{k=1}^{N} q^{(\gamma(k))}, \tag{3.10}$$

$I(N, M)$ consists of all (strictly) increasing injections $\gamma : \{1, ..., N\} \to \{1, ..., M\}$, which enumerate (unordered) correspondences between the input features $(\xi_1, \ldots, \xi_N)$ and a subset of the M random singletons, and

$$\mathcal{Q}^*(\gamma) = \frac{\prod_{j=1}^{M}(1 - q^{(j)})}{\prod_{k=1}^{N}(1 - q^{(\gamma(k))})}. \tag{3.11}$$

Proof Since the singleton events D^j are independent, the belief function for $D = \cup_j D^j$ decomposes into $\beta_D(S) = \prod_{j=1}^{M} \beta_{D^j}(S)$. Next, we employ the product rule for the set derivative (see Def. 9) to obtain the global pdf for D in terms of the singleton belief functions and their first derivatives. Higher derivatives of β_{D^j} are zero since D^j are singletons (see Remark 17). Thus, the product rule yields first derivatives on all (ordered) subsets of the singleton belief functions:

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

By Proposition 16, we have that $\beta_{D^j}(\emptyset) = (1 - q^{(j)})$ and $\frac{\delta \beta_{D^{j_i}}}{\delta \xi_i}(\emptyset) = q_{j_i} p^{(j_i)}(\xi_i)$ and so

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

which nearly resembles Eq. (3.9). To bridge the gap, we describe the choice of indices j_i by an injective function from $\{1, ..., N\}$ into $\{1, ..., M\}$. 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, Π_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 γ as in Eq. (3.10). We reorder the product in order to shift these permutations onto the input variables, obtaining

$$\frac{\delta^N \beta_D}{\delta\xi_1 ... \delta\xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} \sum_{\gamma \in I(N,M)} \mathcal{Q}(\gamma) \prod_{k=1}^{N} p^{(\gamma(k))}(\xi_{\pi(k)}). \tag{3.12}$$

Finally, the global pdf in Eq. (3.9) follows directly from applying Eq. (3.7) to Eq.(3.12). ■

Remark 21 *The global pdf in Eq.* (3.9)*, 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.* (3.12) *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.*

Example 1 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 20 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

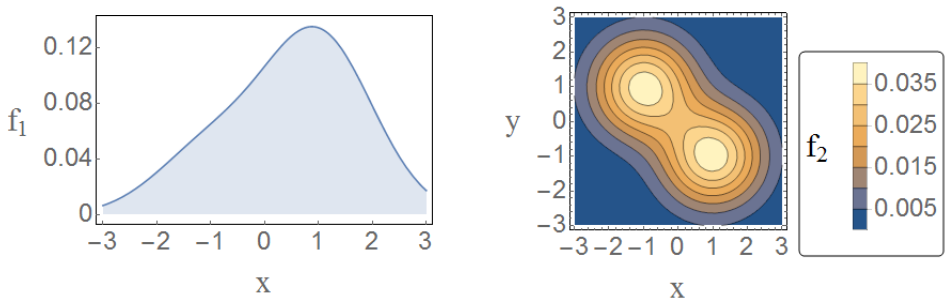


Figure 2: *Left:* Plot of the local density $f_1(x)$ in Eq. (3.13a). *Right:* Contour plot of the local density $f_2(x, y)$ in Eq. (3.13b). These pdfs cover the different possible input dimensions and are symmetric under permutations of the input.

$f_0 = \mathbb{P}[|D| = 0] = (1 - q^{(1)})(1 - q^{(2)}) = 0.08$, $f_1 = f_D\big|_{\mathbb{R}}$, and $f_2 = f_D\big|_{\mathbb{R}^2}$. We sum over permutations and divide by $N!$ ($N = 1, 2$ is the input cardinality) to obtain a symmetric global pdf.

$$\begin{aligned} f_1(x) &= (1 - q^{(2)})q^{(1)}p^{(1)}(x) + (1 - q^{(1)})q^{(2)}p^{(2)}(x) \\ &= \frac{0.12}{\sqrt{2\pi}}e^{-(x+1)^2/2} + \frac{0.32}{\sqrt{2\pi}}e^{-(x-1)^2/2}, \end{aligned} \tag{3.13a}$$

$$\begin{aligned} f_2(x, y) &= \frac{q^{(1)}q^{(2)}}{2}\left[p^{(1)}(x)p^{(2)}(y) + p^{(1)}(y)p^{(2)}(x)\right] \\ &= \frac{0.24}{2\pi}\left(e^{-((x-1)^2+(y+1)^2)/2} + e^{-((x+1)^2+(y-1)^2)/2}\right). \end{aligned} \tag{3.13b}$$

Accounting for each cardinality and following Eq. (3.13a) and Eq. (3.13b), the total probability adds up to

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

as desired. The local densities in Eq. (3.13a) and Eq. (3.13b) are plotted in Fig. 2. Though $f_1(x)$ is the sum of two Gaussians, in Fig. 2 (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.

4. Kernel Density Estimation

Lemma 3.2 yields a definition of global pdf for a random persistence diagram that considers all features individually; however, as seen in Example 1, the computation of Eq. (3.9) can be rather formidable if one considers persistence diagrams with more than two points. To that end, our goal is the construction of a kernel density centered at a persistence diagram $\mathscr{D}$ with a bandwidth $\sigma > 0$ that reduces computational burden by treating some features individually and others collectively.

Generically, persistence diagrams have the majority of their points concentrated close to the diagonal. Consequently, the bandwidth σ is responsible for splitting a persistence diagram into upper and lower portions; see Eq. (4.1) and Fig. 3 (*Left*). The upper portion models the most topologically prominent points, which encompass topological information about the data, and its distribution reflects uncertainty in the precise location for prominent topological features in a persistence diagram. The lower portion models the majority of points in a persistence diagram. These points arise as a result of local noise in the underlying data, and in this fashion its distribution prescribes a noise likelihood model. Moreover, one can evaluate diagrams of any cardinality in the kernel (in this sense, the kernel is a global density). On the other hand, if one fixes the cardinality, one obtains the local kernel.

The construction of the kernel density proceeds by treating the upper and lower parts as independent, which necessitates the establishment of two density functions, one for each portion. The density for the upper part follows the recipe of Lemma 3.2 with a modified Gaussian chosen for $p^j(\xi)$ in Eq. (3.9). To construct the kernel for the lower portion, we use (i) the number of points in $\mathscr{D}^\ell$ to create a pertinent counting measure, and (ii) a modified Gaussian mixture with mean the projection of each point in $\mathscr{D}^\ell$ to the diagonal. When evaluating a persistence diagram in the composite kernel, some of the points are evaluated in the density for the lower while others are evaluated in the density for the upper part. For a particular allocation of points to the upper and lower portions and by independence, the total evaluation follows from multiplying the results of these two evaluations together. However, since it is unknown a priori which input points should be used in each kernel, one must account for every possible partitioning of input points.

Section 4.1 gives a precise construction of of our kernel density estimator. In Section 4.2, we prove that our kernel density estimator converges to the true distribution as the number of persistence diagrams used to create it goes to infinity. Finally, Section 4.3 proposes a new statistic for persistence diagrams, the mean absolute bottleneck deviation (MAD), and establishes convergence of the sample MAD computed with our kernel density estimator.

4.1. Construction

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 20. 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 $\mathscr{D} \subset \mathcal{W}_k = W \times \{k\}$ (see Eq. (2.1)). Since k is fixed, we treat $\mathscr{D} = \{\xi_i\}_{i=1}^M = \{(b_i, d_i)\}_{i=1}^M$ within $W = \big\{(b,d) \in \mathbb{R}^2 : d > b \geq 0\big\}$.

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. 10 (g), (h) or Fig. 14 (Right)), 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, 2018; De Silva and Ghrist, 2007; Xia et al., 2015; Donato et al., 2016; Atienza et al., 2016). Thus, we split $\mathscr{D}$ into

upper and lower portions according to a bandwidth σ as

$$\mathscr{D}^u = \{(b_i, d_i, k) \in \mathscr{D} : d_i - b_i \geq \sigma\} \text{ and } \mathscr{D}^\ell = \{(b_i, d_i, k) \in \mathscr{D} : d_i - b_i < \sigma\}. \tag{4.1}$$

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

Definition 22 *Each feature $\xi_j = (b_j, d_j) \in \mathscr{D}^u$ yields an independent random singleton diagram D^j defined by its chance to be nonempty $q^{(j)}$ (via Eq. (4.3)) 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 20, where each $p^{(j)}$ is given by the pdf associated with $N^*((b_j, d_j), \sigma I)$, which is given by*

$$p^{(j)}(b, d) = \frac{\varphi_j(b, d)}{\int_W \varphi_j(u, v) du\, dv} \mathbb{1}_W(b, d), \tag{4.2}$$

where φ_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

$$q^{(j)} = \mathbb{P}(D^j \neq \emptyset) = \int_{\{v>u\}} \varphi_j(u, v)\, du\, dv. \tag{4.3}$$

Afterward, the Gaussian restricted to T is further restricted to W and renormalized to obtain a probability measure as in Eq. (4.2). This double restriction to both T and W is necessary for proper restriction of the Gaussian pdf and definition of $q^{(j)} = \mathbb{P}(D^j \neq \emptyset)$. 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 23 *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 $(0, d_j)$ with $k = 0$, we instead take*

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

where ϕ_j is the 1-dimensional Gaussian centered at d_j with standard deviation σ.

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) ν and lower density p^ℓ.

Definition 24 *The lower random diagram D^ℓ is defined by choosing a cardinality N according to a pmf ν followed by N i.i.d. draws according to a fixed density p^ℓ. First, take $N_\ell = \left|\mathscr{D}^\ell\right|$ and define $\nu(\cdot)$ with mean N_ℓ and so that $\nu(n) = 0$ for $n > mN_\ell$ for some $m > 0$ independent of N_ℓ. The subsequent density $p^\ell(b,d)$ is given by projecting the lower features $\mathscr{D}^\ell$ of the center diagram $\mathscr{D}$ onto the diagonal $b = d$, then creating a restricted Gaussian kernel density estimation for these features; specifically,*

$$p^\ell(b,d) = \frac{1}{N_\ell} \sum_{(b_i,d_i)\in\mathscr{D}^\ell} \frac{1}{\pi\sigma^2} e^{-\left(\left(b-\frac{b_i+d_i}{2}\right)^2+\left(d-\frac{b_i+d_i}{2}\right)^2\right)/2\sigma^2}. \tag{4.4}$$

Projecting the lower features $\mathscr{D}^\ell$ of the center diagram $\mathscr{D}$ onto the diagonal simplifies later analysis and evaluation of p^ℓ; without projecting, a unique normalization factor, similar to $q^{(j)}$ in Def. 22, would be required for each Gaussian summand in Eq. (4.4). By Proposition 16 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:

$$f_{D^\ell}(\xi_1,...,\xi_N) = \nu(N)\prod_{j=1}^{N} p^\ell(\xi_j). \tag{4.5}$$

Eq. (4.5) provides a noise model for the short-lived features near the diagonal. Combining the expressions for D^ℓ and D^u, we arrive at the following proposition.

Proposition 25 *Fix a center persistence diagram $\mathscr{D}$ and bandwidth $\sigma > 0$. Split $\mathscr{D}$ into $\mathscr{D}^\ell$ and $\mathscr{D}^u$ according to Eq. (4.1). Define D^ℓ with global pdf from Eq. (4.5), and D^u with global pdf from Eq. (3.9). Treating the random persistence diagrams D^u and D^ℓ as independent, define their union D. The following kernel density satisfies Def. 13 as the global pdf of D:*

$$K_\sigma(Z,\mathscr{D}) = \sum_{j=0}^{N_u} \nu(N-j) \sum_{\gamma\in I(j,N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p^\ell(\xi_k), \tag{4.6}$$

where $Z = (\xi_1,...,\xi_N)$ is the input, $\xi_i = (b_i,d_i)$ for $i = 1,...,N$ are the features, and $N_u = |\mathscr{D}^u|$ depends on both $\mathscr{D}$ and σ. Here $\mathcal{Q}(\gamma)$ is given by Eq. (3.10), each $p^{(j)}$ refers to the modified Gaussian pdf as shown in Eq. (4.2) for its matching feature ξ_j in D^u, and p^ℓ is given by Eq. (4.4).

Proof Since D^u and D^ℓ are independent random persistence diagrams, the belief function decomposes into $\beta_D(S) = \beta_{D^u}(S)\beta_{D^\ell}(S)$. Moreover, since derivatives above order N_u vanish for β_{D^u} (see Remark 17), the product rule and binomial-type counting yield

$$\begin{aligned}\frac{\delta^N\beta_D}{\delta\xi_1...\delta\xi_N}(\emptyset) &= \sum_{j=0}^{N_u} \sum_{1\le i_1\neq...\neq i_j\le N} \frac{\delta^j\beta_{D^u}}{\delta\xi_{i_1}...\delta\xi_{i_j}}(\emptyset) \frac{\delta^{N-j}\beta_{D^\ell}}{\delta\xi_1...\delta\hat{\xi}_{i_1}...\delta\hat{\xi}_{i_j}...\delta\xi_N}(\emptyset) \\ &= \sum_{\pi\in\Pi_N} \sum_{j=0}^{N_u} \frac{1}{j!(N-j)!} \frac{\delta^j\beta_{D^u}}{\delta\xi_{\pi(1)}...\delta\xi_{\pi(j)}}(\emptyset) \frac{\delta^{N-j}\beta_{D^\ell}}{\delta\xi_{\pi(j+1)}...\delta\xi_{\pi(N)}}(\emptyset)\end{aligned} \tag{4.7}$$

where $\delta\hat{\xi}_i$ indicates that the given index is skipped in the set derivative (having been allocated to the other factor). Similar to the proof of Lemma 20, the choice of indices i_j is replaced with a

permutation $\pi \in \Pi_N$; however, the ordering within each derivative is unrelated the choice of i_j, leading to $j!$-fold and $(N-j)!$-fold redundancy within each term.

Taking Eq. (4.5) together with Eq. (3.7) yields

$$\frac{\delta \beta_{D^\ell}}{\delta \xi_{\pi(j+1)} \ldots \delta \xi_{\pi(N)}}(\emptyset) = (N-j)! \nu(N-j) \prod_{j=1}^{N-j} p^\ell(\xi_j).$$

Also, Eq. (3.9) and Eq. (3.7) yield

$$\frac{\delta \beta_{D^u}}{\delta \xi_{\pi(1)} \ldots \delta \xi_{\pi(j)}}(\emptyset) = \sum_{\pi^* \in \Pi_j} \sum_{\gamma \in I(j, N_u)} \mathcal{Q}(\gamma) \prod_{k=1}^{j} p^{(\gamma(k))}(\xi_{\pi^*(k)}).$$

We substitute these relations into the final expression of Eq. (4.7). The first of these substitutions is straightforward, while the second has $j!$-fold redundant permutations overtop the existing permutations in Π_N. These substitutions yield that $\frac{\delta^N \beta_D}{\delta \xi_1 \ldots \delta \xi_N}(\emptyset) = \sum_{\pi \in \Pi_N} K_\sigma(Z, \mathscr{D})$ as described in Eq. (4.6) and shows that the kernel $K_\sigma(Z, \mathscr{D})$ satisfies the definition of a global pdf for D (Def. 13). Finally, the sum over permutations is removed according to Eq. (3.7) to obtain the expression for $f_D(Z) = K_\sigma(Z, \mathscr{D})$. ■

Remark 26 *A specific example of the component distributions provided for the kernel in Proposition 25 is presented in Fig. 3. Since the kernel density K_σ of Eq.* (4.6) *is a probability density according to Def. 13, it is a function on $\cup_{N=0}^{M} \mathcal{W}_{0:\mathbb{d}-1}^N$, and so the sum of several such kernels is defined by adding each local pdf layer separately.*

Remark 27 *In the definition of our kernel, a single parameter σ 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 31). In general, the bandwidth parameter σ_2 which refers to the standard deviation used to define the Gaussians (as σ appears in Defs. 22 and 24) need not be equal to the splitting parameter σ_1 which determines which points are in $\mathscr{D}^u$ or $\mathscr{D}^\ell$ (as σ appears in Eq.* (4.1)*). 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 31). For a fixed kernel bandwidth σ_2, increasing C (and thus σ_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.*

Since the kernel density is a probability density function for a random persistence diagram, it has an associated probability hypothesis density (See Def. 18).

Corollary 28 *Fix a center persistence diagram $\mathscr{D}$ and bandwidth $\sigma > 0$. Split $\mathscr{D}$ into $\mathscr{D}^\ell$ and $\mathscr{D}^u$ according to Eq.* (4.1)*. Define D^ℓ with global pdf from Eq.* (4.5)*, and D^u with global pdf from*

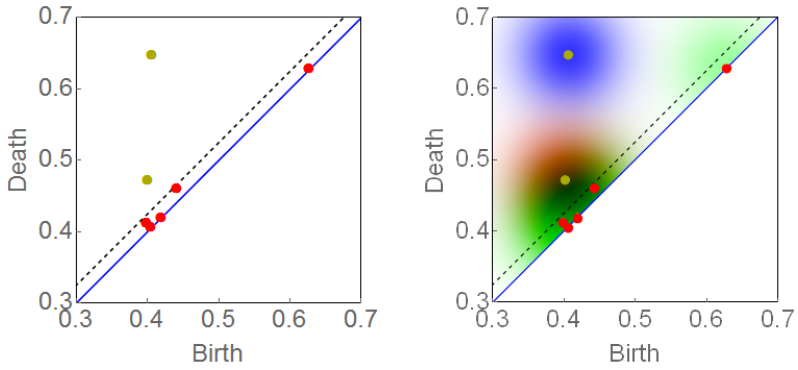


Figure 3: *Left:* A persistence diagram split according to Eq. (4.1). The dashed black line, $d = b+\sigma$, separates the diagram into the red upper points of $\mathscr{D}^u$ and the yellow lower points of $\mathscr{D}^\ell$. *Right:* The red and black gradients represent the upper singleton densities $p^{(1)}$ and $p^{(2)}$ given by Eq. (4.2). The green gradient represents the lower density p^ℓ defined in Eq 4.4. While each of these densities is defined on the wedge $W \subset \mathbb{R}^2$, the global kernel in Eq. (4.6) is defined on $\bigcup_N W^N$ for each input-cardinality N.

Eq. (3.9). *Treating the random persistence diagrams D^u and D^ℓ as independent, the probability hypothesis density (PHD) associated with the kernel density centered at $\mathscr{D}$ with bandwidth σ of Theorem 25 is given by*

$$K_{\sigma,PHD}(\xi, \mathscr{D}) = N_\ell \, p^\ell(\xi) + \sum_{j=1}^{N_u} q^{(j)} p^{(j)}(\xi), \tag{4.8}$$

where the feature ξ is the input and $N_u = |\mathscr{D}^u|$ and $N_\ell = \left|\mathscr{D}^\ell\right|$ depend on both $\mathscr{D}$ and σ. Here each $p^{(j)}$ refers to the modified Gaussian pdf as shown in Eq. (4.2) for its matching singleton feature ξ_j in D^u, $q^{(j)}$ given by (4.3) is the probability each singleton is present, and the lower density p^ℓ is given by Eq. (4.4).

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^ℓ is weighted according to the mean draw cardinality, which was chosen to be $\left|\mathscr{D}^\ell\right|$. ■

Remark 29 (Computational cost) *The kernel density presented in Eq. (4.6) of Proposition 25 has approxmiately $N!2^{N_u}$ terms, necessitating shrewd computational strategies for real world usage. In practice, one may choose to consider only terms that correspond to high probability matchings. Such an implementation may be carried out with a linear assignment algorithm, like Munkres, resulting in a computational complexity of $O(N_u N^3)$. Another approximation for the full kernel density is to consider input features as independent draws from the PHD given in (4.8) of Corollary 28. Evaluation of a diagram in Eq. (4.8) has time complexity $O(N(N_u + N_\ell))$. Finally, sampling from the kernel in Proposition 25 has cost $O(N_u + N_\ell)$.*

Notice that in Corollary 28, the input for the PHD is a single feature ξ as opposed to a list of features $Z = \{\xi_1, ..., \xi_N\}$ for the global kernel in Proposition 25. Furthermore, Proposition 25 extends to the analogue result for a center persistence diagram with features of varied degree of homology.

Corollary 30 *Consider a persistence diagram $\mathscr{D} = \bigcup_{k=0}^{\mathrm{d}-1} \mathscr{D}_k \times \{k\}$ split according to the degrees of homology with associated random persistence diagrams D_k defined according to Eq. (4.6) for each center diagram $\mathscr{D}_k$. Treating each D_k as independent, the full global pdf for $D = \bigcup D_k$ centered at $\mathscr{D}$ with bandwidth σ is given by*

$$K_\sigma(Z, \mathscr{D}) = \Lambda(N) \prod_{k=0}^{\mathrm{d}-1} K_\sigma(Z_k, \mathscr{D}_k), \tag{4.9}$$

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

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

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 $\frac{\delta \beta_{D_k}}{\delta Z}(\emptyset)$ are zero unless $Z \subset \mathcal{W}_k$. Thus, the product rule leaves only the single term $\frac{\delta \beta_D}{\delta Z}(\emptyset) = \prod_{k=0}^{\mathrm{d}-1} \frac{\delta \beta_{D_k}}{\delta Z_k}(\emptyset)$. In turn, each kernel global pdf $K_\sigma(Z_k, \mathscr{D}_k)$ is related to the associated belief function derivative by a sum over permutations Π_{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 $\Lambda(N)$. ■

4.2. Convergence of the Kernel Density Estimator

In this section, to prove the convergence (to the target distribution) of the kernel density estimate defined via the kernel established in Proposition 25, we consider persistence diagrams $\{\mathscr{D}_i\}_{i=1}^n$ 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 $f(\xi_1, ..., \xi_N) \leq C_N \left\|(\xi_1, ..., \xi_N)\right\|^{-2N}$ 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 $\mathbb{R}^{\mathrm{d}}$ of bounded cardinality. The conditions $(A2)$ and $(A3)$ hold for underlying data sampled from a compact set $E \subset \mathbb{R}^{\mathrm{d}}$ perturbed by Gaussian noise. The work (Adler et al., 2014)(see Corollary 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.

Theorem 31 *Consider a random persistence diagram global pdf f satisfying assumptions (A1)-(A3). Define the kernel $K_\sigma(Z, \mathscr{D})$ according to Theorem 25 and consider the kernel density estimate $\hat{f}(Z) = \frac{1}{n}\sum_{i=1}^{n} K_\sigma(Z, \mathscr{D}_i)$, with centers $\mathscr{D}_i$ 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 32 *The value of α_{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 $\sigma_{opt} = A(K)n^{-1/(\mathrm{d}+4)}$, 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 $\alpha \cong 1/(2M+4)$ 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.*

Eq.(4.8) of Corollary 28 could be used as an approximation to the full kernel. The following argument verifies its convergence.

Corollary 33 *Let F_g denote the PHD (Def. 18) of a random persistence diagram with global pdf g. Define $\hat{f}$ as in Thm 31. For a random persistence diagram whose global pdf f satisfies assumptions (A1)-(A3) of Thm 31, one has $F_{\hat{f}} \to F_f$ as $n \to \infty$ almost everywhere.*

Proof Let $C \subset W$ be compact. Define the counting function $\kappa_C(Z) = |Z \cap C|$. Notice by Def. 18 that $|\int_C F_{\hat{f}}(u)du - \int_C F_f(u)du| \leq \int_W \kappa_C(Z)|\hat{f}(Z) - f(Z)|\delta Z \leq M(2^{M+1}-1)\int_C |\hat{f}(Z) - f(Z)|\delta Z$, where the last inequality follows because we assume that random persistence diagrams are bounded by M and κ_C vanishes outside of C. By Thm 31 and boundedness of C, we can choose n sufficiently large to ensure $\int_C |\hat{f}(Z) - f(Z)|\delta Z$ is arbitrarily small. Hence, $\int_C F_{\hat{f}}(u)du \to \int_C F_f(u)du$ as $n \to \infty$ on arbitrary compact subsets C. The result follows immediately by standard results in measure theory. ■

The proof of Thm. 31 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 30.

Throughout the proof we use ξ_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 $\Delta_\alpha^\beta = \{(b, d) \in W : \alpha < d - b < \beta\}$.

Lemma 34 *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Then there exists $C > 0$ so that $\mathbb{E}^f(|\Delta_0^\sigma \cap D|) \leq C\sigma$.*

Proof Consider a region $A \subset W$ and a counting function $\kappa_A(Z) = |Z \cap A|$ such that $\kappa_A(\{\xi_1, ..., \xi_N\}) = \sum_{i=1}^N \mathbb{1}_A(\xi_i)$. It is clear that this set function is well defined and measurable if A is measurable. Using set integration (Def. 12),

$$\mathbb{E}(|\Delta_0^\sigma \cap D|) = \int_W \kappa_{\Delta_0^\sigma}(Z) f(Z) \delta Z = \sum_{N=0}^M \frac{N}{N!} \int_W \mathbb{1}_{\Delta_0^\sigma}(\xi_1) \left[\int f(\xi_1, ...\xi_N) d\xi_2 ... d\xi_N \right] d\xi_1 \tag{4.10}$$

The expressions in Eq. (4.10) can be phrased in terms of the probability hypothesis density from Eq. (3.8), and for any choice of $L > 0$ are bounded by

$$\begin{aligned} \int_{\Delta_0^\sigma} F_D(\xi) d\xi &\leq \int_0^L \int_{y-\sigma}^y F_D(x, y)\, dx\, dy + \int_L^\infty \int_{y-\sigma}^y C_3 y^{-2}\, dx\, dy \\ &\leq LC_2\sigma + 3C_3\sigma/L = (LC_2 + C_3/L)\sigma \end{aligned}$$

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

Lemma 34 yields control over the counting measure ν_i defined in Def. 24 and the coefficients $\mathcal{Q}_i^*(\cdot)$ of Eq. (3.11) which respectively determine the distribution of lower and upper cardinalities for a persistence diagram sampled according to the kernel density $K_\sigma(Z, \mathscr{D}_i)$.

Corollary 35 *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Take ν to be the lower cardinality probability mass function for the kernel density $K_\sigma(Z, D)$ shown in Eq. (4.6). 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, ν is random with respect to f as well. Recall that ν is defined so that $\mathbb{E}^\nu(\boldsymbol{a}) = \left|D^\ell\right|$ for $\boldsymbol{a}$ distributed according to ν and thus $\mathbb{E}^f[\mathbb{E}^\nu(\boldsymbol{a})] \leq C\sigma$ for some $C > 0$ by Lemma 34. Subsequently, the value $\mathbb{E}^f \nu(j_0)$ is controlled by this double expectation so long as $j_0 \neq 0$. Indeed,

$$\mathbb{E}(\boldsymbol{a}) = \sum_{j=0}^\infty j\nu(j) = \sum_{j=1}^\infty j\nu(j) \geq \sum_{j=1}^\infty \nu(j) \geq \nu(j_0)$$

for any $j_0 > 0$ and $\nu_i(j_0) = 0$ for $j_0 < 0$ since it represents a cardinality distribution. ■

In the following lemma, the result of Lemma 34 is used to control the expressions $\mathcal{Q}(\gamma)$ or $\mathcal{Q}^*(\gamma)$, of Eq. (3.10) and Eq. (3.11) respectively, in the kernel density estimate.

Lemma 36 *Consider a random persistence diagram D distributed according to f satisfying assumptions (A1)-(A3). Take $\mathcal{Q}$ of Eq. (3.10) and $\mathcal{Q}^*$ of Eq. (3.11) to be the upper singleton probabilities for the kernel density $K_\sigma(Z, D)$ shown in Eq. (4.6). Then, there exists $C > 0$ so that $\mathbb{E}^f[\mathcal{Q}(\gamma)] \leq \mathbb{E}^f[\mathcal{Q}^*(\gamma)] \leq C\sigma$ for any $\gamma \in I(j, N)$ with $j < N$.*

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

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. (4.3); or in terms of the error function Φ, $q(p) = \frac{1}{2}\left(1 + \Phi\left(\frac{p}{2\sigma}\right)\right)$. Define the minimal persistence as $p_{\min}(Z) = \sup\{p : |\Delta_0^p \cap Z| = \emptyset\}$ which satisfies $p_{\min}(Z) \geq p$ if and only if $|\Delta_0^p \cap Z| = \emptyset$. In turn, we may bound $\mathcal{Q}^*(\gamma) \leq (1 - q(p_{\min}(D))$ independently of γ. By Lemma 34, there is $C > 0$ such that $\mathbb{P}^f[|\Delta_0^\sigma \cap D| \neq \emptyset] \leq \mathbb{E}^f[|\Delta_0^\sigma \cap D|] \leq C\sigma$, which controls the distribution of the minimal persistence.

In particular, $q'(p) = \frac{1}{2\sigma\sqrt{\pi}} e^{-p^2/4\sigma^2}$ by the fundamental theorem of calculus. The control of Lemma 34 and the fact that $p_{\min}(Z) \geq 0$ also allows us to use integration via the probability of sublevel sets. Take $g(p) = 1 - q(p)$ so that $\lim_{p\to\infty} g(p) = 0$. Specifically, since $\mathcal{Q}^*(\gamma) \leq (1 - q(p_{\min}(D))$, and using the fundamental theorem of calculus then Fubini's theorem, we have:

$$\begin{aligned} \mathbb{E}^f[\mathcal{Q}^*(\gamma)] &\leq \int_{\mathcal{W}_{0:\mathrm{d}-1}} g(p_{\min}(Z)) f(Z)\delta Z = \int_{\mathcal{W}_{0:\mathrm{d}-1}} \left(\int_{\infty}^{p_{\min}(Z)} g'(p)dp \right) f(Z)\delta Z \\ &= \int_{\infty}^{0} \left(\int_{\{Z : p_{\min}(Z) < p\}} f(Z)\delta Z \right) g'(p)dp = \int_{0}^{\infty} \left(\mathbb{P}^f[p_{\min} < p] \right) q'(p)dp. \end{aligned} \tag{4.11}$$

We now further bound the expectation in Eq. (4.11). Replacing terms with their definitions and using the bound control from Lemma 34 we obtain:

$$\begin{aligned} \mathbb{E}^f[\mathcal{Q}^*(\gamma)] &\leq \int_0^\infty \mathbb{P}^f(\Delta_0^p \cap D \neq \emptyset) \frac{1}{2\sigma\sqrt{\pi}} e^{-p^2/4\sigma^2} dp \\ &\leq \frac{C}{2\sigma\sqrt{\pi}} \int_0^\infty p e^{-(p/2\sigma)^2} dp = \frac{C}{2\sigma\sqrt{\pi}} \left[-2\sigma^2 e^{-p^2/4\sigma^2} \right]_{p=0}^{\infty} = \frac{C}{\sqrt{\pi}}\sigma. \end{aligned}$$

■

Proof of Theorem 31. For convenience, we denote the upper cardinalities by $N_i = |\mathscr{D}_i^u|$ and total cardinalities by $M_i = |\mathscr{D}_i|$ for the sample persistence diagrams. Denote the set of strictly increasing functions from $\{1, ..., j\}$ into $\{1, ..., N_i\}$ by $I(j, N_i)$. Here we use 'id' to denote the identity map, where $I(N_i, N_i) = \{\mathrm{id}\}$. The proof is organized by splitting the kernel densities into several pieces and then controlling each piece separately.

First, we separate the kernel $K_\sigma(Z, \mathscr{D}_i)$, defined in Eq. (4.6), into three portions, A_i, B_i, and C_i, according to the upper cardinality j:

$$\begin{aligned} K_\sigma(Z, \mathscr{D}_i) &= \sum_{j=0}^{N_i} \nu_i(N-j) \sum_{\gamma \in I(j,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{j} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \\ &= \nu_i(N-N_i)\mathcal{Q}_i(\mathrm{id}) \prod_{k=1}^{N_i} p_i^{(k)}(\xi_k) \prod_{k=N_i+1}^{N} p_i^{\ell}(\xi_k) \\ &+ \sum_{j=0, j\neq N}^{N_i-1} \nu_i(N-j) \sum_{\gamma \in I(j,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{j} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \\ &+ \mathbb{1}_{\{n\in\mathbb{N}: n<N_i\}}(N)\nu_i(0) \sum_{\gamma \in I(N,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k) \\ &= A_i + B_i + C_i, \end{aligned} \tag{4.12}$$

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

$$\sum_{j=0, j\neq N}^{N_i-1} \sum_{\gamma \in I(j,N_i)} \left[\prod_{k=1}^{j} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k) \prod_{k=j+1}^{N} p_i^{\ell}(\xi_k) \right]. \tag{4.13}$$

The bounding sum of Eq. (4.13) consists of restricted $2N$-dimensional Gaussians, with the weights $q_i^{(j)}$ dominating the restriction rescaling in Eq. (4.2). Fix $\pi \in \Pi_N$ and $j \in \{0, ..., M-1\} \setminus \{N\}$. Without loss of generality, we treat the case when the permutation π 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 $M_i = |\mathscr{D}_i|$ are fixed at level $M_i = m \in \{0, ..., M\}$. Now, $m = |\mathscr{D}_i| \geq N_i > j$, so there is some extension for every γ within the sum, $\gamma^* \in \Pi_m$. Recall that this collection is random because each $\mathscr{D}_i$ is randomly distributed according to f, therefore we consider the expectation with respect to this randomness:

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

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

$$\mathbb{E}^f \left[\sum_{\{i:M_i=m\}} \frac{1}{|\{i : M_i = m\}|} \prod_{k=1}^{j} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k) \right] \to \int_{W^{m-j}} f(\xi_1, ..., \xi_m) d\xi_{j+1} ... d\xi_m, \tag{4.14}$$

which is again bounded via (A2) and (A3). Of course, $|\{i : M_i = m\}| \leq n$, so taking Eq. (4.14) into account for every m bounds the averaging sum of the upper product: $\frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{j} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k)$.

Relying on Eq. (4.13), we must also consider the lower product $\prod_{k=j+1}^{N} p_i^{\ell}(\xi_k)$. Since the points ξ_i are fixed, we focus on their minimal persistence $p_{\min} = \min_i(d_i - b_i)$. Thus,

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

and subsequently,

$$\left[\prod_{k=j+1}^{N} p_i^{\ell}(\xi_k)\right] \leq \frac{1}{(2\pi\sigma^2)^N} e^{-Np_{\min}^2/4\sigma^2} \to 0, \tag{4.15}$$

as $\sigma \to 0$, uniformly on any compact subset of W (or $\mathcal{W}_{0:\mathrm{d}-1}$). Altogether, Eqs. (4.14) and (4.15) 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. (4.12). We split the sum $\frac{1}{n}\sum_{i=1}^{n} A_i$ according to the cardinality of $\mathscr{D}_i$. 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 $\mathrm{MM}(n)$ 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 $\mathscr{D}_i$. We obtain the following mismatched term:

$$\frac{1}{n}\sum_{\{i:N_i\neq M_i\}} A_i \leq \left(\frac{MM(n)}{n}\right)\frac{1}{MM(n)}\sum_{\{i:N_i\neq M_i\}}\left[\mathcal{Q}_i(\mathrm{id})\prod_{k=1}^{N_i} p_i^{(k)}(\xi_k)\prod_{k=N_i+1}^{N} p_i^{\ell}(\xi_k)\right] \tag{4.16}$$

The bounding sum in Eq. (4.16) is split into pieces where $M_i = m$ for each m between 0 and M. Using the same strategy yielding Eq. (4.14), with $MM(n)$ 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 $\gamma = \mathrm{id}$. Using the same approach leading to Eq. (4.15), 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 34, $\mathbb{P}^f(M_i \neq N_i) = \mathbb{P}^f(\mathscr{D}_i \cap \Delta_0^{\epsilon\sigma} \neq \emptyset) \leq C_5\sigma$; consequently, $\mathbb{E}^f[MM(n)/n] \to 0$ and the mismatch terms on left hand side of Eq. (4.16) follow.

Now consider the indices for which $N_i = M_i$. In this case, since $\mathscr{D}_i^{\ell}$ 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 kernel density estimate is given by

$$\frac{1}{n}\mathbb{E}^f\sum_{\{i:N_i=M_i\}} A_i = \frac{1}{n}\mathbb{E}^f\left[\sum_{\{i:N_i=M_i\}}\left(\mathcal{Q}_i(\mathrm{id})\prod_{k=1}^{N} p_i^{(k)}(\xi_k)\right)\right] = \frac{1}{n}\mathbb{E}^f\left[\sum_{\{i:N_i=M_i\}}\left(\prod_{k=1}^{N} q_i^{(k)} p_i^{(k)}(\xi_k)\right)\right]. \tag{4.17}$$

As shown, the terms in Eq. (4.17) are restricted $2N$ dimensional Gaussians. It is known (Scott, 2015) that restricted Gaussian kernel density estimates like $\left[\prod_{k=1}^{N} q_i^{(k)} p_i^{(k)}(\xi_k)\right]$ converge (uniformly on compactly contained sets) to the true value of the chosen draws $\mathscr{D}_i$ for a suitable choice of α in $\sigma = O(n^{-\alpha})$ as restricted by $N \leq M$. After correcting for the samples with $N_i < M_i = N$, the samples $\mathscr{D}_i$ are treated as random draws from $f(D\,||D| = N)$. Consequently, we may conclude

that the target distribution associated with $\left[\prod_{k=1}^{N} q_i^{(k)} p_i^{(k)}(\xi_k)\right]$ is the rescaled $\frac{1}{f(N)} f(\xi_1, ..., \xi_N)$, where $f(N) := \mathbb{P}^f(|D| = N)$. This rescaling for the conditional pdf $f(D\,||D| = N)$ is necessary to reweight according to Proposition 16.

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 34. 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. (4.12). We begin by bounding the probability mass functions ν_i by 1 and considering only terms for which the characteristic function is nonzero:

$$\frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{\{i:N<N\}} \nu_i(0) \sum_{\gamma\in I(N,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k) \leq \frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \mathcal{Q}_i(\gamma) \prod_{k=1}^{N} p_i^{(\gamma(k))}(\xi_k). \tag{4.18}$$

Next, we split the term $\mathcal{Q}(\gamma)$ according to Eq. (3.10) and apply Lemma 36 to the upper bound in Eq. (4.18) to obtain the larger upper bound

$$\frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \mathcal{Q}^*(\gamma) \prod_{k=1}^{N} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k) \leq C\left[\frac{1}{n}\sum_{\{i:N<N_i\}} \sum_{\gamma\in I(N,N_i)} \prod_{k=1}^{N} q_i^{(\gamma(k))} p_i^{(\gamma(k))}(\xi_k)\right]\sigma. \tag{4.19}$$

The expectation of the bracketed terms in Eq. (4.19) converges in a fashion identical to the terms $\frac{1}{n}\sum_{i=1}^{n} A_i$. Since these terms are multiplied by σ, altogether $\left[\frac{1}{n}\sum_{i=1}^{n} C_i\right]$ vanishes in the limit as $n \to \infty$. Putting together the limits of each portion built from $K_\sigma(Z, \mathscr{D}_i) = A_i + B_i + C_i$, the theorem follows. ■

4.3. A Measure of Dispersion

Theorem 31 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 (Def. 5) 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 37 *The mean absolute bottleneck deviation (MAD) from origin diagram $\mathscr{D}$ with respect to a global pdf f is given by*

$$MAD_f(\mathscr{D}) = \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}, Z) f(Z) \delta Z \tag{4.20}$$

The following proposition and lemma aid in proving convergence of MAD kernel estimates. Their proofs are delegated to the supplementary materials.

Proposition 38 *Consider D distributed according to the kernel density $K_\sigma(\cdot, \mathscr{D})$ with center diagram $\mathscr{D}$ and bandwidth σ. Fix $\delta \geq 1$. Then,*

$$\mathbb{P}\left[W_\infty(D, \mathscr{D}) < \delta\sigma\right] \geq \left(\int_{B(\mathbf{0},\delta)} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dx\, dy\right)^M \tag{4.21}$$

where M is the maximal cardinality of D (a multiple of $|\mathscr{D}|$). 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 $\mathcal{W}_{0:\mathrm{d}-1}$ and tighten the decay control from assumption $(A3)$. Formally,

$(A2)^*$ The local density $f_N : \mathcal{W}_{0:\mathrm{d}-1}^N \to \mathbb{R}$ is bounded for each $N \in \{1, ..., M\}$.

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

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

Lemma 39 *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 $\mathscr{D}$.*

Similar to assumption $(A3)$ (given prior to Theorem 31), $(A3)^*$ holds for a random persistence diagram associated with underlying data sampled from a compact set perturbed by Gaussian noise. One may also replace Lemma 39 and its assumptions by directly assuming that the maximal persistence moment is bounded; with this, the results of Lemma 39 follow immediately from Eq. (A.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 40 *Consider a distribution of persistence diagrams with bounded global pdf, f, satisfying assumptions $(A1)$, $(A2)^*$, and $(A3)^*$. Let $\hat{f}(Z) = \frac{1}{n}\sum_{i=1}^n K_\sigma(Z, \mathscr{D}_i)$ be a kernel density estimate with centers $\mathscr{D}_i$ 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,*

$$\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z)\hat{f}(Z)\delta Z \to \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) f(Z)\delta Z \tag{4.22}$$

as $n \to \infty$ for any origin diagram $\mathscr{D}_0$.

Proof The MAD of f with origin $\mathscr{D}_0$ is finite by Lemma 39. 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

$$\begin{aligned}
&\left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) f(Z)\delta Z - \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z)\hat{f}(Z)\delta Z\right| \\
&\quad\leq \left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) f(Z)\delta Z - \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, \mathscr{D}_i) K_\sigma(Z, \mathscr{D}_i)\delta Z\right| \\
&\quad+ \frac{1}{n}\sum_{i=1}^n \left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z - \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, \mathscr{D}_i) K_\sigma(Z, \mathscr{D}_i)\delta Z\right|.
\end{aligned} \tag{4.23}$$

The term of the upper bound in Eq. (4.23) trivially simplfies to obtain the sample estimator for the MAD:

$$\begin{aligned}\left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) f(Z)\delta Z - \sum_{i=1}^{n} \frac{1}{n}\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, \mathscr{D}_i) K_\sigma(Z, \mathscr{D}_i)\delta Z\right| \\ = \left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) f(Z)\delta Z - \frac{1}{n}\sum_{i=1}^{n} W_\infty(\mathscr{D}_0, \mathscr{D}_i)\right|.\end{aligned} \tag{4.24}$$

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. (4.23) is further bounded via the reverse triangle inequality; specifically,

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{n}\left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z - \int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_0, \mathscr{D}_i) K_\sigma(Z, \mathscr{D}_i)\delta Z\right| \\ \leq \sum_{i=1}^{n} \frac{1}{n}\left|\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_i, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z\right|.\end{aligned} \tag{4.25}$$

Toward bounding Eq. (4.25), choose a threshold parameter $a = O(\sigma^\beta)$ for some $\beta \in (0,1)$, so that $a \to 0$ but $a/\sigma \to \infty$ in the sample size (and bandwidth) limit. Next, take $A_i = \{Z \subset W : W_\infty(Z, \mathscr{D}_i) \leq a\}$ and split the integral between A_i and its complement as

$$\int_{\mathcal{W}_{0:\mathrm{d}-1}} W_\infty(\mathscr{D}_i, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z = \int_{A_i} W_\infty(\mathscr{D}_i, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z + \int_{A_i^c} W_\infty(\mathscr{D}_i, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z.$$

The integral over A_i is trivially bounded by a. Integration over the complementary events is controlled via layered integration along with Proposition 38. For $a/\sigma > 1$, which occurs when n is large enough, we obtain

$$\begin{aligned}\int_{A_i^c} W_\infty(\mathscr{D}_i, Z) K_\sigma(Z, \mathscr{D}_i)\delta Z &= a\,\mathbb{P}^i\left[W_\infty(\mathscr{D}_i, Z) > a\right] + \int_a^\infty \mathbb{P}^i\left[W_\infty(\mathscr{D}_i, Z) > b\right] db \\ &\leq a\left(\mathbb{P}[|z| > a/\sigma]^M\right) + \int_a^\infty \left(\mathbb{P}[|z| > b/\sigma]^M\right) db,\end{aligned} \tag{4.26}$$

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

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

$$\begin{aligned}\int_{A_i^c} g(Z) K_\sigma(Z, \mathscr{D}_i)\delta Z &= \int_{\{Z : g(Z) > a\}} \left(\int_0^{g(Z)} db\right) K_\sigma(Z, \mathscr{D}_i)\delta Z \\ &= \int_0^\infty \int_{\{Z : g(Z) > a \text{ and } g(Z) > b\}} K_\sigma(Z, \mathscr{D}_i)\delta Z db \\ &= \int_0^\infty \mathbb{P}^f[g(Z) > \max\{a, b\}] db \\ &= a\mathbb{P}^f[g(Z) > a] + \int_a^\infty \mathbb{P}^f[g(Z) > b] db.\end{aligned}$$

Applying Proposition 38 changes the probabilities on $g(Z)$ to normal tail probabilities. Thus, both bounding terms in Eq. (4.23) converge to zero and thus the kernel estimate converges to the true mean absolute deviation. ■

5. 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 $((b_1, d_1), ..., (b_N, d_N))$, we consider the kernel density evaluated at $(b_1, d_1) \in W$ with (b_i, d_i) fixed for $i \geq 2$. For clarity, the unique symmetric pdf $f_{sym}(\xi_1, ..., \xi_N) = \frac{1}{N!} \sum_{\pi \in \Pi_N} f(\xi_{\pi(1)}, ..., \xi_{\pi(N)})$ is used in the contour plots (see Remark 15). For explicit computation, we choose the probability mass function

$$\nu(N) = \max \left\{ \frac{N_\ell + 1 - |N_\ell - N|}{(N_\ell + 1)^2}, 0 \right\} \tag{5.1}$$

when evaluating the lower density in Eq. (4.5), where $N_\ell = \left|\mathscr{D}^\ell\right|$ is the lower cardinality of the center diagram. This probability mass function is chosen to satisfy the requirements of Def. 24, and specifically has the property that $\nu(N) > 0$ for $0 \leq N \leq 2\left|\mathscr{D}^\ell\right|$.

Example 2 Consider the center persistence diagram $\mathscr{D} = \{(1, 3), (2, 4), (1, 1.3), (3, 3.2)\} \subset W$ and bandwidth $\sigma = 1/2$. We construct the associated kernel density $K_\sigma(Z, \mathscr{D})$ according to Theorem 25 and follow with some plots and analysis of the kernel density. The random persistence diagram D associated with the kernel density $K_\sigma(Z, \mathscr{D})$ has a variable number of features $N = |D|$; consequently, the input diagram $Z = \{\xi_1, ..., \xi_N\}$ must have variable length and therefore the kernel density has local definitions (see Rmk. 14) on W^N for each possible input cardinality N.

Since each modified Gaussian $p^{(j)}$ (Def. 22) and the lower density p^ℓ (Def. 24) integrate to 1 over the wedge W, an expression for the probability mass function (pmf) $\mathbb{P}[|D| = N]$ can be expressed solely in terms of ν and $q^{(j)}$:

$$\begin{aligned} \mathbb{P}[|D| = N] = &\left[q^{(1)} q^{(2)}\right] \nu(N-2) \\ &+ \left[q^{(1)}\left(1 - q^{(2)}\right) + q^{(2)}\left(1 - q^{(1)}\right)\right] \nu(N-1) \\ &+ \left[\left(1 - q^{(1)}\right)\left(1 - q^{(2)}\right)\right] \nu(N) \end{aligned} \tag{5.2}$$

The plot of this pmf is shown in Fig. 4. Recall that $D = D^u \cup D^\ell$, so that $|D| = |D^u| + \left|D^\ell\right|$; since $q^{(j)} \approx 1$ for $j = 1, 2$, $|D^u| = 2$ with high probability and the pmf $\mathbb{P}[|D| = N]$ is nearly the pmf for $\left|D^\ell\right|$, ν, shifted up by 2 units. Fig. 4 suggests that understanding the kernel density requires investigation into higher cardinality inputs. In general, it is important to consider input diagrams Z with $|Z| \geq |\mathscr{D}^u|$.

First, we describe the random diagram associated to the lower features $\mathscr{D}^\ell = \{(1, 1.3), (3, 3.2)\}$ of the center diagram $\mathscr{D}$. The lower random diagram D^ℓ is described in Def. 24 according to a probability mass function (pmf) ν for the cardinality of D^ℓ and a single probability density $p^\ell(b, d)$ for the subsequent features' locations in the wedge W. The pmf ν is defined according to Eq. (5.1) with $N_\ell = 2$; that is, $\nu(\{0, 1, 2, 3, 4\}) = \{1/9, 2/9, 3/9, 2/9, 1/9\}$ respectively, and zero otherwise.

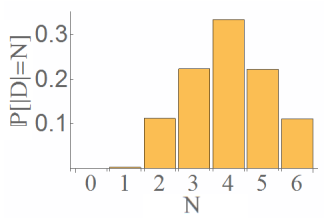


Figure 4: Cardinality probabilities $\mathbb{P}[|D| = N]$ for random diagram D distributed according to global pdf $K_\sigma(\cdot, \mathscr{D})$ in Ex. 2. In general, we have that $0 \leq |D^u| \leq |\mathscr{D}^u|$ and according to Eq. (5.1), $\nu(N) \neq 0$ for $0 \leq N \leq 2\left|\mathscr{D}^\ell\right|$. Thus, the cardinality $|D| = |D^u| + \left|D^\ell\right|$ takes on values between 0 and $6 = |\mathscr{D}^u| + 2\left|\mathscr{D}^\ell\right|$.

Following Def. 24, we project the features of $\mathscr{D}^\ell$ onto the diagonal to obtain $\{(1.15, 1.15), (3.1, 3.1)\}$. Relying on Eq. (4.4), the resulting lower density is given by

$$p^\ell(b,d) = \frac{2}{\pi}\left[e^{-\left((b-1.15)^2+(d-1.15)^2\right)} + e^{-\left((b-3.1)^2+(d-3.1)^2\right)}\right]. \tag{5.3}$$

restricted to the wedge W. The coefficient $\frac{2}{\pi}$ is obtained by a direct substitution into Eq. (4.4).

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 $W^N \subset \mathbb{R}^{2N}$ 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. 5. Recall that the integral of the PHD over a region U yields the expected number of features in U (see definition 18). The kernel's corresponding PHD is a sum of Gaussians as described in Corollary 28.

$$\begin{aligned} K_{\sigma,PHD}((b,d),\mathscr{D}) &= 2p^\ell(b,d) + q^{(1)}p^{(1)}(b,d) + q^{(2)}p^{(2)}(b,d) \\ &= 1.273\left(e^{-2\left((b-3.1)^2+(d-3.1)^2\right)} + e^{-2\left((b-1.15)^2+(d-1.15)^2\right)}\right) \\ &\quad + 0.635e^{-2\left((b-2)^2+(d-4)^2\right)} + 0.635e^{-2\left((b-1)^2+(d-3)^2\right)}. \end{aligned} \tag{5.4}$$

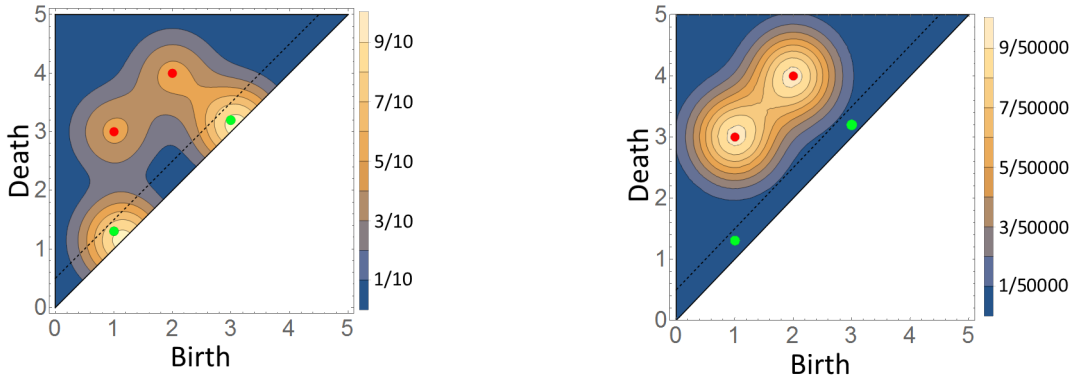
Next, for input of cardinality $|Z| = 1$, we obtain an easily viewable 2-dimensional distribution. Theorem 25 yields the following expression:

$$\begin{aligned} K_\sigma((b_1,d_1),\mathscr{D}) &= \nu(0)\left[(1-q^{(2)})q^{(1)}p^{(1)}(b_1,d_1) + (1-q^{(1)})q^{(2)}p^{(2)}(b_1,d_1)\right] \\ &\quad + \nu(1)\left[(1-q^{(1)})(1-q^{(2)})p^\ell(b_1,d_1)\right]. \\ &= 7.74\times10^{-2}\left(e^{-2\left((b_1-2)^2+(d_1-4)^2\right)} + e^{-2\left((b_1-1)^2+(d_1-3)^2\right)}\right). \\ &\quad + 1.65\times10^{-4}p^\ell(b_1,d_1). \end{aligned} \tag{5.5}$$

The kernel is treated as a global pdf as in Proposition 13 and Rmk. 14; 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.). Since the values of $q^{(j)}$ are very close to 1, terms which include the upper pdfs $p^{(j)}$ have much larger total mass.

Contour plots of the densities expressed in Eqs. (5.4) and (5.5) (restricted to W) are respectively shown in Figs. 5(a) and 5(b). In Fig. 5(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. 5(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. (4.6)) so that the number of points near the diagonal is fluid (by our choice of ν), 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.



(a) (b)

Figure 5: Contour maps for (a) the probability hypothesis density associated to the kernel density (Eq. (5.4)) and (b) the kernel density restricted to a single input feature (Eq. (5.5)). 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. (4.6), we obtain:

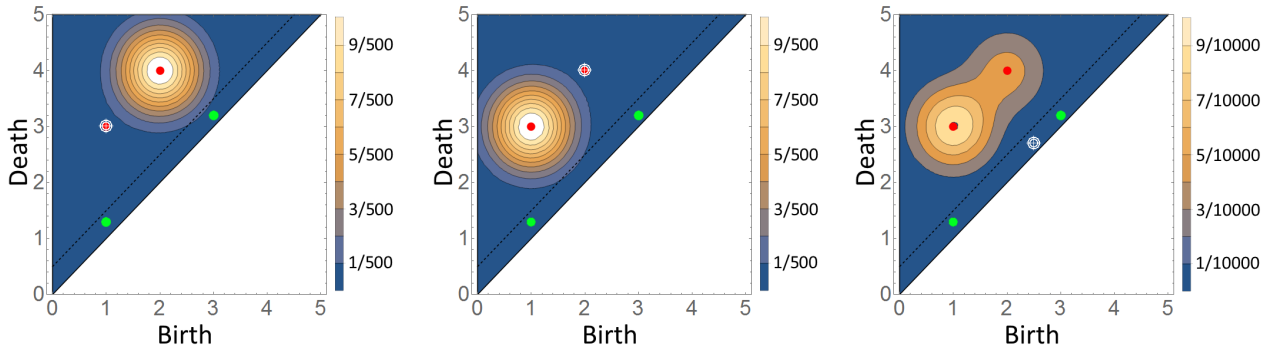
$$\begin{aligned} K_\sigma((\xi_1,\xi_2),\mathscr{D}) &= \nu(0)q^{(1)}q^{(2)}p^{(1)}(b_1,d_1)p^{(2)}(b_2,d_2) \\ &\quad + \nu(1)\left[(1-q^{(2)})q^{(1)}p^{(1)}(b_1,d_1) + (1-q^{(1)})q^{(2)}p^{(2)}(b_1,d_1)\right]p^{\ell}(b_2,d_2) \\ &\quad + \nu(2)(1-q^{(1)})(1-q^{(2)})p^{\ell}(b_1,d_1)p^{\ell}(b_2,d_2) \\ &= 4.5\times 10^{-2}e^{-2\left((b_1-2)^2+(d_1-4)^2\right)}e^{-2\left((b_1-1)^2+(d_1-3)^2\right)} \\ &\quad + 2.11\times 10^{-4}\left[e^{-2\left((b_1-2)^2+(d_1-4)^2\right)} + e^{-2\left((b_1-1)^2+(d_1-3)^2\right)}\right]p^{\ell}(b_2,d_2) \\ &\quad + 7.39\times 10^{-7}p^{\ell}(b_1,d_1)p^{\ell}(b_2,d_2). \end{aligned} \tag{5.6}$$

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. 6; a general description of slicing is given in Rmk. 41.

Remark 41 *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 $j = 2, ..., N$, and views the density on the corresponding hyperplane $W \times \{(b'_2, d'_2)\} \times ... \times \{(b'_N, d'_N)\} \subset W^N$. 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. 1.

If we consider the density evaluated along slices as $K_\sigma\left(((b,d),(1,3)),\mathscr{D}\right)$ or $K_\sigma\left(((b,d),(2,4)),\mathscr{D}\right)$ (Fig. 6 (a) or (b), respectively), the restricted plot is a Gaussian centered at the other upper feature. If the fixed feature is instead close to the diagonal, as in Fig. 6 (c), the density slice is close to a mixture between the two upper Gaussians $p^{(1)}$ and $p^{(2)}$.



(a) (b) (c)

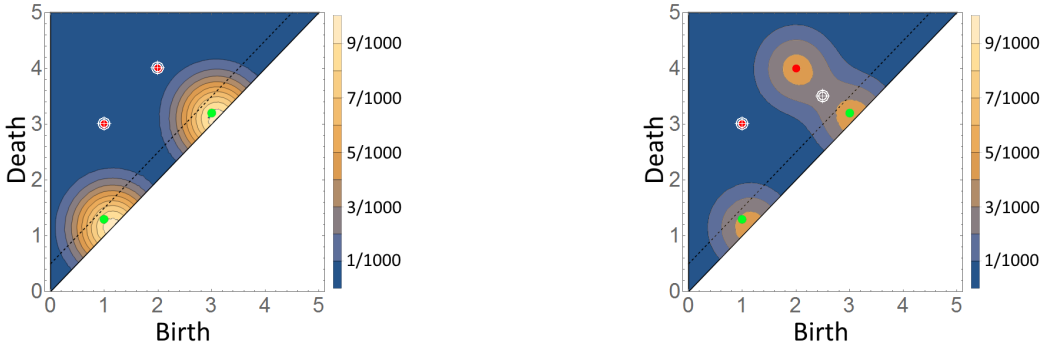
Figure 6: Contour maps for slices of the kernel density $K_\sigma((\xi,\xi_2'),\mathscr{D})$ with input cardinality 2. A single feature ξ_2', indicated by white crosshairs, is fixed to restrict to a 2D subspace as follows: (a) $\xi_2' = (1,3)$ (b) $\xi_2' = (2,4)$ and (c) $\xi_2' = (2.5, 2.7)$. The center 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.

In a similar fashion, we also express the kernel density with input cardinality $|Z| = 3$. Since there are only 2 upper features in $\mathscr{D}$, this and further expressions are not markedly more complicated than Eq. (5.6). From Eq. (4.6), we obtain:

$$\begin{aligned} K_\sigma((\xi_1,\xi_2,\xi_3),\mathscr{D}) &= \nu(1)\left[q^{(1)}q^{(2)}p^{(1)}(b_1,d_1)p^{(2)}(b_2,d_2)\right]p^\ell(b_3,d_3) \\ &\quad + \nu(2)(1-q^{(2)})q^{(1)}p^{(1)}(b_1,d_1)p^\ell(b_2,d_2)p^\ell(b_3,d_3) \\ &\quad + \nu(2)(1-q^{(1)})q^{(2)}p^{(2)}(b_1,d_1)p^\ell(b_2,d_2)p^\ell(b_3,d_3) \\ &\quad + \nu(3)(1-q^{(1)})(1-q^{(2)})p^\ell(b_1,d_1)p^\ell(b_2,d_2)p^\ell(b_3,d_3). \\ &= 9.01\times 10^{-2}p^\ell(b_3,d_3)e^{-2\left((b_1-1)^2+(d_1-3)^2\right)}e^{-2\left((b_2-2)^2+(d_2-4)^2\right)} \\ &\quad + 4.96\times 10^{-4}p^\ell(b_2,d_2)p^\ell(b_3,d_3)e^{-2\left((b_1-2)^2+(d_1-4)^2\right)} \\ &\quad + 4.96\times 10^{-4}p^\ell(b_2,d_2)p^\ell(b_3,d_3)e^{-2\left((b_1-1)^2+(d_1-3)^2\right)} \\ &\quad + 1.22\times 10^{-6}p^\ell(b_1,d_1)p^\ell(b_2,d_2)p^\ell(b_3,d_3). \end{aligned} \tag{5.7}$$

One may notice that Eq. (5.7) has the same 4 terms as Eq. (5.6), but with another factor of p^ℓ 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^ℓ. Contour plot slices of this local kernel are shown in Fig. 7, following Rmk. 41. 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. (5.7), the heaviest

weighted term consists of both upper features' densities as well as the lower density $p^{\ell}(b_3, d_3)$. Indeed, Fig. 7(a) shows the slice $K_\sigma(((b,d),(1,3),(2,4)),\mathscr{D})$, which leaves both upper features fixed, and the resulting slice is nearly proportional to the lower density p^{ℓ}. Fig. 7 (b) shows the slice $K_\sigma(((b,d),(1,3),(2.5,3.5)),\mathscr{D})$, which fixes one of the upper features of $\mathscr{D}$ 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/Π_3 makes the slice look multi-modal, depending on whether $(2.5, 3.5)$ is assigned to $p^{(2)}$ or p^{ℓ}. Other assignments have negligible mass. Thus, Fig. 7 (b) resembles a mixture of these two densities.



(a) (b)

Figure 7: Contour maps for slices of the kernel density $K_\sigma((\xi, \xi_2', \xi_3'), \mathscr{D})$ with input cardinality 3. A pair of features ξ_2' and ξ_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.

The terms $(1 - q^{(k)})$ within the $\mathcal{Q}^*$ expression (see Eq. (3.11)) are very small and appear in terms for which the corresponding upper feature is unassigned. These terms are so small because both upper features have very long persistence in this example (four times the bandwidth), and so the terms in Eqs. (5.5), (5.6), and (5.7) which do not include one or both upper Guassians $p^{(1)}$ 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. 5, 6, 7, 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, $N((0,0), \left(\frac{1}{50}\right)^2 I_2)$. 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.

(a) (b)

Figure 8: 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.

KDE	(1)	(2)	(3)	(4)
n	100	300	1000	5000
σ	0.03	0.025	0.020	0.015

Table 1: Choices of sample size n (number of persistence diagrams) and bandwidth σ for each kernel density estimate $\hat{f}_{n,\sigma}(Z)$ shown in Fig. 9.

8. 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 (σ) as shown in Table 1). The bandwidth was chosen to scale according to Silverman's rule of thumb (Silverman, 1986) (see Rmk. 32).

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 41. For example, $g(\xi) = \hat{f}_{n,\sigma}(\xi, \xi'_2, ..., \xi'_N)$ for fixed ξ'_j $(j = 2, ..., N)$ 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. 9, wherein the same slices (i.e., the same features are fixed) are viewed for each choice of (n, σ). 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. 9 (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. 9 (right). Indeed, the persistence diagram shown in Fig. 8 is typical for this example. Overall, by scanning from top to bottom, Fig. 9 demonstrates the convergence of the KDEs as n increases and σ 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.

Example 4 Electroencephalography (EEG) monitors electrical activity in the brain by measuring changes in voltage over time at particular locations on the scalp. To obtain data, researchers place arrays of electrodes on subjects' heads that record fluctuations in voltage as time series. It is known that EEG in the 1-100 Hz range is heavily involved in cognition (Boothe et al., 2017); moreover, specific bands in the 1-100 Hz range are hypothesized to be associated with certain tasks or brain states. For example, alpha range EEG, which has a spectral peak in the 8-12 Hz range, is thought to be important in inhibition and excitation during decision problems (Kilmesch, 2012). Collecting EEG is a noninvasive procedure, however, measurements are often obscured by noise arising from electrical activity in the environment, movement, or other physiological processes like heartbeat. Consequently, a critical problem is the need to detect EEG with the same underlying dynamics, e.g. EEG that is predominantly composed of 8-12 Hz oscillations, in the presence of varying levels of noise (Repovs, 2010).

In this example, we consider the widely–used autoregressive EEG model introduced in (Franaszczuk and Bilnowaska, 1985), and we employ the KDE established in Proposition 25 to statistically analyze EEG. The authors in (Franaszczuk and Bilnowaska, 1985) model an EEG time series $(x_{t_i})_{i=0}^{L}$ of time length L as a convolution of white noise with a linear filter function given in Eq. (5.8),

$$h(t_i) = \sum_{j=1}^{p} e^{-\beta_j t_i} \cos(\omega_j t_i), \tag{5.8}$$

where ω_j correspond to centers of peaks in the power spectral density of $(x_{t_i})_{i=0}^{L}$ while the parameters β_j are approximately equal to 1/2 of their respective widths (both ω_j and β_j are given in Hz). Recall that the power spectral density describes the contribution of each frequency to the total power of $(x_{t_i})_{i=0}^{L}$ after decomposing $(x_{t_i})_{i=0}^{L}$ into a series of oscillatory functions. For example, a power spectral density with a narrow peak at 10 Hz corresponds to a time series that heavily resembles a function oscillating at 10 Hz. A broader peak at 10 Hz in essence means that the time series has a greater contribution from more frequencies surrounding 10 Hz, diminishing the resemblance (for comparison, the power spectral density of white noise is completely flat). This view of the power spectral density means one can effectively simulate EEG comprised of oscillations in a desired band of frequencies by selecting appropriate parameters in Eq. (5.8).

We focus on alpha range (8-12 Hz) EEG. Specifically, we simulate 200 EEG signals in the alpha range by first generating 200 white noise vectors of length $L = 1,024$ through independent draws from $\mathcal{N}(0, 1)$ then convolving them with the linear filter described by Eq. (5.8) with $p = 1, \beta_1 = 3.7$, and $\omega_1 = 10.5$. We corrupt 100 signals by additive noise $\mathcal{N}(0, 10^{-1/20})$, while the rest are corrupted by $\mathcal{N}(0, 10^{-5/20})$. This yields two collections of EEG signals with signal-to-noise ratios (SNRs) of 1 and 5, denoted by SNR_1 and SNR_5, respectively.

Next, we convert each EEG signal into a persistence diagram using the methodology of (Perea and Harer, 2015). Namely, we transform EEG signals to point clouds in $\mathbb{R}^2$ using delay embeddings where the delay parameter was determined by the sampling rate (100 Hz) along with the dominant underlying frequency of the signals (10 Hz); we then center and scale the point clouds by their variances along the vertical and horizontal axes; see Fig. 10 (e) and (f). Once we obtain point clouds, we compute persistence diagrams for 1-dimensional homological features using Rips

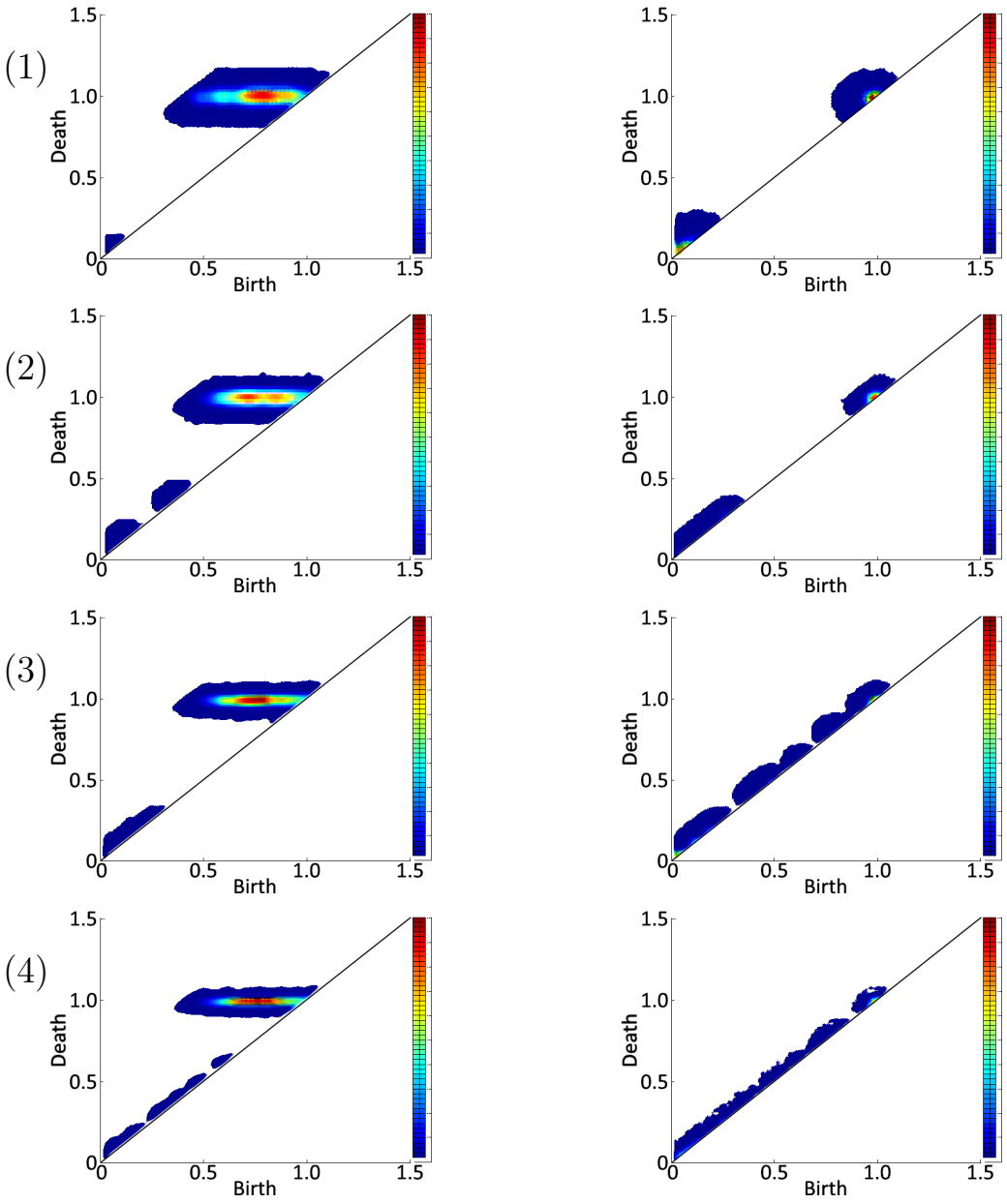


Figure 9: Plots of persistence diagram KDEs for Ex. 3. Each plot is presented as a heat map where color indicates the probability density. White regions above the diagonal indicate portions of very low probability density. Each column is a particular slice, while each row is a particular global KDE with n and σ 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 $\hat{f}_{n,\sigma}((b,d),(0.77,0.98))$ 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.

filtrations; see Section 2. We choose to focus solely on 1-dimensional homological features since they relate to periodicity in the underlying time series, which is the defining characteristic of our signals.

Denote the family of persistence diagrams created from SNR_i for $i = 1, 5$ by D_{SNR_i}, and let f_{SNR_i} be their global probability densities. Our goal is to verify that SNR_1 and SNR_5 EEG have the same underlying dynamics. A sensible strategy is to select a quantity created from persistence diagrams that is robust to noise, approximate its distribution for SNR_i using D_{SNR_i}, and then compare the two empirical distributions. To this end, we start by approximating f_{SNR_i} with the kernel density estimators $\hat{f}_{SNR_i}(Z) := 10^{-2} \sum_{\mathscr{D} \in D_{SNR_i}} K_\sigma(Z, \mathscr{D})$. For each fixed $i = 1, 5$ the persistence diagrams $\mathscr{D} \in D_{SNR_i}$ are the 100 diagrams of each SNR_i case. For the noise likelihood model related to the lower part of a persistence diagram, $\mathscr{D}^\ell$, we use Eq. (5.1) as the cardinality distribution. Given that features with higher persistence generally describe global topology that is more resilient to noise, and relying on these kernels $\hat{f}_{SNR_i}$, we take $S = 1,000$ sample persistence diagrams and compute their bottleneck distance $W_\infty(\emptyset, S_i^j) = \max_{(b,d) \in S_i^j} d - b$, where S_i^j is the jth sample persistence ($j = 1, \ldots, S$) diagram distributed according to $\hat{f}_{SNR_i}$, $i = 1, 5$. These distances create empirical distributions, one for each SNR_i EEG denoted by F_{SNR_i}. We formally proceed with hypothesis testing

$$H_0 : F_{SNR_1} = F_{SNR_5} \text{ vs } H_1 : F_{SNR_1} \neq F_{SNR_5}.$$

Failure to reject H_0 in this case is evidence that D_{SNR_1} and D_{SNR_5} have similar behavior for the features less affected by noise, which in turn implies that SNR_1 and SNR_5 have similar underlying dynamics. Finally, we compare these distributions with a two-sided Kolmogorov-Smirnov (KS) Test (Simard, 2011) that yields a $p-$value=0.72.

	KS-Test P-value	Time (s)
KDE MP	0.72	0.047
PI L_∞	6.15×10^{-9}	0.042
PL L_∞	0.79	0.048

Table 2: The p-values and run times for each method (KDE, PI, and PL) used for the hypothesis test of Eq. (4).

	Sample MAD
SNR_1	1.040
SNR_5	1.035

Table 3: The sample MADs for SNR_1 and SNR_5 computed by taking the means of the distributions in Fig 11(e) and Fig 11(f), repsectively.

For the sake of comparison to other TDA methods, we also compute persistence images (PIs) with resolution 50×50 and spread 0.2 using the ramp function to produce weights, (Adams et al., 2017), and persistence landscapes (PLs) from D_{SNR_i}, (Bubenik, 2015). We examine the L_∞-norm as a summary for each of these vectorizations (the L_∞-norm of the first landscape in particular for PLs) since this measurement is also associated with high persistence features. After computing L_∞-norms for each of the PIs and PLs obtained from D_{SNR_i}, we resample each L_∞ empirical distribution 1,000 times to create bootstrapped distributions with size matching those of the W_∞

distributions obtained from the kernel density estimators; see Fig. 11 (c),(d),(e), and (f). In the end, we also compare the bootstrapped distributions with a two-sided KS-test. Table 2 shows the KS-test p-values and a standardized run time for each method.

Notice the kernel density max persistence and landscape L_∞ correctly fail to reject H_0 at the most commonly used significance levels ($p-$value $= 0.79$). In particular, our method is competitive with landscapes (with a slight edge on computational time). On the other hand, the persistence image L_∞ incorrectly rejects H_0 ($p-$value close to 0). Failure of PIs to recognize different dynamics may be a result of the fact that in addition to accounting for the max persistence (through the use of the ramp function for weights), the PI L_∞ also considers the cardinality of each diagram, although the contribution of cardinality diminishes for higher resolutions and smaller spreads.

Finally, we report estimates for $MAD_{f_{SNR_1}}(\emptyset)$ and $MAD_{f_{SNR_5}}(\emptyset)$ by taking the means of our empirical distributions for the max persistences; see Table 3. Notice the estimates are very close numerically and by appealing to Theorem 40, one could argue they are close to their true values. Hence, the MAD offers more evidence that SNR_1 and SNR_5 are statistically indistinguishable.

6. 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. Our kernel density function arises by creating a noise model for persistence diagrams which simplifies treatment of features near the diagonal. Consequently, the kernel density can also be used in measuring likelihood when observing small perturbations of persistence diagrams. 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 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. 9 (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, examples of kernel density estimation are presented in Section 5. In one 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. Moreover, our example with EEG data verifies that our kernel density estimator is accessible to data in a realistic setting while remaining competitive with or outperforming pre-existing TDA techniques.

In the context of Fig. 3, 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 $\mathscr{D}$. In contrast, precise evaluation of the kernel global pdf at a diagram requires the more thorough computations shown in Eq. (4.6). This evaluation is made tractable due to the separation of the center diagram into upper and lower portions: $\mathscr{D} = \mathscr{D}^u \cup \mathscr{D}^\ell$ as described in Eq. (4.1). In practice, diagrams should split so that $|\mathscr{D}^u|$ is small while $\left|\mathscr{D}^\ell\right|$ is large. Evaluation of individual feature pdfs on the multi-wedge $\mathcal{W}_{0:\mathrm{d}-1}$ only scales quadratically on the cardinality $|\mathscr{D}|$ and higher degree calculations are required only for combinatorics on the large persistence features in the upper diagram $\mathscr{D}^u$. Consequently, these calculations are tractable so long as $\mathscr{D}^u$ does not grow too much in cardinality, while an increased cardinality for $\mathscr{D}^\ell$ 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 (Theorem 31) 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.

Acknowledgements

The authors would like to thank 3 anonymous reviewers for their comments, which improved the manuscript significantly. Moreover, we would like to thank Drs. David Boothe, Piotr Franasczcuk, and Jason Sherwin for their useful insight and discussions about EEG data. Research has been partially supported by the Army Research Office (ARO) Grant # W911NF-17-1-0313, the National Science Foundation (NSF) Grant # DMS-1821241, and the Thor Industries/Army Research Lab (ARL) Grant # W911NF-17-2-0141.

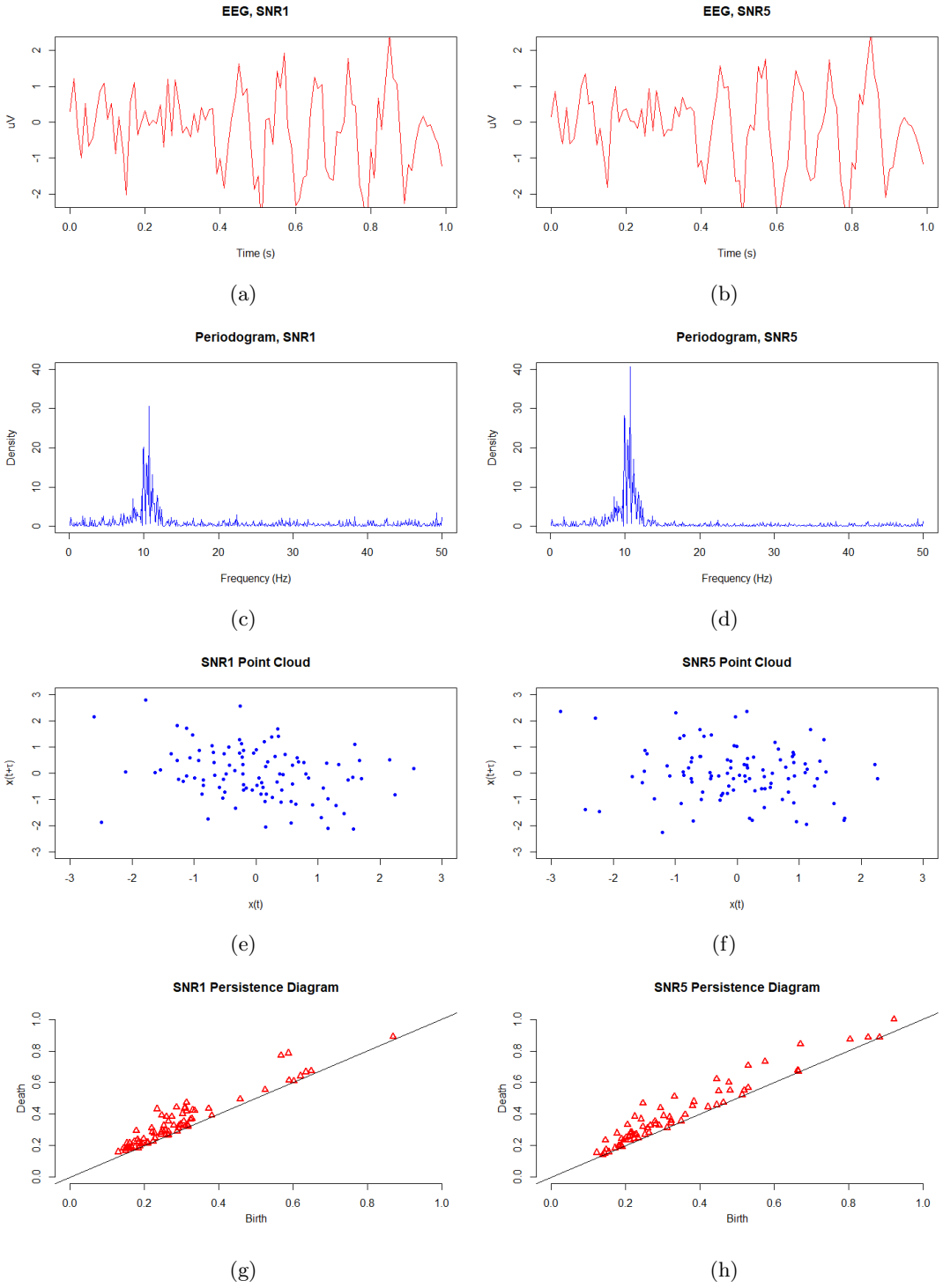


Figure 10: A segment of an EEG signal with (a) SNR_1, and (b) SNR_5, respectively, along with (c) and (d): their corresponding periodograms (estimates of the power spectral densities). The associated point clouds are given in (e) and (f), respectively, and (g) and (h) their resulting persistence diagrams.

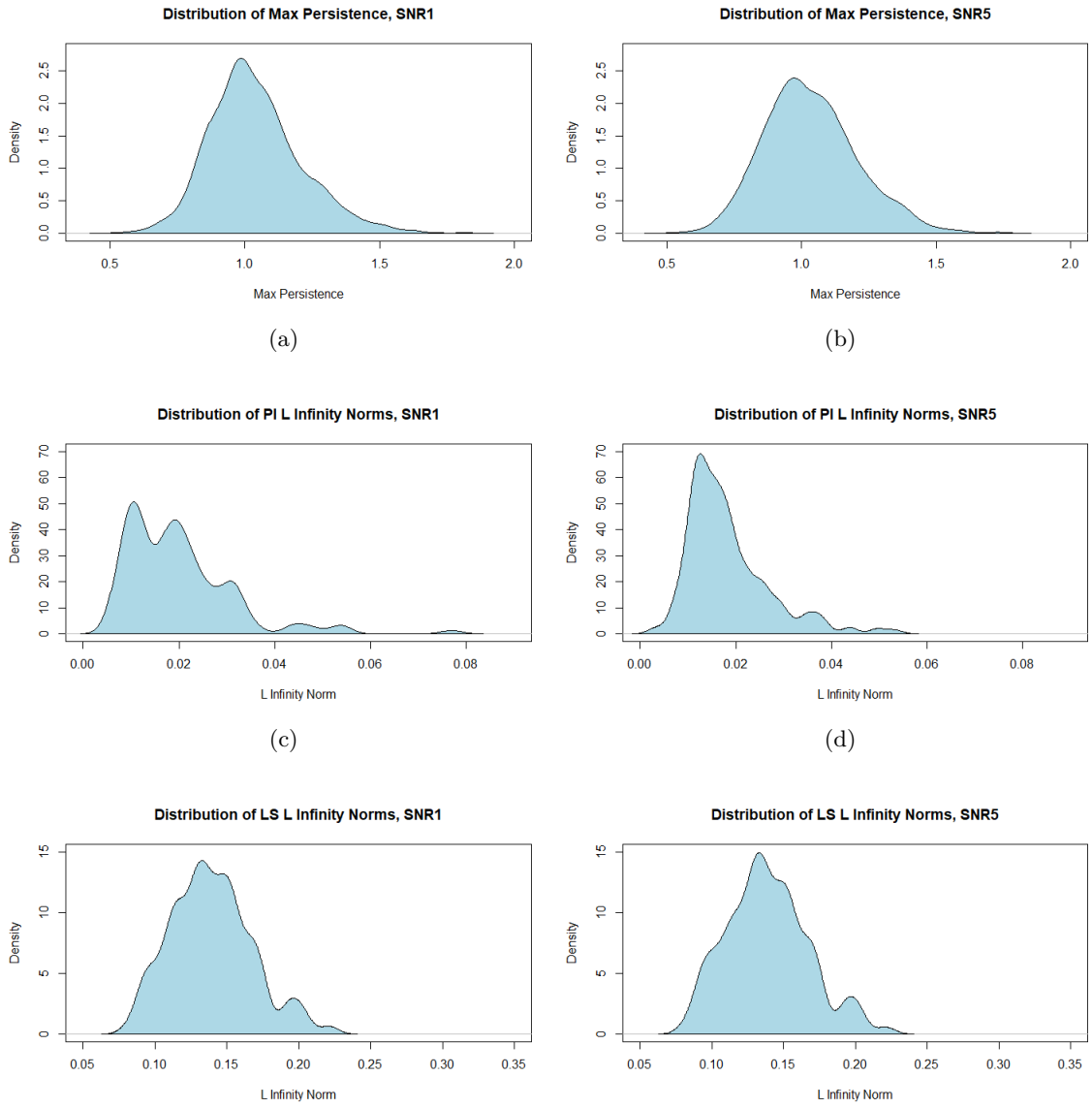


Figure 11: This figure shows the distribution for each statistic we considered when comparing SNR_1 to SNR_5. Each column represents a class, either SNR_1 or SNR_5, and each row a particular statistic.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L.(2015). Persistence images: A stable vector representation of persistent homology *The Journal of Machine Learning Research* 18(1):218–252.

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*.

Boothe, D., Yu, A., Kudela, P., Anderson, W., Vettel, J., and Franaszczuk, P.(2017). Impact of neuronal membrane damage on local field potential in large-scale simulation of cerebral cortex. *Frontiers in Neurology*, volume 8, pages 236

Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research* 16(1):77-102.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015). Subsampling methods for persistent homology. In *International Conference on Machine Learning* (pp. 2143–2151).

Chazal, F. and Divol, V. (2018). The density of expected persistence diagrams and its kernel based estimation. In *The Symposium of Computational Geometry 2018*

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 l 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.

FRANASZCZUK, P. AND BILNOWASKA, K. Linear model of brainactivity-EEG as a superposition of damped oscillatory modes *Biological Cybernetics* 53:19-25 Spring-Verlag

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.

GUILLEMARD, 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.

Kilmesch, W.(2012). Alpha band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606-617

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016). Persistence weighted Gaussian kernel for topological data analysis. In *Proceedings of the* 33^{rd} *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.

Mahler, R. P. (2003). Multi-target Bayes filtering via first-order multi-target moments. In *IEEE Trans. AES*, volume 37, pages 1152-1178.

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

Marchese, A. and Maroulas, V. (2018). Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification*, 12 (3), pp. 657–682.

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.

Maroulas, V., Micucci, C., and Spannaus, A. (2018). A stable cardinality distance for topological classification. https://arxiv.org/abs/1812.01664.pdf

Maroulas, V., Nasrin, F., and Oballe, C. (2019). A Bayesian Framework for Persistent Homology https://arxiv.org/pdf/1901.02034.pdf

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.

Pereira, 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*

Repovs, G.(2010). Dealing with noise in EEG recording and data analysis. *Informatica Medica Slovenica*, 15(1):18-25

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.

Simard, R. and L'Ecuyer, P.(2011). *Computing the two-sided Kolmogorov Smirnov distribution.* Journal of Statistical Software 39(11):1-18

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.

Appendix A. Proofs from Section 4.3

A.1. Proof of Proposition 38

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 $W_\infty(D, \mathscr{D}) < \delta\sigma$, it is sufficient that each constituent feature does not stray too far from either its corresponding center or the diagonal (see Fig. 3 for reference). Specifically, we follow Def. 5 to build a correspondence between D and $\mathscr{D}$ 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 ζ while features in $\mathscr{D}$ are denoted using ξ.

Consider each feature $\xi^j \in \mathscr{D}^u = \mathscr{D} \cap \{d - b \geq \sigma\}$ and its associated random singleton diagram $D^j = \{\zeta^j\}$ or $\emptyset$ as in Def. 22. 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 $\zeta^j \in B_j$ (or equivalently $z^j \in B((0,0),\delta)$). Thus, we obtain $\mathbb{P}\left[\zeta^j \in B(\xi^j, \delta\sigma)\right] \geq \mathbb{P}\left[|z| \leq \delta\right]$ for the probability that ζ^j can be mapped to ξ^j in a bounding correspondence. If $B_j \nsubseteq W$, this probability is even higher because ξ^j can be mapped to the diagonal and thus the case $D^j = \emptyset$ is included.

Now take into account the features in $\mathscr{D}^\ell$ and the associated random features D^ℓ as in Def. 24. Although the features in D^ℓ 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^ℓ 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^ℓ and features in $\mathscr{D}^\ell$ are close enough to the diagonal (within $\delta\sigma$). Indeed, the features in $\mathscr{D}^\ell$ 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 $p^\ell(b,d) = \frac{1}{\pi N_\ell \sigma^2} \sum_{j=1}^{N_\ell} e^{-\left(\left(x - \frac{b_j+d_j}{2}\right)^2 + \left(y - \frac{b_j+d_j}{2}\right)^2\right)/2\sigma^2}$. Consider the sets $U_j = B\left(\left(\frac{b_j+d_j}{2}, \frac{b_j+d_j}{2}\right), \delta\sigma\right)$ and $\mathcal{U} = \bigcup_{j=1}^{N_\ell} U_j$. 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 $\mathbb{P}[d - b \leq \delta\sigma] = \int_{\Delta_0^{\delta\sigma}} p^\ell(x,y)\,dx\,dy \geq \int_{W \cap \mathcal{U}} p^\ell(x,y)\,dx\,dy$ since $W \cap U \subset \Delta_0^{\delta\sigma} = \{(b,d) \in W : d - b \leq \delta\sigma\}$. Next, we restrict the lower bounding integral for each term of p^ℓ to its matching subset U_j and change variables to attain the desired form:

$$\begin{aligned}\int_{W\cap\mathcal{U}} p^\ell(x,y)\,dx\,dy &\geq \sum_{j=1}^{N_\ell} \int_{U_j} \frac{1}{2\pi N_\ell \sigma^2} e^{-\left(\left(x - \frac{b_j+d_j}{2}\right)^2 + \left(y - \frac{b_j+d_j}{2}\right)^2\right)/2\sigma^2}\,dx\,dy \\ &= \int_{B((0,0),\delta)} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\,dx\,dy.\end{aligned}$$

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

A.2. Proof of Lemma 39

Choose an arbitrary persistence diagram $\mathscr{D}$. Since bottleneck distance is defined according to the sup-norm (see Eq. (2.2)), 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 $Z = \{\xi_1, ..., \xi_N\}$ with $\xi_i = (b_i, d_i, k_i)$, we have:

$$\int_{\mathcal{W}_{0:\mathrm{cl}-1}} \max(d_i - b_i)\delta Z \leq \int_{\mathcal{W}_{0:\mathrm{cl}-1}} \|Z\|\, f(Z)\delta Z \tag{A.1}$$

since $\max(d_i - b_i) \leq \max\left(\|(b_i, d_i)\|\right) \leq \|Z\|$. Consider a compact set $K \subset \mathcal{W}_{0:\mathrm{cl}-1}$ which contains a neighborhood of the origin. Given assumptions $(A2)^*$ and $(A3)^*$, Eq. (A.1) is bounded by the following finite expression.

$$\int_{\mathcal{W}_{0:\mathrm{cl}-1}} \|Z\|\, f(Z)\delta Z \leq \int_K C_2 \|Z\|\, \delta Z + \sum_{N=1}^{M} \int_{h_N^{-1}(h_N(K)^c)} C_3 \|Z\|^{-2N-1}\, d\xi_1 ... d\xi_N. \tag{A.2}$$

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 $\mathscr{D}_0$ is bounded

as follows.

$$\int_{\mathcal{W}_{0:\mathfrak{cl}-1}} W_\infty(\mathscr{D}_0, Z) f(Z) \delta Z \leq \int_{\mathcal{W}_{0:\mathfrak{cl}-1}} W_\infty(\mathscr{D}_0, \emptyset) f(Z) \delta Z + \int_{\mathcal{W}_{0:\mathfrak{cl}-1}} W_\infty(\emptyset, Z) f(Z) \delta Z \qquad \text{(A.3)}$$

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 $W_\infty(\mathscr{D}_0, \emptyset)$, while the second integral is finite according to Eq. (A.2).

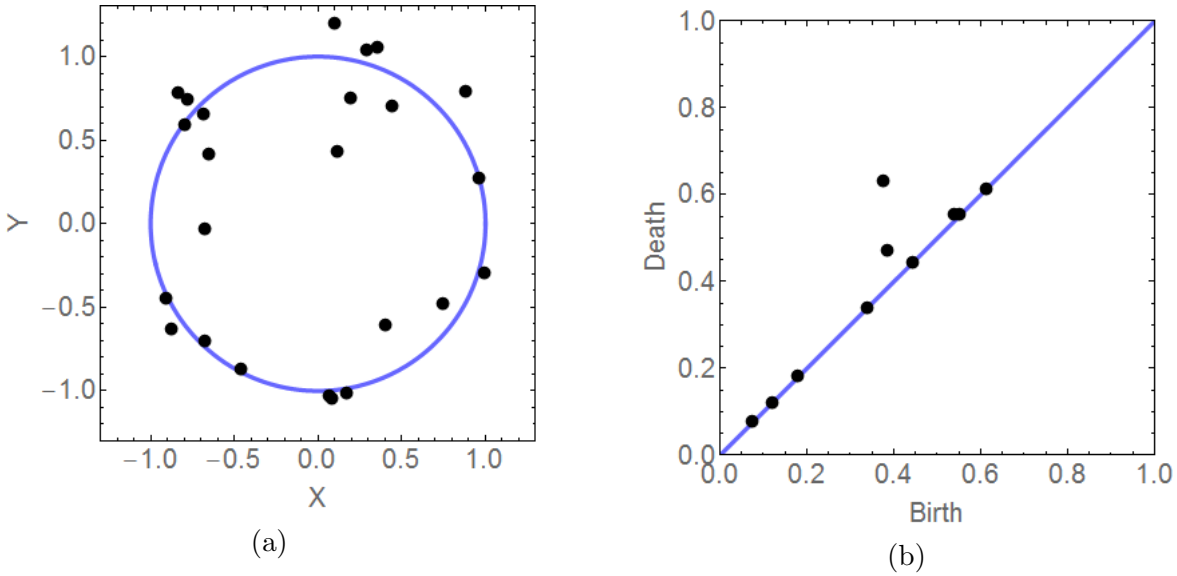
Appendix B. Extra Examples

Here we present two more examples of constructing a kernel density estimator (KDE) according to the kernel given in Eq. (4.6). 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 5. 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 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. 41). This allows one to track the convergence of the KDE as the sample size of persistence diagrams, n, increases and the bandwidth, σ, decreases.

Example 5 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. 12.

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. 41. For example, $g(\xi) = \hat{f}_{n,\sigma}(\xi, \xi'_2, ..., \xi'_N)$ for fixed ξ'_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. 13, wherein the same slices are viewed for each choice of n and σ. 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. 13 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 $\hat{f}(\xi, \xi'_j)$.



(a) (b)

Figure 12: (a) An example of the underlying datasets generated for Ex. 5. 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

While the underlying dataspace is the unit circle in both Ex. 3 and Ex. 5, the precise presentation of the underlying data effects 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' 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. 9 is $d = 0.98 \approx 1$). 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. 9 is $b = 0.77 >> 0$).

In comparison to Ex. 3, Ex. 5 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. 13 is $d = 0.8 < 0.98$). It is evident from Fig. 13 that while the noise is additive on the underlying data, its precise effect on the random persistence diagram is nonlinear. Moreover, Ex. 5 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. 13 is $b = 0.4 < 0.77$). In addition, larger noise and sample size both result in more features near the diagonal in Ex. 5 as compared to Ex. 3.

Example 6 While Ex. 5 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

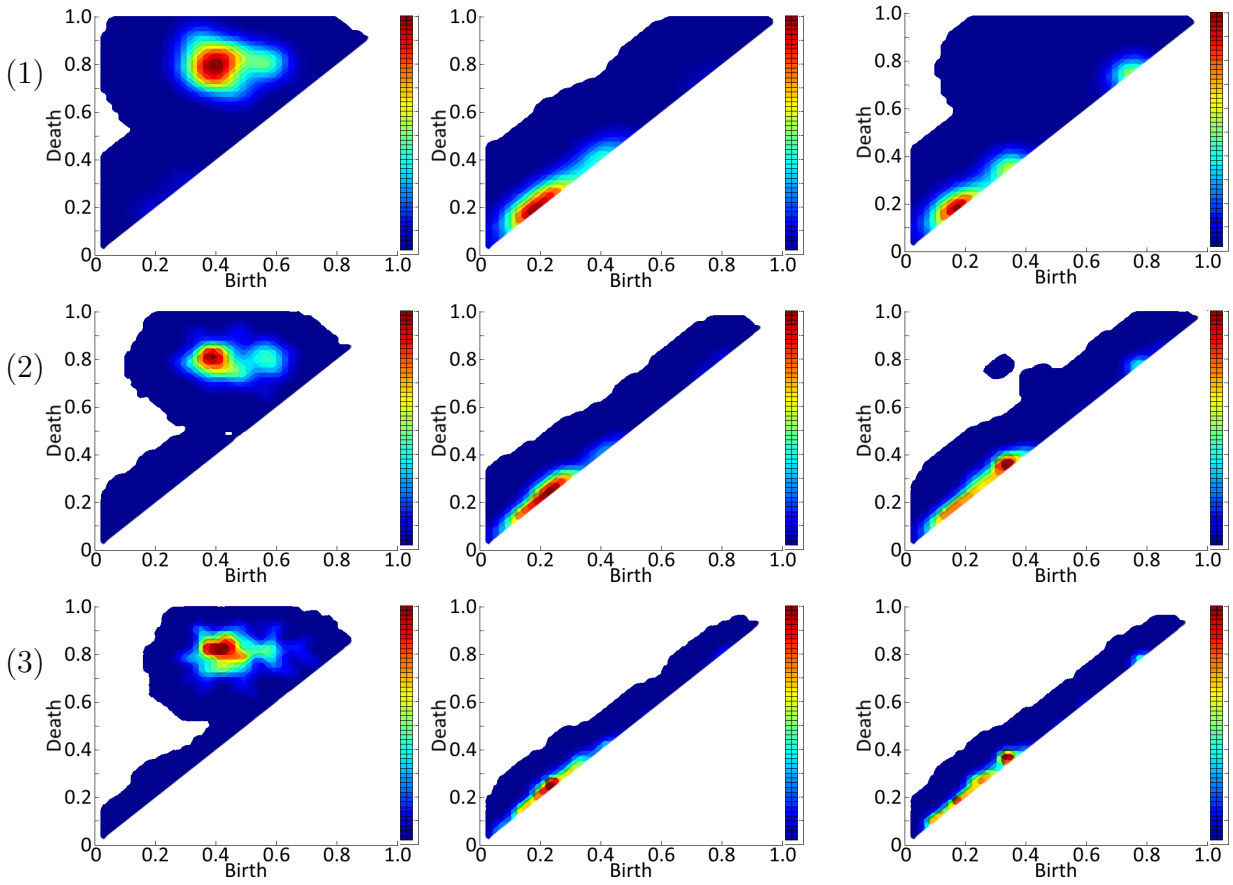
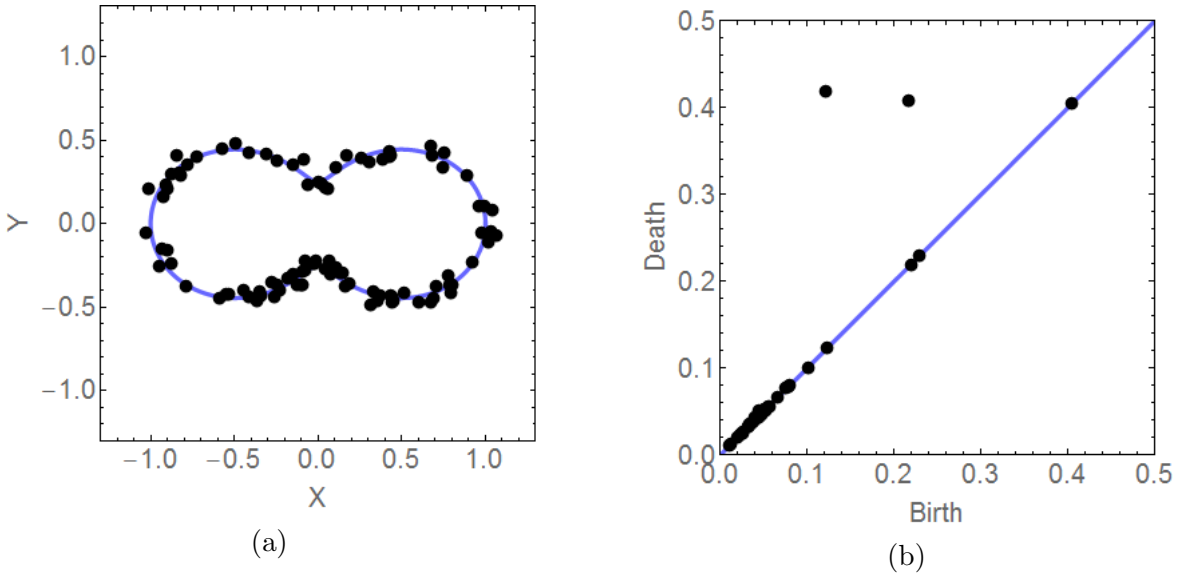


Figure 13: Key slices of persistence diagram KDEs for Ex. 5. Each column is a particular slice, while each row is a particular KDE: (1) $n = 20$ and $\sigma = 0.05$, (2) $n = 100$ and $\sigma = 0.03$, and (3) $n = 300$ and $\sigma = 0.02$. The first column are the local KDEs $\hat{f}_{n,\sigma}((b, d))$ 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.



(a) (b)

Figure 14: (a) An example of the underlying datasets generated for Ex. 5. 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.

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 persistence diagram for $k = 1$ are shown in Fig. 14.

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. 41. For example, $g(\xi) = \hat{f}_{n,\sigma}(\xi, \xi_2', ..., \xi_N')$ for fixed ξ_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. 15, wherein the same slices are viewed for each choice of n and σ. 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. 15 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. 15, 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.

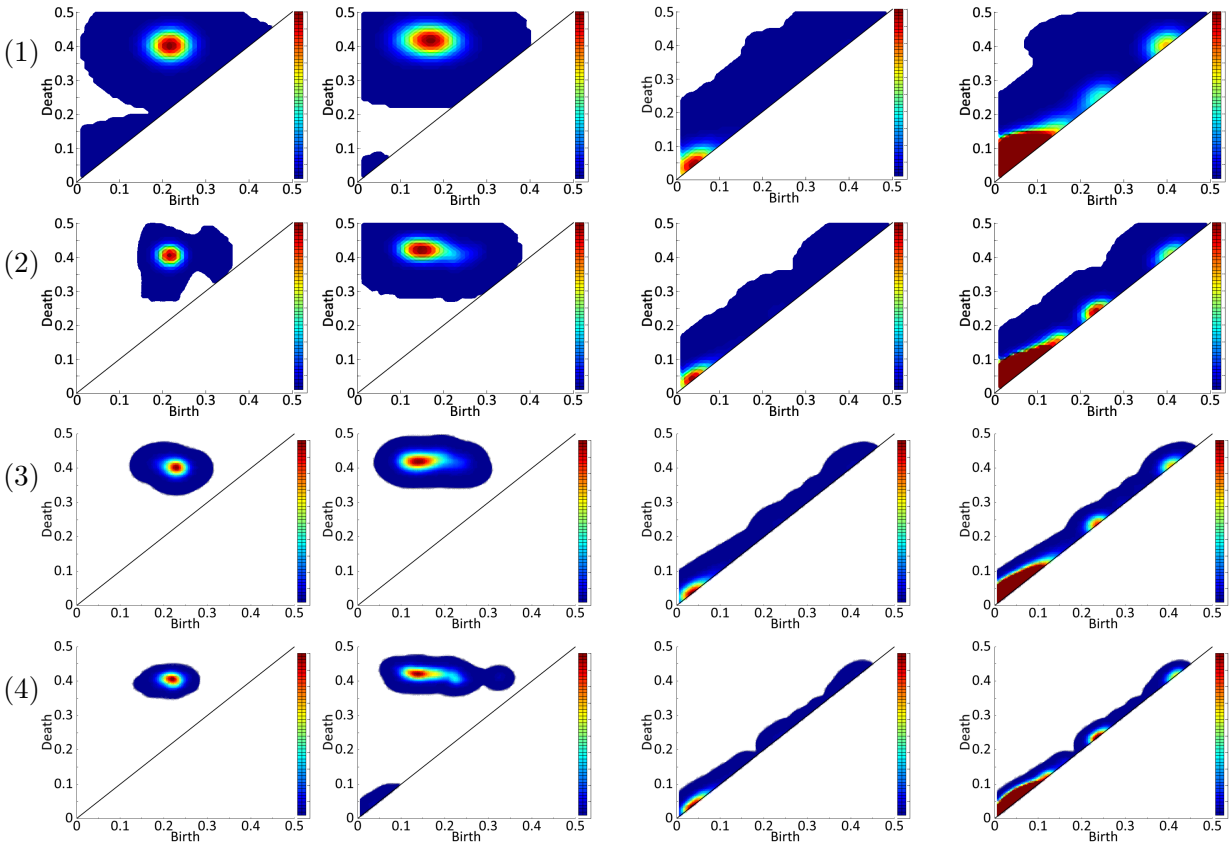


Figure 15: Key slices of persistence diagram KDEs for Ex. 6. Each column is a particular slice, while each row is a particular KDE (1) $n = 20$ and $\sigma = 0.03$, (2) $n = 100$ and $\sigma = 0.02$, (3) $n = 300$ and $\sigma = 0.015$, and (4) $n = 1000$ and $\sigma = 0.01$. The first column are the 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.

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.