Robust Persistence Diagrams using Reproducing Kernels

Siddharth Vishwanath[†], Kenji Fukumizu[‡*],
Satoshi Kuriki[‡*], and Bharath Sriperumbudur[†*]

[†]*Department of Statistics, The Pennsylvania State University, University Park, PA, USA*
[‡]*The Institute of Statistical Mathematics, Tokyo, Japan*

June 18, 2020

Abstract

Persistent homology has become an important tool for extracting geometric and topological features from data, whose multi-scale features are summarized in a persistence diagram. From a statistical perspective, however, persistence diagrams are very sensitive to perturbations in the input space. In this work, we develop a framework for constructing robust persistence diagrams from superlevel filtrations of robust density estimators constructed using reproducing kernels. Using an analogue of the influence function on the space of persistence diagrams, we establish the proposed framework to be less sensitive to outliers. The robust persistence diagrams are shown to be consistent estimators in bottleneck distance, with the convergence rate controlled by the smoothness of the kernel—this in turn allows us to construct uniform confidence bands in the space of persistence diagrams. Finally, we demonstrate the superiority of the proposed approach on benchmark datasets.

1 Introduction

Given a set of points $\mathbb{X}_n = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ observed from a probability distribution $\mathbb{P}$ on an input space $\mathcal{X} \subseteq \mathbb{R}^d$, understanding the shape of $\mathbb{X}_n$ sheds important insights on low-dimensional geometric and topological features which underlie $\mathbb{P}$, and this question has received increasing attention in the past few decades. To this end, Topological Data Analysis (TDA), with a special emphasis on persistent homology (Edelsbrunner et al., 2000; Zomorodian and Carlsson, 2005), has become a mainstay for extracting the shape information from data. In statistics and machine-learning, persistent homology has facilitated the development of novel methodology (e.g., Chazal et al. 2013; Chen et al. 2019; Brüel-Gabrielsson et al. 2018), which has been widely used in a variety of applications dealing with massive, unconventional forms of data (e.g., Gameiro et al. 2015; Bendich et al. 2016; Xu et al. 2019).

Informally speaking, persistent homology detects the presence of topological features across a range of resolutions by examining a nested sequence of spaces, typically referred to as a *filtration*. The filtration encodes the birth and death of topological features as the resolution varies, and is presented in the form of a concise representation—a persistence diagram or barcode. In the context of data-analysis, there are two different methods for obtaining filtrations. The first is computed from the pairwise Euclidean distances of $\mathbb{X}_n$, such as the Vietoris-Rips, Čech, and Alpha filtrations

*Authors arranged alphabetically

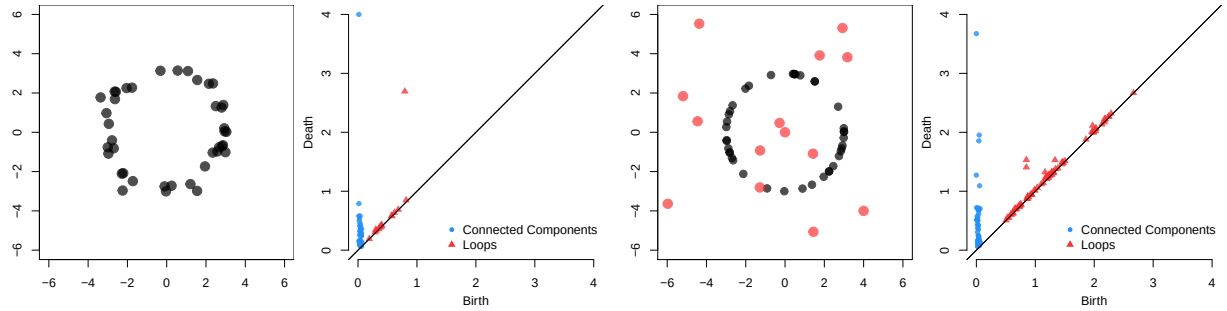


Figure 1: (Left) $\mathbb{X}_n$ is sampled from a circle with small perturbations to each point. The persistence diagram detects the presence of the loop, as guaranteed by the stability of persistence diagrams Chazal et al. (2016); Cohen-Steiner et al. (2007). (Right) $\mathbb{X}_n$ is sampled from a circle but with just a few outliers. The resulting persistence diagram changes dramatically — the persistence of the main loop plummets, and other spurious loops appear, as elaborated in Section 2.

Edelsbrunner et al. (2000). The second approach is based on choosing a function on $\mathcal{X}$ that reflects the density of $\mathbb{P}$ (or its approximation based on $\mathbb{X}_n$), and, then, constructing a filtration. While the two approaches explore the topological features governing $\mathbb{P}$ in different ways, in essence, they generate equivalent insights.

Despite obvious advantages, the adoption of persistent homology in mainstream statistical methodology is still limited. An important limitation among others, in the statistical context, is that the resulting persistent homology is highly sensitive to outliers. While the stability results of Chazal et al. (2016); Cohen-Steiner et al. (2007) guarantee that small perturbations on all of $\mathbb{X}_n$ induce only small changes in the resulting persistence diagrams, a more pathological issue arises when a small fraction of $\mathbb{X}_n$ is subject to very large perturbations. Figure 1 illustrates how inference from persistence diagrams can change dramatically when $\mathbb{X}_n$ is contaminated with only a few outliers. Another challenge is the mathematical difficulty in performing sensitivity analysis in a formal statistical context. Since the space of persistence diagrams have an unusual mathematical structure, it falls victim to issues such as non-uniqueness of Fréchet means and bounded curvature of geodesics (Mileyko et al., 2011; Turner et al., 2014; Divol and Lacombe, 2019). With this background, the *central objective* of this paper is to develop outlier robust persistence diagrams, develop a framework for examining the sensitivity of the resulting persistence diagrams to noise, and establish statistical convergence guarantees. To the best of our knowledge, not much work has been carried out in this direction, except for Bendich et al. (2011) where robust persistence diagrams are constructed from Vietoris-Rips or Čech filtrations on $\mathbb{X}_n$ by replacing the Euclidean distance with diffusion distance. However, no sensitivity analysis of the resultant diagrams are carried out in (Bendich et al., 2011) to demonstrate their robustness.

Contributions. The main contributions of this work are threefold. 1) We propose robust persistence diagrams constructed from filtrations induced by an RKHS-based robust KDE (kernel density estimator) Kim and Scott (2012) of the underlying density function of $\mathbb{P}$ (Section 3). While this idea of inducing filtrations by an appropriate function—(Fasy et al., 2014; Chazal et al., 2017; Phillips et al., 2015) use KDE, distance-to-measure (DTM) and kernel distance (KDist), respectively—has already been explored, we show the corresponding persistence diagrams to be less robust compared to our proposal. 2) In Section 4.1, we generalize the notions of *influence function* and *gross error sensitivity*—which are usually defined for normed spaces—to the space of persistence diagrams, which lack the vector space structure. Using these generalized notions, we investigate the sensitivity of persistence diagrams constructed from filtrations induced by different functions (e.g., KDE, robust KDE, DTM) and demonstrate the robustness of the proposed method, both theoretically (Remark 4.1)

and numerically (Section 5). 3) We establish the statistical consistency of the proposed robust persistence diagrams and provide uniform confidence bands by deriving exponential concentration bounds for the uniform deviation of the robust KDE (Section 4.2).

Definitions and Notations. For a metric space $(\mathcal{X}, \varrho)$, the ball of radius r centered at $\boldsymbol{x} \in \mathcal{X}$ is denoted by $B_{\mathcal{X}}(\boldsymbol{x}, r)$. $L_p(\mathcal{X}, \mu)$ is the Banach space of functions of p^{th}-power μ-integrable functions with norm $\|\cdot\|_p$, where μ is a Borel measure defined on $\mathcal{X}$. $\mathcal{P}(\mathbb{R}^d)$ is the set of all Borel probability measures on $\mathbb{R}^d$, and $\mathcal{M}(\mathbb{R}^d)$ denotes the set of probability measures on $\mathbb{R}^d$ with compact support and *tame*, bounded density function. For bandwidth $\sigma > 0$, $\mathcal{H}_\sigma$ denotes a reproducing kernel Hilbert space (RKHS) with $K_\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as its reproducing kernel. We assume that K_σ is radial, i.e., $K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \sigma^{-d}\psi(\|\boldsymbol{x} - \boldsymbol{y}\|_2/\sigma)$ with $\psi(\|\cdot\|_2)$ being a probability density function on $\mathbb{R}^d$, where $\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^d x_i^2$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We denote $\|K_\sigma\|_\infty \doteq \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d} K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \sigma^{-d}\psi(0)$. For $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$, $\mu_{\mathbb{P}} \doteq \int K_\sigma(\cdot, \boldsymbol{y})d\mathbb{P}(\boldsymbol{y}) \in \mathcal{H}_\sigma$ is called the mean embedding of $\mathbb{P}$, and $\mathcal{D}_\sigma \doteq \{\mu_{\mathbb{P}} : \mathbb{P} \in \mathcal{P}(\mathbb{R}^d)\}$ is the space of all mean embeddings. $\delta_{\boldsymbol{x}}$ denotes a Dirac measure at $\boldsymbol{x}$.

2 Persistent Homology: Preliminaries

We present the necessary background on persistent homology for completeness. See Chazal and Michel (2017); Wasserman (2018) for a comprehensive introduction.

Persistent Homology. Let $\phi : \mathcal{X} \to \mathbb{R}$ be a non-negative function on the metric space $(\mathcal{X}, d)$. At level $r > 0$, the *sublevel* set $\mathcal{X}_r = \phi^{-1}([0, r]) = \{\boldsymbol{x} \in \mathcal{X} : \phi(\boldsymbol{x}) \le r\}$ encodes the topological information in $\mathcal{X}$. For $0 \le r < s \le \infty$, the sublevel sets are nested, i.e., $\mathcal{X}_r \subseteq \mathcal{X}_s$. The sequence $\{\mathcal{X}_r\}_{0\le r\le\infty}$ is a nested sequence of topological spaces, called a *filtration*, denoted by $\mathrm{Sub}(\phi)$, and ϕ is called the *filter function*. As the level r varies, the evolution of the topology is captured in the filtration. Roughly speaking, new cycles (i.e., connected components, loops, voids and higher order analogues) can appear or existing cycles can merge. A new k-dimensional feature is said to be born at $b \in \mathbb{R}$ when a nontrivial k-cycle appears in $\mathcal{X}_b$. The same k-cycle dies at level $d > b$ when it disappears in all $\mathcal{X}_{d+\epsilon}$ for $\epsilon > 0$. Persistent homology, $PH_*(\phi)$, is an algebraic module which tracks the *persistence pairs* (b, d) of births b and deaths d across the entire filtration $\mathrm{Sub}(\phi)$. Mutatis mutandis, a similar notion holds for superlevel sets $\mathcal{X}^r = \phi^{-1}([r, \infty))$, inducing the filtration $\mathrm{Sup}(\phi)$. For $r < s$, the inclusion $\mathcal{X}^r \supseteq \mathcal{X}^s$ is reversed and a cycle born at b dies at a level $d < b$, resulting in the persistence pair (d, b) instead. Figure 2 shows three connected components in the superlevel set for $r = 8$. The components were born as r swept through the blue points, and die when r approaches the red points. We refer the reader to Appendix B for more details. Figure 2 is described in more detail in Figure 9.

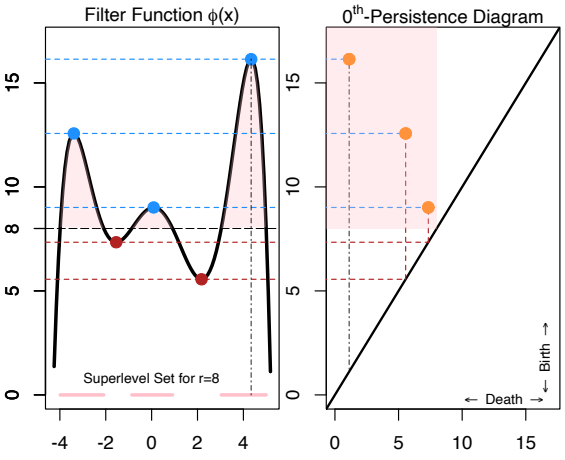


Figure 2: $\mathsf{Dgm}(\mathrm{Sup}(\phi))$ for $\phi : \mathbb{R} \to \mathbb{R}$

Persistence Diagrams. By collecting all persistence pairs, the persistent homology is concisely represented as a persistence diagram $\mathsf{Dgm}(\mathrm{Sub}(\phi)) \doteq \{(b, d) \in \mathbb{R}^2 : 0 \le b < d \le \infty\}$. A similar definition carries over to $\mathsf{Dgm}(\mathrm{Sup}(\phi))$, using (d, b) instead. See Figure 2 for an illustration. When

the context is clear, we drop the reference to the filtration and simply write $\mathsf{Dgm}(\phi)$. The k^{th} persistence diagram is the subset of $\mathsf{Dgm}(\phi)$ corresponding to the k-dimensional features. The space of persistence diagrams is the locally-finite multiset of points on $\Omega = \{(x, y) : 0 \leq x < y \leq \infty\}$, endowed with the family of p-Wasserstein metrics W_p, for $1 \leq p \leq \infty$. We refer the reader to Edelsbrunner and Harer (2010); Divol and Lacombe (2019) for a thorough introduction. W_∞ is commonly referred to as the *bottleneck distance*.

Definition 2.1. *Given two persistence diagrams D_1 and D_2, the bottleneck distance is given by*

$$W_\infty(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1 \cup \Delta} \|p - \gamma(p)\|_\infty,$$

where $\Gamma = \{\gamma : D_1 \cup \Delta \to D_2 \cup \Delta\}$ is the set of all bijections from D_1 to D_2, including the diagonal $\Delta = \{(x, y) \in \mathbb{R}^2 : 0 \leq x = y \leq \infty\}$ with infinite multiplicity.

An assumption we make at the outset is that the filter function f is *tame*. Tameness is a metric regularity condition which ensures that the number of points on the persistence diagrams are finite, and, in addition, the number of nontrivial cycles which share identical persistence pairings are also finite. Tame functions satisfy the celebrated stability property w.r.t. the bottleneck distance.

Proposition 2.1 (Stability of Persistence Diagrams Cohen-Steiner et al., 2007; Chazal et al., 2016)**.** *Given two tame functions $f, g : \mathcal{X} \to \mathbb{R}$,*

$$W_\infty(\mathsf{Dgm}(f), \mathsf{Dgm}(g)) \leq \|f - g\|_\infty.$$

The space of persistence diagrams is, in general, challenging to work with. However, the stability property provides a handle on the persistence space through the function space of filter functions.

3 Robust Persistence Diagrams

Given $\mathbb{X}_n = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\} \subseteq \mathbb{R}^d$ drawn iid from a probability distribution $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$ with density f, the corresponding persistence diagram can be obtained by considering a filter function $\phi_n : \mathbb{R}^d \to \mathbb{R}$, constructed from $\mathbb{X}_n$ as an approximation to its population analogue, $\phi_\mathbb{P} : \mathbb{R}^d \to \mathbb{R}$, that carries the topological information of $\mathbb{P}$. Commonly used $\phi_\mathbb{P}$ include the (i) kernelized density, f_σ, (ii) Kernel Distance (KDist), $d_\mathbb{P}^{K_\sigma}$, and (iii) distance-to-measure (DTM), $d_{\mathbb{P},m}$, which are defined as:

$$f_\sigma(\boldsymbol{x}) \doteq \int_\mathcal{X} K_\sigma(\boldsymbol{x}, \boldsymbol{y}) d\mathbb{P}(\boldsymbol{y}) \; ; \qquad d_\mathbb{P}^{K_\sigma} \doteq \|\mu_{\delta_{\boldsymbol{x}}} - \mu_\mathbb{P}\|_{\mathcal{H}_\sigma} \; ; \qquad d_{\mathbb{P},m}(\boldsymbol{x}) \doteq \sqrt{\frac{1}{m} \int_0^m F_{\boldsymbol{x}}^{-1}(u) du},$$

where $F_{\boldsymbol{x}}(t) = \mathbb{P}(\|\boldsymbol{X} - \boldsymbol{x}\|_2 \leq t)$ and $\sigma, m > 0$. For these $\phi_\mathbb{P}$, the corresponding empirical analogues, ϕ_n, are constructed by replacing $\mathbb{P}$ with the empirical measure, $\mathbb{P}_n \doteq \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{X}_i}$. For example, the empirical analogue of f_σ is the familiar kernel density estimator (KDE), $f_\sigma^n = \frac{1}{n} \sum_{i=1}^n K_\sigma(\cdot, \boldsymbol{X}_i)$. While KDE and KDist capture the topological information of supp($\mathbb{P}$) by approximating the density f (the sublevel sets of KDist are simply rescaled versions of the superlevel sets of KDE (Phillips et al., 2015; Chazal et al., 2017)), DTM, on the other hand, approximates the distance function to supp($\mathbb{P}$).

Since ϕ_n is based on $\mathbb{P}_n$, it is sensitive to outliers in $\mathbb{X}_n$, which, in turn affect the persistence diagrams (as illustrated in Figure 1). To this end, in this paper, we propose *robust persistence diagrams* constructed using superlevel filtrations of a robust density estimator of f, i.e., the filter function, ϕ_n

is chosen to be a robust density estimator of f. Specifically, we use the robust KDE, $f^n_{\rho,\sigma}$, introduced by Kim and Scott (2012) as the filter function, which is defined as a solution to the following M-estimation problem:

$$f^n_{\rho,\sigma} \doteq \arg\inf_{g\in\mathcal{G}} \int_{\mathcal{X}} \rho\left(\|\Phi_\sigma(\boldsymbol{y}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}_n(\boldsymbol{y}), \tag{1}$$

where $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a robust loss function, $\Phi_\sigma(\boldsymbol{x}) = K_\sigma(\cdot, \boldsymbol{x}) \in \mathcal{H}_\sigma$ is the feature map associated with K_σ for a fixed $\sigma > 0$, and $\mathcal{G} = \mathcal{H}_\sigma \cap \mathcal{D}_\sigma = \mathcal{D}_\sigma$ is the hypothesis class. Observe that when $\rho(z) = \frac{1}{2}z^2$, the unique solution to Eq. (1) is given by the KDE, f^n_σ. Therefore, a robust KDE is obtained by replacing the square loss with a *robust loss*, which satisfies the following assumptions. These assumptions, which are similar to those of Kim and Scott (2012); Vandermeulen and Scott (2013) guarantee the existence and uniqueness of $f^n_{\rho,\sigma}$ (if ρ is convex; Kim and Scott 2012), and are satisfied by most robust loss functions, including the Huber loss, $\rho(z) = \frac{1}{2}z^2\mathbb{1}\{z \leq 1\} + \left(z - \frac{1}{2}\right)\mathbb{1}\{z > 1\}$ and the Charbonnier loss, $\rho(z) = \sqrt{1+z^2} - 1$.

($\mathcal{A}1$) ρ is strictly-increasing and M-Lipschitz, with $\rho(0) = 0$.

($\mathcal{A}2$) $\rho'(x)$ is continuous and bounded with $\rho'(0) = 0$.

($\mathcal{A}3$) $\varphi(x) = \rho'(x)/x$ is bounded, L-Lipschitz and continuous, with $\varphi(0) < \infty$.

($\mathcal{A}4$) ρ'' exists, with ρ'' and φ nonincreasing.

Unlike for squared loss, the solution $f^n_{\rho,\sigma}$ cannot be obtained in a closed form. However, it can be shown to be the fixed point of an iterative procedure, referred to as KIRWLS algorithm (Kim and Scott, 2012). The KIRWLS algorithm starts with initial weights $\{w_i^{(0)}\}_{i=1}^n$ such that $\sum_{i=1}^n w_i^{(0)} = 1$, and generates the iterative sequence of estimators $\{f^{(k)}_{\rho,\sigma}\}_{k\in\mathbb{N}}$ as

$$f^{(k)}_{\rho,\sigma} = \sum_{i=1}^n w_i^{(k-1)} K_\sigma(\cdot, \boldsymbol{X}_i) \;\; ; \qquad\qquad w_i^{(k)} = \frac{\varphi(\|\Phi_\sigma(\boldsymbol{X}_i) - f^{(k)}_{\rho,\sigma}\|_{\mathcal{H}_\sigma})}{\sum_{j=1}^n \varphi(\|\Phi_\sigma(\boldsymbol{X}_j) - f^{(k)}_{\rho,\sigma}\|_{\mathcal{H}_\sigma})}.$$

Intuitively, note that if $\boldsymbol{X}_i$ is an outlier, then the corresponding weight w_i is small (since φ is nonincreasing) and therefore less weightage is given to the contribution of $\boldsymbol{X}_i$ in the density estimator. Hence, the weights serve as a measure of *inlyingness*—smaller (*resp.* larger) the weights, lesser (*resp.* more) inlying are the points. When $\mathbb{P}_n$ is replaced by $\mathbb{P}$, the solution of Eq. (1) is its population analogue, $f_{\rho,\sigma}$. Although $f_{\rho,\sigma}$ does not admit a closed form solution, it can be shown (Kim and Scott, 2012) that there exists a non-negative real-valued function w_σ satisfying $\int_{\mathbb{R}^d} w_\sigma(\boldsymbol{x})\, d\mathbb{P}(\boldsymbol{x}) = 1$ such that

$$f_{\rho,\sigma} = \int_{\mathbb{R}^d} K_\sigma(\cdot, \boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) = \int_{\mathbb{R}^d} \frac{\varphi(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma})}{\int_{\mathbb{R}^d} \varphi(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}) d\mathbb{P}(\boldsymbol{y})} K_\sigma(\cdot, \boldsymbol{x})\; d\mathbb{P}(\boldsymbol{x}), \tag{2}$$

where w_σ acts as a population analogue of the weights in KIRWLS algorithm.

To summarize our proposal, the fixed point of the KIRWLS algorithm, which yields the robust density estimator $f^n_{\rho,\sigma}$, is used as the filter function to obtain a robust persistence diagram of $\mathbb{X}_n$. On the computational front, note that $f^n_{\rho,\sigma}$ is computationally more complex than the KDE, f^n_σ, requiring $O(n\ell)$ computations compared to $O(n)$ of the latter, with ℓ being the number of iterations required to reach the fixed point of KIRWLS. However, once these filter functions are computed, the corresponding persistence diagrams have similar computational complexity as both require computing superlevel sets, which, in turn, require function evaluations that scale as $O(n)$ for both $f^n_{\rho,\sigma}$ and f^n_σ.

4 Theoretical Analysis of Robust Persistence Diagrams

In this section, we investigate the theoretical properties of the proposed robust persistence diagrams. First, in Section 4.1, we examine the sensitivity of persistence diagrams to outlying perturbations through the notion of *metric derivative* and compare the effect of different filter functions. Next, in Section 4.2, we establish consistency and convergence rates for the robust persistence diagram to its population analogue. These results allow to construct uniform confidence bands for the robust persistence diagram. The proofs of the results are provided in Section 6.

4.1 A measure of sensitivity of persistence diagrams to outliers

The influence function and gross error sensitivity are arguably the most popular tools in robust statistics for diagnosing the sensitivity of an estimator to a single adversarial contamination Hampel et al. (2011); Huber (2004). Given a statistical functional $T : \mathcal{P}(\mathcal{X}) \to (V, \|\cdot\|_V)$, which takes an input probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ on the input space $\mathcal{X}$ and produces a statistic $\mathbb{P} \mapsto T(\mathbb{P})$ in some normed space $(V, \|\cdot\|_V)$, the *influence function* of $\boldsymbol{x} \in \mathcal{X}$ at $\mathbb{P}$ is given by the Gâteaux derivative of T at $\mathbb{P}$ restricted to the space of signed Borel measures with zero expectation:

$$IF(T; \mathbb{P}, \boldsymbol{x}) \doteq \frac{\partial}{\partial \epsilon} T\Big((1-\epsilon)\mathbb{P} + \epsilon\delta_{\boldsymbol{x}}\Big)\Big|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{T\left((1-\epsilon)\mathbb{P} + \epsilon\delta_{\boldsymbol{x}}\right) - T(\mathbb{P})}{\epsilon},$$

and the *gross error sensitivity* at $\mathbb{P}$ is given by $\Gamma(T; \mathbb{P}) \doteq \sup_{\boldsymbol{x} \in \mathcal{X}} \|IF(T; \mathbb{P}, \boldsymbol{x})\|_V$. However, a persistence diagram (which is a statistical functional) does not take values in a normed space, and, therefore, the notion of influence function has to be generalized to metric spaces through the concept of a metric derivative: Given a complete metric space (X, d_X) and a curve $s : [0,1] \to X$, the *metric derivative* at $\epsilon = 0$ is given by

$$\left|s'\right|(0) \doteq \lim_{\epsilon \to 0} \frac{1}{\epsilon} d_X(s(0), s(\epsilon)).$$

Using this generalization, we have the following definition, which allows to examine the influence an outlier has on the persistence diagram obtained from a filtration.

Definition 4.1. *Given a probability measure $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and a filter function $\phi_{\mathbb{P}}$ depending on $\mathbb{P}$, the* persistence influence *of a perturbation $\boldsymbol{x} \in \mathbb{R}^d$ on $\mathsf{Dgm}\left(\phi_{\mathbb{P}}\right)$ is defined as*

$$\Psi\left(\phi_{\mathbb{P}}; \boldsymbol{x}\right) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} W_\infty\left(\mathsf{Dgm}\left(\phi_{\mathbb{P}^\epsilon_{\boldsymbol{x}}}\right), \mathsf{Dgm}\left(\phi_{\mathbb{P}}\right)\right),$$

where $\mathbb{P}^\epsilon_{\boldsymbol{x}} \doteq (1-\epsilon)\mathbb{P} + \epsilon\delta_{\boldsymbol{x}}$, and the gross-influence *is defined as $\Gamma(\phi_{\mathbb{P}}) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} \Psi\left(\phi_{\mathbb{P}}; \boldsymbol{x}\right)$.*

The following result (proved in Section 6.1) bounds the persistence influence for the persistence diagram induced by the filter function $f_{\rho,\sigma}$, which is the population analogue of robust KDE.

Theorem 4.1. *For a loss ρ satisfying $(\mathcal{A}1)$–$(\mathcal{A}3)$, and $\sigma > 0$, the persistence influence of $\boldsymbol{x} \in \mathbb{R}^d$ on $\mathsf{Dgm}\left(f_{\rho,\sigma}\right)$ satisfies*

$$\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right) \left(\int_{\mathbb{R}^d} \zeta\Big(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\Big) d\mathbb{P}(\boldsymbol{y})\right)^{-1}, \tag{3}$$

where $\zeta(z) = \varphi(z) - z\varphi'(z)$.

Remark 4.1. *We make the following observations from Theorem 4.1.*

(i) Choosing $\rho(z) = \frac{1}{2}z^2$ *and noting that* $\varphi(z) = \rho''(z) = 1$*, Eq. (3) yields a bound for the persistence influence of the KDE as*

$$\Psi\left(f_\sigma; \boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \left\|\Phi_\sigma(\boldsymbol{x}) - f_\sigma\right\|_{\mathcal{H}_\sigma}.$$

On the other hand, for robust loss functions, the term involving ρ' *is bounded because of* $(\mathcal{A}2)$*, making them less sensitive to very large perturbations. In fact, for nonincreasing* φ*, it can be shown (see Section 6.2) that*

$$\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} w_\sigma(\boldsymbol{x}) \left\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma},$$

where, in contrast to KDE, the measure of inlyingness, w_σ*, weighs down extreme outliers.*

(ii) For the generalized Charbonnier loss (a robust loss function), given by $\rho(z) = \left(1 + z^2\right)^{\alpha/2} - 1$ *for* $1 \leq \alpha \leq 2$*, the persistence influence satisfies*

$$\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \left(1 + \left\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}^2\right)^{\frac{\alpha-1}{2}} \left(1 + \int_{\mathbb{R}^d} \left\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}^2 d\mathbb{P}(\boldsymbol{y})\right)^{\frac{1-\alpha}{2}}.$$

Note that for $\alpha = 1$*, the bound on the persistence influence* $\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right)$ *does not depend on how extreme the outlier* $\boldsymbol{x}$ *is. Similarly, for the Cauchy loss, given by* $\rho(z) = \log(1 + z^2)$*, we have*

$$\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \left(1 + \int_{\mathbb{R}^d} \left\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}^2 d\mathbb{P}(\boldsymbol{y})\right).$$

This shows that for large perturbations, the gross error sensitivity for the Cauchy and Charbonnier losses are far more stable than that of KDE. This behavior is also empirically illustrated in Figure 3. The experiment is detailed in Section 5.2.

(iii) For the DTM function, it can be shown that

$$\Psi\left(d_{\mathbb{P},m}; \boldsymbol{x}\right) \leq \frac{2}{\sqrt{m}} \sup\left\{\left|f(\boldsymbol{x}) - \int_{\mathbb{R}^d} f(\boldsymbol{y}) d\mathbb{P}(\boldsymbol{y})\right| : \|\nabla f\|_{L_2(\mathbb{P})} \leq 1\right\}. \tag{4}$$

While $d_{\mathbb{P},m}$ *cannot be compared to both* f_σ *and* $f_{\rho,\sigma}$*, as it captures topological information at a different scale, determined by* m*, we point out that when* $\operatorname{supp}(\mathbb{P})$ *is compact,* $\Psi\left(d_{\mathbb{P},m}; \boldsymbol{x}\right)$ *is not guaranteed to be bounded, unlike in* $\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right)$*. We refer the reader to Section 6.2 for more details.*

It follows from Remark 4.1 that as $\sigma \to 0$, the persistence influence of both the KDE and robust KDE behave as $O(\sigma^{-d})$, showing that the robustness of robust persistence diagrams manifests only in cases where $\sigma > 0$. However, robustness alone has no bearing if the robust persistence diagram and the persistence diagram from the KDE are fundamentally different, i.e., they estimate different quantities as $\sigma \to 0$. The following result (proved in Section 6.3) shows that as $\sigma \to 0$, $\mathsf{Dgm}\left(f_{\rho,\sigma}\right)$ recovers the same information as that in $\mathsf{Dgm}\left(f_\sigma\right)$, which is same as $\mathsf{Dgm}\left(f\right)$, where f is the density of $\mathbb{P}$.

Theorem 4.2. *For a strictly-convex loss* ρ *satisfying* $(\mathcal{A}1)$*–*$(\mathcal{A}4)$*, and* $\sigma > 0$*, suppose* $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$ *with density* f*, and* $f_{\rho,\sigma}$ *is the robust KDE. Then*

$$W_\infty\left(\mathsf{Dgm}\left(f_{\rho,\sigma}\right), \mathsf{Dgm}\left(f\right)\right) \to 0 \qquad \text{as} \quad \sigma \to 0.$$

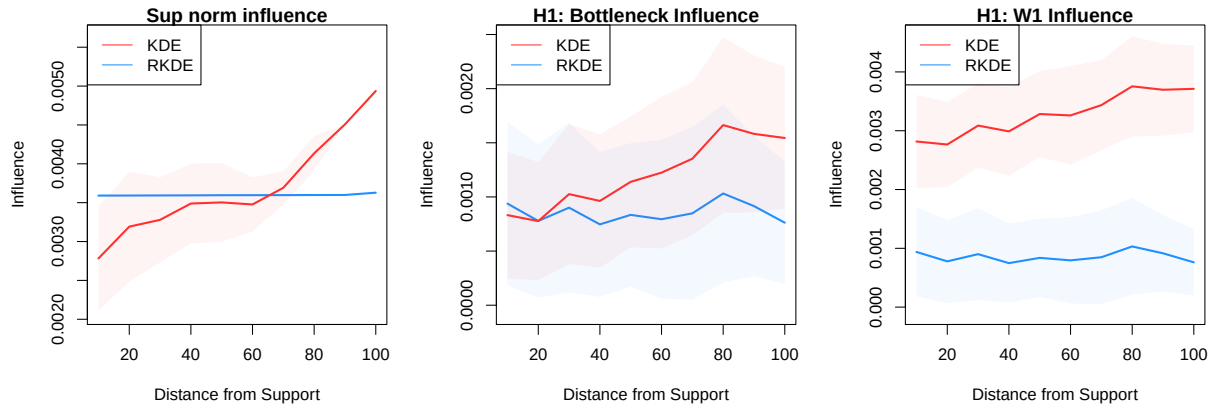


Figure 3: Points $\mathbb{X}_n$ are sampled from $\mathbb{P}$ with nontrivial 1^{st}-order homological features and outliers $\mathbb{Y}_m$ are added at a distance r from the support of $\mathbb{P}$. (Left) The average L_∞ distance between the density estimators computed using $\mathbb{X}_n$ and $\mathbb{X}_n \cup \mathbb{Y}_m$ as r increases. (Center) The average W_∞ distance between the corresponding persistence diagrams for the 1^{st}-order homological features. (Right) The W_1 distance (defined in Eq. B.1 in Appendix B) between the same persistence diagrams. The results show that the outliers $\mathbb{Y}_m$ have little influence on the persistence diagrams from the robust KDEs. In contrast, as the outliers become more extreme (i.e., r increases) their influence on the persistence diagrams from the KDE becomes more prominent.

Suppose $\mathbb{P} = (1-\pi)\mathbb{P}_0 + \pi\mathbb{Q}$, where $\mathbb{P}_0$ corresponds to the true signal which we are interested in studying, and $\mathbb{Q}$ manifests as some ambient noise with $0 < \pi < \frac{1}{2}$. In light of Theorem 4.2, by letting $\sigma \to 0$, along with the topological features of $\mathbb{P}_0$, we are also capturing the topological features of $\mathbb{Q}$, which may obfuscate any statistical inference made using the persistence diagrams. In a manner, choosing $\sigma > 0$ suppresses the noise in the resulting persistence diagrams, thereby making them more stable. On a similar note, the authors in Fasy et al. (2014) state that for a suitable bandwidth $\sigma > 0$, the level sets of f_σ carry the same topological information as $\text{supp}(\mathbb{P})$, despite the fact that some subtle details in f may be omitted. In what follows, we consider the setting where robust persistence diagrams are constructed for a fixed $\sigma > 0$.

4.2 Statistical properties of robust persistence diagrams from samples

Suppose $\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right)$ is the robust persistence diagram obtained from the robust KDE on a sample $\mathbb{X}_n$ and $\mathsf{Dgm}\left(f_{\rho,\sigma}\right)$ is its population analogue obtained from $f_{\rho,\sigma}$. The following result (proved in Section 6.4) establishes the consistency of $\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right)$ in the W_∞ metric.

Theorem 4.3. *For convex loss ρ satisfying $(\mathcal{A}1)$–$(\mathcal{A}4)$, and fixed $\sigma > 0$, suppose $\mathbb{X}_n$ is observed iid from a distribution $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$ with density f. Then*

$$W_\infty\left(\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right), \mathsf{Dgm}\left(f_{\rho,\sigma}\right)\right) \xrightarrow{p} 0 \quad \text{as } n \to \infty.$$

We present the convergence rate of the above convergence in Theorem 4.4, which depends on the smoothness of $\mathcal{H}_\sigma$. In a similar spirit to Fasy et al. (2014), this result paves the way for constructing uniform confidence bands. Before we present the result, we first introduce the notion of *entropy numbers* associated with an RKHS.

Definition 4.2 (Entropy Number). *Given a metric space (T, d) the n^{th} entropy number is defined as*

$$e_n(T,d) \doteq \inf\left\{\epsilon > 0 : \exists\ \{t_1, t_2, \ldots, t_{2^{n-1}}\} \subset T \ \text{ such that } T \subset \bigcup_{i=1}^{2^{n-1}} B_d(t_i, \epsilon)\right\}.$$

Further, if $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ are two normed spaces and $L : V \to W$ is a bounded, linear operator, then $e_n(L) = e_n(L : V \to W) \doteq e_n\left(L(B_V), \|\cdot\|_W\right)$, where B_V is a unit ball in V.

Loosely speaking, entropy numbers are related to the eigenvalues of the integral operator associated with the kernel K_σ, and measure the capacity of the RKHS in approximating functions in $L_2(\mathbb{R}^d)$. In our context, the entropy numbers will provide useful bounds on the covering numbers of sets in the hypothesis class $\mathcal{G}$. We refer the reader to Steinwart and Christmann (2008) for more details. With this background, the following theorem (proved in Section 6.5) provides a method for constructing uniform confidence bands for the persistence diagram constructed using the robust KDE on $\mathbb{X}_n$.

Theorem 4.4. *For convex loss ρ satisfying $(\mathcal{A}1)$–$(\mathcal{A}4)$, and fixed $\sigma > 0$, suppose the kernel K_σ satisfies $e_n\left(\mathrm{id} : \mathcal{H}_\sigma \to L_\infty(\mathcal{X})\right) \le a_\sigma n^{-\frac{1}{2p}}$, where $a_\sigma > 1$, $0 < p < 1$ and $\mathcal{X} \subset \mathbb{R}^d$. Then, for a fixed confidence level $0 < \alpha < 1$,*

$$\sup_{\mathbb{P}\in\mathcal{M}(\mathcal{X})} \mathbb{P}^{\otimes n}\left\{W_\infty\Big(\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right), \mathsf{Dgm}\left(f_{\rho,\sigma}\right)\Big) > \frac{2M\left\|K_\sigma\right\|_\infty^{\frac{1}{2}}}{\mu}\left(\xi(n,p) + \delta\sqrt{\frac{2\log\left(1/\alpha\right)}{n}}\right)\right\} \le \alpha,$$

where $\xi(n,p)$ is given by

$$\xi(n,p) = \begin{cases} \gamma\frac{a_\sigma^p}{(1-2p)} \cdot \frac{1}{\sqrt{n}} & \text{if } 0 < p < \frac{1}{2}, \\ \gamma C\sqrt{a_\sigma} \cdot \frac{\log(n)}{\sqrt{n}} & \text{if } p = \frac{1}{2}, \\ \gamma\frac{p\sqrt{a_\sigma}}{2p-1} \cdot \frac{1}{n^{1/4p}} & \text{if } \frac{1}{2} < p < 1, \end{cases}$$

for fixed constants $\gamma > \frac{12}{\sqrt{\log 2}}$, $C > 3 - \log(9a_\sigma)$ and $\mu = 2\min\left\{\varphi(2\left\|K_\sigma\right\|_\infty^{\frac{1}{2}}), \rho''(2\left\|K_\sigma\right\|_\infty^{\frac{1}{2}})\right\}$.

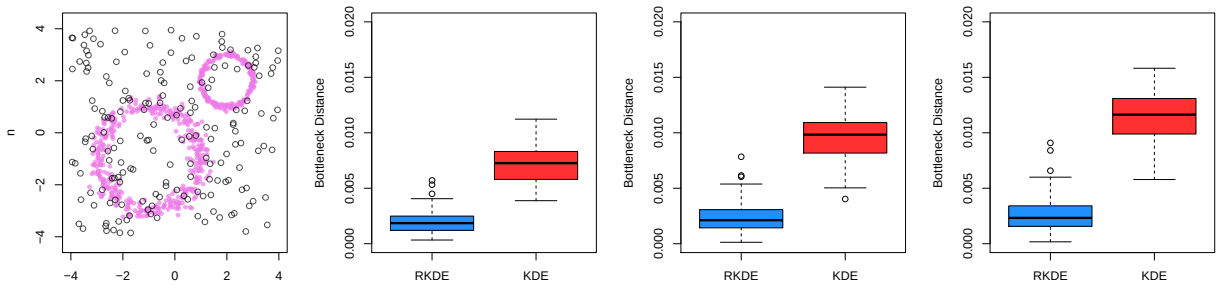
Remark 4.2. *We highlight some salient observations from Theorem 4.4.*

(i) If $diam(\mathcal{X}) = r$, and the kernel K_σ is m-times differentiable, then from (Steinwart and Christmann, 2008, Theorem 6.26), the entropy numbers associated with K_σ satisfy

$$e_n\left(\mathrm{id} : \mathcal{H}_\sigma \to L_\infty(\mathcal{X})\right) \le cr^m n^{-\frac{m}{d}}.$$

In light of Theorem 4.4, for $p = \frac{d}{2m}$, we can make two important observations. First, as the dimension of the input space $\mathcal{X}$ increases, we have that the rate of convergence decreases; which is a direct consequence from the curse of dimensionality. Second, for a fixed dimension of the input space, the parameter p in Theorem 4.4 can be understood to be inversely proportional to the smoothness of the kernel. Specifically, as the smoothness of the kernel increases, the rate of convergence is faster, and we obtain sharper confidence bands. This makes a case for employing smoother kernels.

(ii) A similar result is obtained in (Fasy et al., 2014, Lemma 8) for persistence diagrams from the KDE, with a convergence rate $O_p(n^{-1/2})$, where the proof relies on a simple application of Hoeffding's inequality, unlike the sophisticated tools the proof of Theorem 4.4 warrants for the robust KDE.



(a) $\mathbb{X}_n$ (in •) and $\mathbb{Y}_m$ (in ∘) (b) $\pi = 20\%$, $p = 4 \times 10^{-60}$ (c) $\pi = 30\%$, $p = 2 \times 10^{-72}$ (d) $\pi = 40\%$, $p = 2.5 \times 10^{-75}$

Figure 4: (a) A realization of $\mathbb{X}_n \cup \mathbb{Y}_m$. (b, c, d) As the noise level π increases, boxplots for $W_\infty\left(\mathbf{D}_{\rho,\sigma}, \mathcal{D}_\sigma^{\#}\right)$ in blue and $W_\infty\left(\mathbf{D}_\sigma, \mathcal{D}_\sigma^{\#}\right)$ in red show that the robust persistence diagram recovers the underlying signal better.

5 Experiments

We illustrate the performance of robust persistence diagrams in machine learning applications through synthetic and real-world experiments. In all the experiments, the kernel bandwidth σ is chosen as the median distance of each $\boldsymbol{x}_i \in \mathbb{X}_n$ to its k^{th}–nearest neighbour using the Gaussian kernel with the Hampel loss (similar setting as in Kim and Scott, 2012)—we denote this bandwidth as $\sigma(k)$. Since DTM is closely related to the k-NN density estimator (Biau et al., 2011), we choose the DTM smoothing parameter as $m(k) = k/n$.

5.1 Bottleneck Simulation

The objective of this experiment is to assess how the robust KDE persistence diagram compares to the KDE persistence diagram in recovering the topological features of the underlying signal. $\mathbb{X}_n$ is observed uniformly from two circles and $\mathbb{Y}_m$ is sampled uniformly from the enclosing square such that $m = 200$ and $m/n = \pi \in \{20\%, 30\%, 40\%\}$—shown in Figure 4 (a). For each noise level π, and for each of $N = 100$ realizations of $\mathbb{X}_n$ and $\mathbb{Y}_m$, the robust persistence diagram $\mathbf{D}_{\rho,\sigma}$ and the KDE persistence diagram $\mathbf{D}_\sigma$ are constructed from the noisy samples $\mathbb{X}_n \cup \mathbb{Y}_m$. In addition, we compute the KDE persistence diagram $\mathcal{D}_\sigma^{\#}$ on $\mathbb{X}_n$ alone as a proxy for the target persistence diagram one would obtain in the absence of any contamination. The bandwidth $\sigma(k) > 0$ is chosen for $k = 5$. For each realization i, bottleneck distances $U_i = W_\infty\left(\mathbf{D}_{\rho,\sigma}, \mathcal{D}_\sigma^{\#}\right)$ and $V_i = W_\infty\left(\mathbf{D}_\sigma, \mathcal{D}_\sigma^{\#}\right)$ are computed for 1^{st}-order homological features. The boxplots and p-values for the one-sided hypothesis test $H_0 : U - V = 0$ vs. $H_1 : U - V < 0$ are reported in Figures 4 (b, c, d). The results demonstrate that the robust persistence diagram is noticeably better in recovering the true homological features, and in fact demonstrates superior performance when the noise levels are higher.

5.2 Persistence-Influence Experiment

Points $\mathbb{X}_n$ are sampled from an annular region inside $[-5, 5]^2$ along with some uniform noise in the ambient space, corresponding to the black points in Figure 5 (a). $\mathbb{X}_n$ has interesting 1^{st}-order homological features. We compute the robust KDE $f_{\rho,\sigma}^n$ and the KDE f_σ^n on the points $\mathbb{X}_n$ along with the corresponding persistence diagrams $\mathsf{Dgm}\left(f_{\rho,\sigma}^n\right)$ and $\mathsf{Dgm}\left(f_\sigma^n\right)$. Outliers $\mathbb{Y}_m$ are added to the original points at a distance r from the origin, the number of points roughly equal to r. Figure 5 (a) depicts these outliers in orange when $r = 20$.

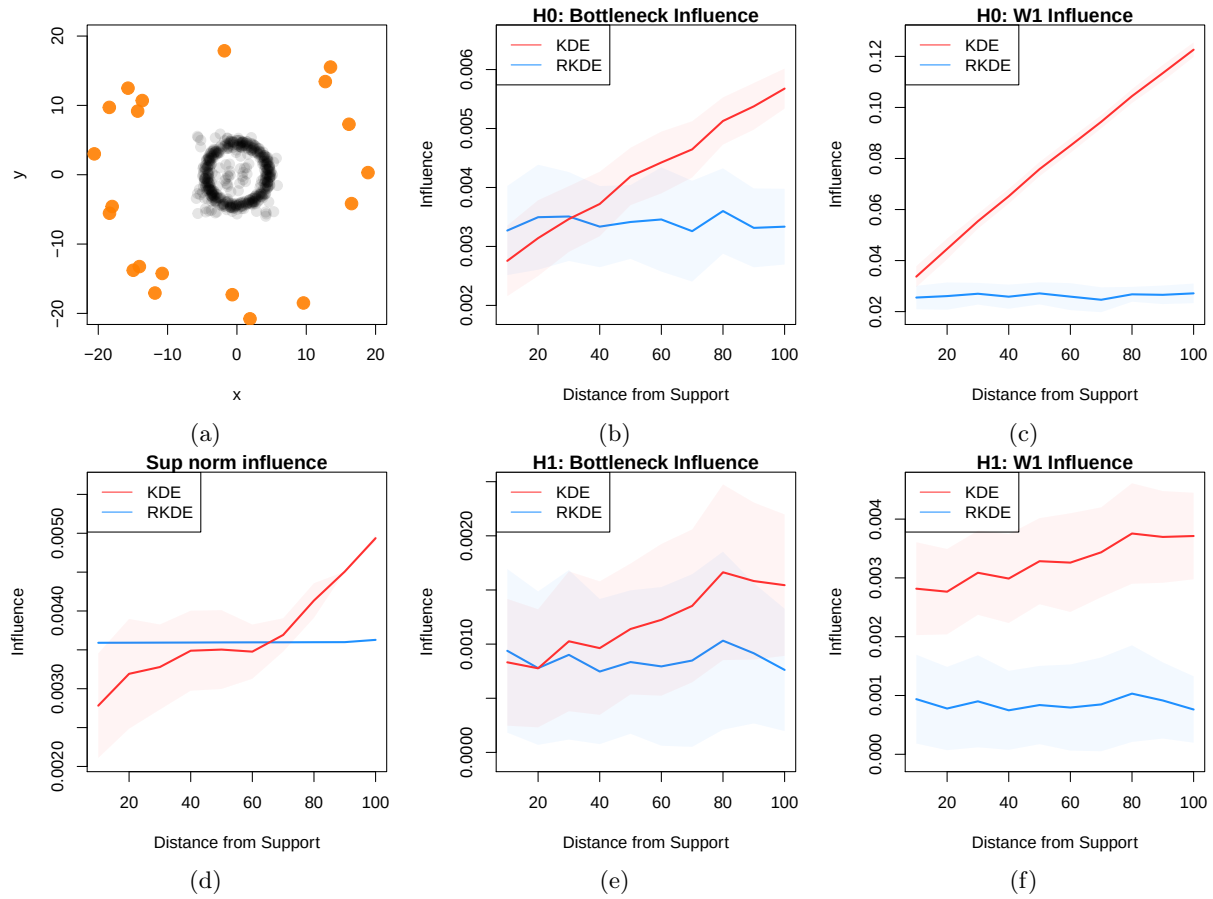


Figure 5: (a) An example of $\mathbb{X}_n$ in blue and the contamination $\mathbb{Y}_m$ when $r = 10$. (d) The L_∞ influence of $\mathbb{Y}_m$ on the KDE and robust KDE. (b, e) The bottleneck influence of $\mathbb{Y}_m$. (c, f) The 1-Wasserstein influence of $\mathbb{Y}_m$ as the distance r increases.

The robust KDE $f^{n+m}_{\rho,\sigma}$ and f^{n+m}_{σ} are now computed on the composite sample $\mathbb{X}_n \cup \mathbb{Y}_m$ along with the persistence diagrams $\mathsf{Dgm}\left(f^{n+m}_{\rho,\sigma}\right)$ and $\mathsf{Dgm}\left(f^{n+m}_{\sigma}\right)$. The bandwidth $\sigma(k)$ is chosen for $k = 5$. We then compute the L_∞ influence of $\mathbb{Y}_m$ i.e., $\|f^{n+m} - f^n\|_\infty$ as shown in Figure 5 (d). Additionally for each of the 0^{th}-order and 1^{st}-order persistence diagrams, we compute the persistence influence of $\mathbb{Y}_m$, i.e., $W_\infty\left(\mathsf{Dgm}\left(f^{n+m}\right), \mathsf{Dgm}\left(f^n\right)\right)$ as shown in Figures 5 (b, e), and the 1-Wasserstein influence, i.e., $W_1\left(\mathsf{Dgm}\left(f^{n+m}\right), \mathsf{Dgm}\left(f^n\right)\right)$ as shown in Figures 5 (c, f). We refer the reader to Eq. (B.1) in Appendix B for the definition of W_1 metric.

For each value of r, we generate 100 such samples and report the average in Figure 5. Figure 5 (d, e, f) correspond to the experiment illustrated in Figure 3. The results indicate that the robust persistence diagrams, $\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right)$, are relatively unperturbed when the outliers are added. It exhibits stability even as r become very large. The KDE persistence diagrams, $\mathsf{Dgm}\left(f^n_{\sigma}\right)$, on the other hand, are unstable as the outlying noise becomes more extreme.

As discussed in the Remark 4.1 (iii), the persistence influence for DTM has a much weaker bound as the outliers become more extreme, and in general is not guaranteed to be bounded. In Figure 6 we illustrate the results from the same experiment when the persistence diagrams from DTM is contrasted with the persistence diagrams from the KDE. This analysis is for the same data as that used in Figure 3. We remark that even though DTM is highly sensitive to extreme outliers, DTM

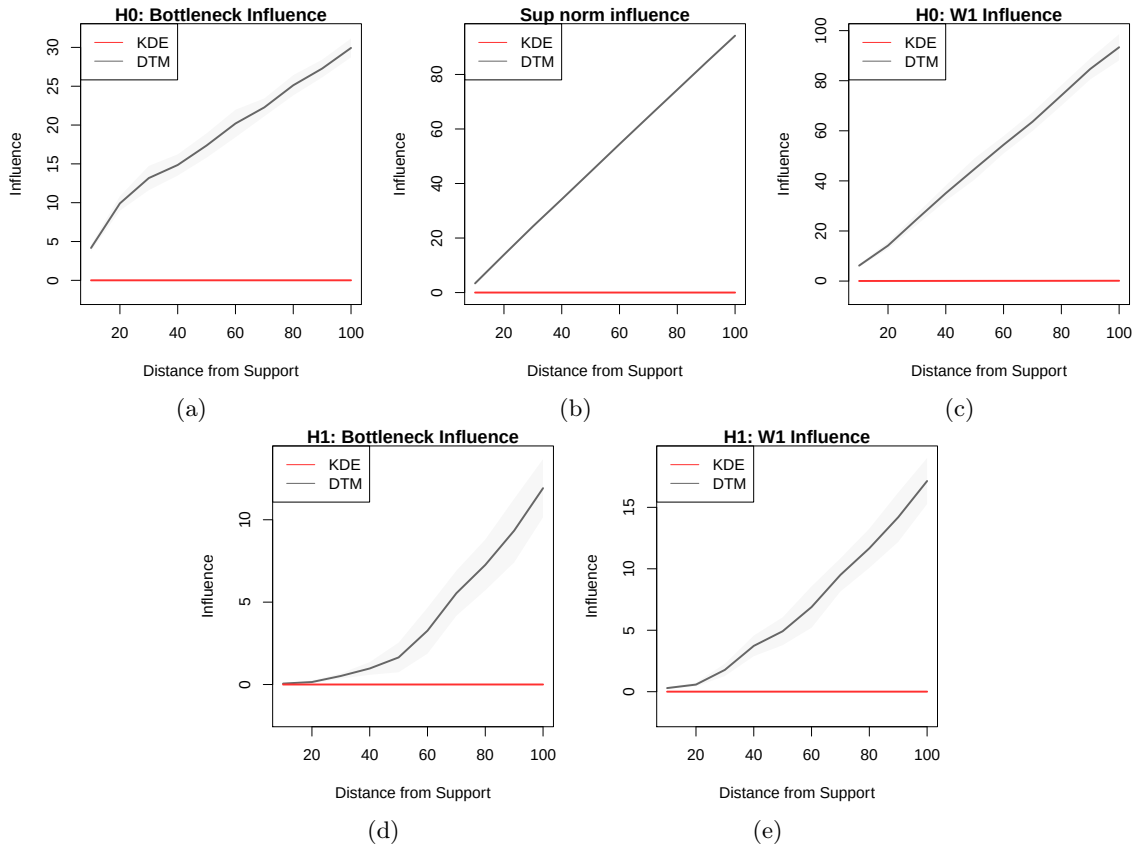
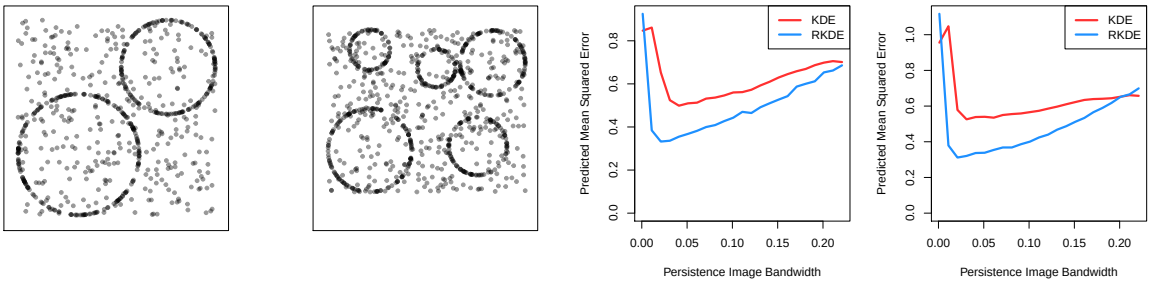


Figure 6: For the same data in Figure 5, (a, d) depicts the bottleneck influence for the DTM in contrast to the KDE – the red line is the same as the one from Figure 5 (b, e). Similarly, in (c, e) we see the W_1 persistence influence of $\mathbb{Y}_m$ for the DTM in contrast to the KDE. (b) shows the L_∞ influence of $\mathbb{Y}_m$ on the DTM. The robust KDE lines are omitted from all plots as it appears to almost merge with the KDE at this scale.

based filtrations have other remarkable properties, as described in Chazal et al. (2017). They are very useful for analyzing persistent homology when one has access to just a single collection of points $\mathbb{X}_n$. For DTM the smoothing parameter is chosen as $m(k) = k/n$ with $k = 5$.

5.3 Random Circles

The objective of this simulation is to evaluate the performance of persistence diagrams in a supervised learning task. We select circles $\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_{\mathbf{N}}$ randomly in $\mathbb{R}^2$ with centers inside $[0, 2]^2$, with the number of such circles, $\mathbf{N}$ uniformly sampled from $\{1, 2, \ldots, 5\}$. Conditional on $\mathbf{N} = N$, $\mathbb{X}_n$ is sampled uniformly from $\mathbb{S}_1, \ldots, \mathbb{S}_N$ with 50% noise in the enclosing square. Two such point clouds are shown in Figure 7 (a, b). Persistence diagrams $\mathsf{Dgm}\left(f^n_\sigma\right)$ and $\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right)$ are constructed for bandwidth $\sigma(k)$ selected from $k = 5, 7$, and vectorized in the form of persistence images $\mathsf{Img}\left(f^n_\sigma, h\right)$, and $\mathsf{Img}\left(f^n_{\rho,\sigma}, h\right)$ for varying bandwidths h (Adams et al., 2017). With $\mathbf{N}$ as the response and the persistence images as the input, results from a support vector regression, averaged over 50 random splits, is shown in Figure 7 (c, d). For a fixed h the robust persistence diagram seems to always contain more predictive information, as observed in the envelope it forms in Figure 7 (c, d).

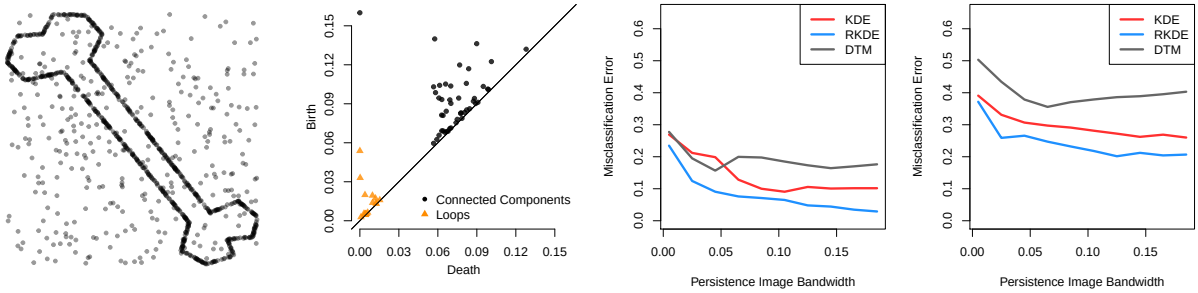


(a) $\mathbb{X}_n$ when $\mathbf{N} = 2$ (b) $\mathbb{X}_n$ when $\mathbf{N} = 5$ (c) $k = 5$ (d) $k = 7$

Figure 7: (a, b) A realization $\mathbb{X}_n$ when $\mathbf{N} = 2$ and $\mathbf{N} = 5$. (c, d) The predicted mean-squared error vs. the persistence image bandwidth for persistence diagrams in support vector regression.

5.4 MPEG7

In this experiment, we examine the performance of persistence diagrams in a classification task using the `MPEG7` dataset Latecki et al. (2000). For simplicity, we only consider five classes: *beetle*, *bone*, *spring*, *deer* and *horse*. We first extract the boundary of the images using a Laplace convolution, and sample $\mathbb{X}_n$ uniformly from the boundary of each image, adding uniform noise ($\pi = 15\%$) in the enclosing region. Persistence diagrams $\mathsf{Dgm}\left(f_\sigma^n\right)$ and $\mathsf{Dgm}\left(f_{\rho,\sigma}^n\right)$ from the KDE and robust KDE are constructed. In addition, owing to their ability to capture nuanced multi-scale features, we also construct $\mathsf{Dgm}\left(d_{n,m}\right)$ from the DTM filtration. The smoothing parameters $\sigma(k)$ and $m(k)$ are chosen as earlier for $k = 5$. The persistence diagrams are normalized to have a max persistence $\max\{|d - b| = 1 : (b, d) \in \mathsf{Dgm}(\phi)\}$, and then vectorized as persistence images for various bandwidths h. A linear SVM classifier is then trained on the resulting persistence images. In the first experiment we only consider the first three classes, and in the second experiment we consider all five classes. The results for the classification error, shown in Figure 8, demonstrate the advantage of the proposed method.



(a) (b) (c) (d)

Figure 8: (a) $\mathbb{X}_n$ is sampled from the image boundary of a *bone*, and uniform noise is added. (b) The resulting persistence diagram from the robust KDE. The persistence diagram picks up the 1^{st}–order features near the joints of the cartoon bone. The misclassification error for the KDE, robust KDE and DTM as the persistence image bandwidth increases, (c) for the three-class classification and, (d) for the five-class classification.

6 Proofs

In what follows, for a fixed loss ρ, we will use the notation $\ell_g(\cdot) = \ell(\cdot, g) = \rho\left(\|\Phi(\cdot) - g\|_{\mathcal{H}_\sigma}\right)$ in order to emphasize the dependency of the loss on the choice of $g \in \mathcal{G}$. Borrowing some notation from empirical process theory, we define the empirical risk-functional in Eq. (1) as

$$\mathcal{J}_n(g) \doteq \mathbb{P}_n \ell_g = \sum_{i=1}^{n} \rho\left(\|\Phi_\sigma(\boldsymbol{X}_i) - g\|_{\mathcal{H}_\sigma}\right),$$

and, similarly, the population risk functional $\mathcal{J}(g)$ is given by

$$\mathcal{J}(g) \doteq \mathbb{P}\ell_g = \int_{\mathbb{R}^d} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}).$$

6.1 Proof of Theorem 4.1

For $\epsilon > 0$, define the risk functional associated with $\mathbb{P}^\epsilon_{\boldsymbol{x}}$ to be

$$\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) = \mathbb{P}^\epsilon_{\boldsymbol{x}} \ell_g = (1-\epsilon)\mathcal{J}(g) + \epsilon\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right),$$

and let $f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} = \inf_{g\in\mathcal{G}} \mathcal{J}_{\epsilon,\boldsymbol{x}}(g)$ be its minimizer. From the stability result of Proposition 2.1 we have that

$$\Psi\left(f_{\rho,\sigma}; \boldsymbol{x}\right) = \lim_{\epsilon\to 0} \frac{1}{\epsilon} W_\infty\left(\mathsf{Dgm}\left(f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma}\right), \mathsf{Dgm}\left(f_{\rho,\sigma}\right)\right) \leq \lim_{\epsilon\to 0} \frac{1}{\epsilon} \left\|f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty.$$

Using Propositions A.2 and A.3, we know that the sequence $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ is equi-coercive, and $\mathcal{J}_{\epsilon,\boldsymbol{x}}$ Γ–converges to $\mathcal{J}$ as $\epsilon \to 0$. From the fundamental theorem of Γ–convergence (Braides, 2002) we have that $\left\|f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} \to 0$, and, consequently, $\left\|f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty \to 0$ as $\epsilon \to 0$. Thus,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon} \left\|f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty = \left\|\lim_{\epsilon\to 0} \frac{f^{\epsilon,\boldsymbol{x}}_{\rho,\sigma} - f_{\rho,\sigma}}{\epsilon}\right\|_\infty. \tag{5}$$

Let the limit in the right hand side of Eq. (5) be denoted by $\dot{f}_{\rho,\sigma}$. Although $\dot{f}_{\rho,\sigma}$ does not admit a closed-form solution, from (Kim and Scott, 2012, Theorem 8) we have that $\dot{f}_{\rho,\sigma}$ satisfies $V = a\dot{f}_{\rho,\sigma} + B$, where

$$\begin{aligned}
V &= \varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right) \cdot \left(\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\right), \\
a &= \int_{\mathbb{R}^d} \varphi\left(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{y}), \quad \text{and} \\
B &= \int_{\mathbb{R}^d} \left(\frac{\varphi'\left(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}} \left\langle \dot{f}_{\rho,\sigma}, \Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma} \right\rangle_{\mathcal{H}_\sigma} \cdot \left(\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\right) \right) d\mathbb{P}(\boldsymbol{y}).
\end{aligned}$$

For brevity, we adopt the notation $z(\boldsymbol{y}) = \|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}$ and $u(\cdot, \boldsymbol{y}) = \frac{\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}}{\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}} \in \mathcal{H}_\sigma$. Then note that $a \in \mathbb{R}$ and $B \in \mathcal{H}_\sigma$ are given by

$$\begin{aligned}
a &= \int_{\mathbb{R}^d} \varphi\left(z(\boldsymbol{y})\right) d\mathbb{P}(\boldsymbol{y}), \quad \text{and} \\
B &= \int_{\mathbb{R}^d} z(\boldsymbol{y}) \varphi'\left(z(\boldsymbol{y})\right) \left\langle \dot{f}_{\rho,\sigma}, u(\cdot, \boldsymbol{y}) \right\rangle_{\mathcal{H}_\sigma} u(\cdot, \boldsymbol{y}) \, d\mathbb{P}(\boldsymbol{y}).
\end{aligned}$$

Using the reverse triangle inequality we have

$$\|V\|_{\mathcal{H}_\sigma} \geq a \left\|\dot{f}_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} - \|B\|_{\mathcal{H}_\sigma}. \tag{6}$$

We now look for an upper bound on $\|B\|_{\mathcal{H}_\sigma}$. By noting that

$$\left\langle \dot{f}_{\rho,\sigma}, u(\cdot,\boldsymbol{x})\right\rangle_{\mathcal{H}_\sigma}\left\langle \dot{f}_{\rho,\sigma}, u(\cdot,\boldsymbol{y})\right\rangle_{\mathcal{H}_\sigma}\left\langle u(\cdot,\boldsymbol{x}), u(\cdot,\boldsymbol{y})\right\rangle_{\mathcal{H}_\sigma} \leq \left\|\dot{f}_{\rho,\sigma}\right\|^2_{\mathcal{H}_\sigma},$$

we have

$$\begin{aligned}\|B\|^2_{\mathcal{H}_\sigma} = \left\langle B, B\right\rangle_{\mathcal{H}_\sigma} &\leq \iint z(\boldsymbol{x})\varphi'(z(\boldsymbol{x}))z(\boldsymbol{y})\varphi'(z(\boldsymbol{y}))\left\|\dot{f}_{\rho,\sigma}\right\|^2_{\mathcal{H}_\sigma} d\mathbb{P}(\boldsymbol{x})\, d\mathbb{P}(\boldsymbol{y}) \\ &= \left\|\dot{f}_{\rho,\sigma}\right\|^2_{\mathcal{H}_\sigma}\left(\int_{\mathbb{R}^d} z(\boldsymbol{y})\varphi'(z(\boldsymbol{y}))\, d\mathbb{P}(\boldsymbol{y})\right)^2.\end{aligned}$$

Plugging this back into Eq. (6) we get

$$\begin{aligned}\|V\|_{\mathcal{H}_\sigma} &\geq \left\|\dot{f}_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}\int_{\mathbb{R}^d} \varphi(z(\boldsymbol{y})) - z(\boldsymbol{y})\varphi'(z(\boldsymbol{y}))\, d\mathbb{P}(\boldsymbol{y}) \\ &= \left\|\dot{f}_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}\int_{\mathbb{R}^d} \zeta(z(\boldsymbol{y}))\, d\mathbb{P}(\boldsymbol{y}),\end{aligned} \tag{7}$$

where $\zeta(z) = \varphi(z) - z\varphi'(z)$. Similarly, by using the definition of φ, it follows that

$$\|V\|_{\mathcal{H}_\sigma} = \varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)\cdot\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} = \rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right).$$

Combining this with Eq. (7) we get

$$\left\|\dot{f}_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} \leq \frac{\rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)}{\int_{\mathbb{R}^d}\zeta\left(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)d\mathbb{P}(\boldsymbol{y})}.$$

By noting that $\left\|\dot{f}_{\rho,\sigma}\right\|_\infty \leq \|K_\sigma\|_\infty^{\frac{1}{2}}\left\|\dot{f}_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}$ and $\Psi(f_{\rho,\sigma};\boldsymbol{x}) \leq \left\|\dot{f}_{\rho,\sigma}\right\|_\infty$, the result follows. ■

6.2 Supplementary Results for the Persistence Influence

In this section, we collect the proofs for the results on persistence influence established in Remark 4.1 from Section 4.1. The following result shows that when φ is nonincreasing, the persistence influence in Eq. (3) can be written in a more succinct form.

Proposition 6.1. *Under the conditions of Theorem 4.1, if φ is nonincreasing, then the persistence influence of $\boldsymbol{x} \in \mathbb{R}^d$ on $\mathsf{Dgm}\,(f_{\rho,\sigma})$ satisfies*

$$\Psi\left(f_{\rho,\sigma};\boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}}\, w_\sigma(\boldsymbol{x})\,\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma},$$

where w_σ is the measure of inlyingness from Eq. (2).

Proof. From Theorem 4.1 we have that the persistence influence satisfies

$$\Psi\left(f_{\rho,\sigma};\boldsymbol{x}\right) \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)\left(\int_{\mathbb{R}^d} \zeta\Big(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\Big) d\mathbb{P}(\boldsymbol{y})\right)^{-1}, \tag{8}$$

where $\zeta(z) = \varphi(z) - z\varphi'(z)$. When φ is nonincreasing, observe that $z\varphi'(z) \leq 0$ for all $0 \leq z < \infty$. Consequently, ζ can be bounded below by φ, and the r.h.s. in Eq. (8) can be bounded above by

$$\begin{aligned}
\Psi\left(f_{\rho,\sigma};\boldsymbol{x}\right) &\overset{(i)}{\leq} \|K_\sigma\|_\infty^{\frac{1}{2}} \frac{\rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)}{\int_{\mathbb{R}^d} \varphi\Big(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\Big) d\mathbb{P}(\boldsymbol{y})} \\
&\overset{(ii)}{=} \|K_\sigma\|_\infty^{\frac{1}{2}} \frac{\varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\right)}{\int_{\mathbb{R}^d} \varphi\Big(\|\Phi_\sigma(\boldsymbol{y}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma}\Big) d\mathbb{P}(\boldsymbol{y})} \|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} \\
&\overset{(iii)}{=} \|K_\sigma\|_\infty^{\frac{1}{2}} w_\sigma(\boldsymbol{x}) \|\Phi_\sigma(\boldsymbol{x}) - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma},
\end{aligned}$$

where (i) follows from the fact that $\zeta(z) \geq \varphi(z)$, (ii) follows from the definition of φ, i.e., $\rho'(z) = z\varphi(z)$, and (iii) follows from the definition of w_σ in Eq. (2), yielding the desired result. ■

The following result establishes the bound for the distance-to-measure described in Eq. (4).

Proposition 6.2. *For $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$, the persistence influence for the distance-to-measure function is given by*

$$\Psi\left(d_{\mathbb{P},m};\boldsymbol{x}\right) \leq \frac{2}{\sqrt{m}} \sup\left\{\left|f(\boldsymbol{x}) - \int_{\mathbb{R}^d} f(\boldsymbol{y}) d\mathbb{P}(\boldsymbol{y})\right| : \|\nabla f\|_{L_2(\mathbb{P})} \leq 1\right\}$$

where $\|\nabla f\|_{L_2(\mathbb{P})}$ is a modified, weighted Sobolev norm (Villani, 2003; Peyre, 2018).

Proof. From (Chazal et al., 2011, Theorem 3.5) the following stability result holds:

$$\left\|d_{\mathbb{P},m} - d_{\mathbb{P}^\epsilon_{\boldsymbol{x}},m}\right\|_\infty \leq \frac{1}{\sqrt{m}} W_2\left(\mathbb{P}, \mathbb{P}^\epsilon_{\boldsymbol{x}}\right).$$

From (Peyre, 2018, Theorem 1) we have that

$$W_2\left(\mathbb{P}, \mathbb{P}^\epsilon_{\boldsymbol{x}}\right) \leq 2 \|\mathbb{P} - \mathbb{P}^\epsilon_{\boldsymbol{x}}\|_{\dot{H}^{-1}(\mathbb{P})},$$

where the weighted, homogeneous Sobolev norm $\|\cdot\|_{\dot{H}^{-1}(\mu)}$ for a signed measure ν w.r.t. a positive measure μ is given by

$$\|\nu\|_{\dot{H}^{-1}(\mu)} = \sup\left\{\left|\int_{\mathbb{R}^d} f(\boldsymbol{x}) d\nu(\boldsymbol{x})\right| : \|\nabla f\|_{L_2(\mu)} \leq 1\right\}.$$

Observe that $\mathbb{P} - \mathbb{P}^\epsilon_{\boldsymbol{x}} = \epsilon\left(\delta_{\boldsymbol{x}} - \mathbb{P}\right)$ and since $\|\cdot\|_{\dot{H}^{-1}(\mu)}$ defines a norm, we have that

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left\|d_{\mathbb{P},m} - d_{\mathbb{P}^\epsilon_{\boldsymbol{x}},m}\right\|_\infty &\leq \frac{1}{\sqrt{m}} \lim_{\epsilon\to 0} \frac{1}{\epsilon} W_2\left(\mathbb{P}, \mathbb{P}^\epsilon_{\boldsymbol{x}}\right) \\
&\leq \frac{2}{\sqrt{m}} \lim_{\epsilon\to 0} \frac{1}{\epsilon} \|\epsilon\left(\delta_{\boldsymbol{x}} - \mathbb{P}\right)\|_{\dot{H}^{-1}(\mathbb{P})} \\
&= \frac{2}{\sqrt{m}} \|\left(\delta_{\boldsymbol{x}} - \mathbb{P}\right)\|_{\dot{H}^{-1}(\mathbb{P})} \\
&= \frac{2}{\sqrt{m}} \sup\left\{\left|f(\boldsymbol{x}) - \int_{\mathbb{R}^d} f(\boldsymbol{y}) d\mathbb{P}(\boldsymbol{y})\right| : \|\nabla f\|_{L_2(\mathbb{P})} \leq 1\right\}.
\end{aligned}$$

From the stability for persistence diagrams, we have that

$$\Psi\left(d_{\mathbb{P},m};\boldsymbol{x}\right) \leq \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left\|d_{\mathbb{P},m}-d_{\mathbb{P}^{\epsilon}_{\boldsymbol{x}},m}\right\|_{\infty}$$

and the result follows. ■

6.3 Proof of Theorem 4.2

Using the triangle inequality we can break our problem down as follows

$$\left\|f^{n}_{\rho,\sigma}-f\right\|_{\infty} \leq \underbrace{\left\|f_{\sigma}-f\right\|_{\infty}}_{\text{(a)}}+\underbrace{\left\|f_{\rho,\sigma}-f_{\sigma}\right\|_{\infty}}_{\text{(b)}},$$

where, $f_{\sigma}=\int_{\mathbb{R}^d}K_{\sigma}(\cdot,\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})$ is the population level KDE. For term (a), when $\mathbb{P}\in\mathcal{M}(\mathbb{R}^d)$, it is well known (Chen, 2017) that the approximation error for the KDE vanishes, i.e.,

$$\left\|f_{\sigma}-f\right\|_{\infty}\to 0,$$

as $\sigma\to 0$. So, it remains to verify that (b) vanishes, i.e., $\left\|f_{\rho,\sigma}-f_{\sigma}\right\|_{\infty}\to 0$. With this in mind, consider the map $T_{\sigma}:\mathcal{G}\to\mathcal{G}$ given by

$$T_{\sigma}(g)=\int\limits_{\mathbb{R}^d}\frac{\varphi\left(\left\|\Phi_{\sigma}\left(\boldsymbol{x}\right)-g\right\|_{\mathcal{H}_{\sigma}}\right)}{\int\limits_{\mathbb{R}^d}\varphi\left(\left\|\Phi_{\sigma}\left(\boldsymbol{x}\right)-g\right\|_{\mathcal{H}_{\sigma}}\right)d\mathbb{P}(\boldsymbol{x})}\Phi_{\sigma}(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x}).$$

Our approach to verifying that (b) vanishes is similar to Vandermeulen and Scott (2013, Lemma 9), where we show that the map T_{σ} is a contraction map when restricted to the subspace

$$\mathcal{Q}_{\sigma}\doteq B_{\mathcal{H}_{\sigma}}\left(\mathbf{0},\delta\nu_{\sigma}\right)\cap\mathcal{D}_{\sigma}.$$

A key difference is that we work with $\left\|\cdot\right\|_{\infty}$–norm, requiring us to obtain a sharper bound for the Lipschitz constant associated with the contraction.

For brevity, we adopt the notation $m(\boldsymbol{x},g)=\varphi\left(\left\|\Phi_{\sigma}\left(\boldsymbol{x}\right)-g\right\|_{\mathcal{H}_{\sigma}}\right)$. Kim and Scott (2012) show that $f_{\rho,\sigma}$ is a fixed point of the map T_{σ}, i.e., $T_{\sigma}(f_{\rho,\sigma})=f_{\rho,\sigma}$, and that f_{σ} is the image of $\mathbf{0}$ under T_{σ}, i.e., $T_{\sigma}(\mathbf{0})=f_{\sigma}$. Additionally, from Lemma A.2, we know that $\left\|f_{\sigma}\right\|_{\mathcal{H}_{\sigma}}\leq\delta\nu_{\sigma}$, for some $0<\delta<1$. Thus, we can rewrite $f_{\rho,\sigma}-f_{\sigma}=T_{\sigma}(f_{\rho,\sigma})-T_{\sigma}(\mathbf{0})$.

Let $g,h\in\mathcal{Q}_{\sigma}$. Then we have that

$$\begin{aligned}
T_{\sigma}(g)-T_{\sigma}(h) &= \int\limits_{\mathbb{R}^d}\frac{m(\boldsymbol{x},g)}{\int\limits_{\mathbb{R}^d}m(\boldsymbol{y},g)d\mathbb{P}(\boldsymbol{y})}\Phi_{\sigma}(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})-\int\limits_{\mathbb{R}^d}\frac{m(\boldsymbol{u},h)}{\int\limits_{\mathbb{R}^d}m(\boldsymbol{v},h)d\mathbb{P}(\boldsymbol{v})}\Phi_{\sigma}(\boldsymbol{u})d\mathbb{P}(\boldsymbol{u})\\
&= \frac{1}{\alpha\beta}\cdot\left(\beta\int\limits_{\mathbb{R}^d}m(\boldsymbol{x},g)\Phi_{\sigma}(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x})-\alpha\int\limits_{\mathbb{R}^d}m(\boldsymbol{u},h)\Phi_{\sigma}(\boldsymbol{u})d\mathbb{P}(\boldsymbol{x})\right)\\
&= \frac{1}{\alpha\beta}\cdot\xi, \qquad (9)
\end{aligned}$$

where $\alpha\doteq\int_{\mathbb{R}^d}m(\boldsymbol{y},g)d\mathbb{P}(\boldsymbol{y})\in\mathbb{R}$, $\beta\doteq\int_{\mathbb{R}^d}m(\boldsymbol{v},h)d\mathbb{P}(\boldsymbol{v})\in\mathbb{R}$ and the numerator $\xi\in\mathcal{H}_{\sigma}$.

By Tonelli's theorem

$$\begin{aligned}
\xi &= \beta \int_{\mathbb{R}^d} m(\boldsymbol{x}, g)\Phi_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) - \alpha \int_{\mathbb{R}^d} m(\boldsymbol{u}, h)\Phi_\sigma(\boldsymbol{u}) d\mathbb{P}(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^d} m(\boldsymbol{x}, g)\Phi_\sigma(\boldsymbol{x}) \left(\int_{\mathbb{R}^d} m(\boldsymbol{v}, h) d\mathbb{P}(\boldsymbol{v}) \right) d\mathbb{P}(\boldsymbol{x}) \\
&\qquad - \int_{\mathbb{R}^d} m(\boldsymbol{u}, h)\Phi_\sigma(\boldsymbol{u}) \left(\int_{\mathbb{R}^d} m(\boldsymbol{y}, g) d\mathbb{P}(\boldsymbol{y}) \right) d\mathbb{P}(\boldsymbol{x}) \\
&= \iint_{\mathbb{R}^d \times \mathbb{R}^d} m(\boldsymbol{x}, g) m(\boldsymbol{v}, h)\Phi_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{v}) d\mathbb{P}(\boldsymbol{x}) - \iint_{\mathbb{R}^d \times \mathbb{R}^d} m(\boldsymbol{u}, h) m(\boldsymbol{y}, g)\Phi_\sigma(\boldsymbol{u}) d\mathbb{P}(\boldsymbol{y}) d\mathbb{P}(\boldsymbol{u}) \\
&= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \Phi_\sigma(\boldsymbol{x}) \left[m(\boldsymbol{x}, g) m(\boldsymbol{y}, h) - m(\boldsymbol{x}, h) m(\boldsymbol{y}, g) \right] d\mathbb{P}(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{y}). \qquad (10)
\end{aligned}$$

Then by adding and subtracting $m(\boldsymbol{x}, h)m(\boldsymbol{y}, h)$ to the term inside, we get

$$\begin{aligned}
m(\boldsymbol{x}, g)m(\boldsymbol{y}, h) - m(\boldsymbol{x}, h)m(\boldsymbol{y}, g) = m(\boldsymbol{y}, h) \left\{ m(\boldsymbol{x}, g) - m(\boldsymbol{x}, h) \right\} \\
+ m(\boldsymbol{x}, h) \left\{ m(\boldsymbol{y}, h) - m(\boldsymbol{y}, g) \right\}.
\end{aligned}$$

Plugging this back into Eq. (10), we get $\xi = \xi_1 + \xi_2$ where

$$\begin{aligned}
\xi_1 &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \Phi_\sigma(\boldsymbol{x}) \left\{ m(\boldsymbol{x}, g) - m(\boldsymbol{x}, h) \right\} m(\boldsymbol{y}, h) d\mathbb{P}(\boldsymbol{y}) d\mathbb{P}(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^d} m(\boldsymbol{y}, h) d\mathbb{P}(\boldsymbol{y}) \int_{\mathbb{R}^d} \Phi_\sigma(\boldsymbol{x}) \left\{ m(\boldsymbol{x}, g) - m(\boldsymbol{x}, h) \right\} d\mathbb{P}(\boldsymbol{x}) \\
&= \beta \int_{\mathbb{R}^d} K_\sigma(\cdot, \boldsymbol{x}) \left\{ m(\boldsymbol{x}, g) - m(\boldsymbol{x}, h) \right\} d\mathbb{P}(\boldsymbol{x}) \\
&\overset{(i)}{=} \beta \cdot \left[\psi_\sigma * \left((m(\cdot, g) - m(\cdot, h)) f(\cdot) \right) \right],
\end{aligned}$$

where (i) follows from the fact that the kernel $K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \psi_\sigma(\boldsymbol{x} - \boldsymbol{y}) \doteq \sigma^{-d}\psi(\|\boldsymbol{x} - \boldsymbol{y}\|_2/\sigma)$ is translation invariant and f is the density associated with $\mathbb{P}$. Similarly,

$$\begin{aligned}
\xi_2 &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \Phi_\sigma(\boldsymbol{x}) m(\boldsymbol{x}, h) \left\{ m(\boldsymbol{y}, h) - m(\boldsymbol{y}, g) \right\} d\mathbb{P}(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{y}) \\
&= \int_{\mathbb{R}^d} \left[m(\boldsymbol{y}, h) - m(\boldsymbol{y}, g) \right] d\mathbb{P}(\boldsymbol{y}) \int_{\mathbb{R}^d} \Phi_\sigma(\boldsymbol{x}) m(\boldsymbol{x}, h) d\mathbb{P}(\boldsymbol{x}) \\
&\leq \|m(\cdot, h) - m(\cdot, g)\|_\infty \cdot \left[\psi_\sigma * (m(\cdot, h) f(\cdot)) \right].
\end{aligned}$$

The upper bound for $\|\xi_1\|_\infty$ is as follows

$$\begin{aligned}\|\xi_1\|_\infty &= \beta \|\psi_\sigma * ((m(\cdot, g) - m(\cdot, h))\, f(\cdot))\|_\infty \\ &\overset{(i)}{\leq} \beta \|\psi_\sigma\|_1 \|(m(\cdot, g) - m(\cdot, h))\, f(\cdot)\|_\infty \\ &\overset{(ii)}{\leq} \beta \|m(\cdot, g) - m(\cdot, h)\|_\infty \|f\|_\infty , \end{aligned} \tag{11}$$

where (i) follows from Young's inequality (Hewitt and Ross, 1979, Theorem 20.18) and (ii) follows from the fact that $\|fg\|_\infty \leq \|f\|_\infty \|g\|_\infty$. Similarly, for ξ_2 we have

$$\begin{aligned}\|\xi_2\|_\infty &\leq \|m(\cdot, h) - m(\cdot, g)\|_\infty \|\psi_\sigma * (m(\cdot, h) f(\cdot))\|_\infty \\ &\overset{(i)}{\leq} \|m(\cdot, h) - m(\cdot, g)\|_\infty \|\psi_\sigma\|_1 \|m(\cdot, h) f(\cdot)\|_\infty \\ &\overset{(ii)}{\leq} \|m(\cdot, h) - m(\cdot, g)\|_\infty \|m(\cdot, h)\|_\infty \|f\|_\infty . \end{aligned} \tag{12}$$

From the proof of (Vandermeulen and Scott, 2013, Lemma 9, Page 20–22), for $g, h \in \mathcal{Q}_\sigma$ for fixed constants $c_1, c_2 > 0$ we have the following two bounds:

$$\alpha, \beta \geq \frac{1}{c_1 \nu_\sigma}, \tag{13}$$

and

$$\|m(\cdot, h) - m(\cdot, g)\|_\infty \leq \|g - h\|_{\mathcal{H}_\sigma} c_2 \nu_\sigma^{-2}, \tag{14}$$

where the last inequality follows from the Lipschitz property of φ and fact that ρ is strictly convex. Additionally, for $c_3 = \|\rho'\|_\infty < \infty$ we have

$$\begin{aligned} m(\boldsymbol{x}, g) &= \varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) \\ &= \frac{\rho'\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}} \\ &\leq \frac{c_3}{\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}} \\ &\overset{(iii)}{\leq} \frac{c_3}{\left|\|\Phi_\sigma(\boldsymbol{x})\|_{\mathcal{H}_\sigma} - \|g\|_{\mathcal{H}_\sigma}\right|} \\ &= \frac{c_3}{(1-\delta)\nu_\sigma}, \end{aligned} \tag{15}$$

where (iii) follows from reverse triangle inequality. Plugging the bounds in equations (13), (14) and (15) back into equations (11) and (12) we get,

$$\|\xi_1\|_\infty + \|\xi_2\|_\infty \leq \|f\|_\infty \left(\beta c_2 \nu_\sigma^{-2} \|g - h\|_{\mathcal{H}_\sigma} + \frac{c_2 c_3}{(1-\delta)} \nu_\sigma^{-3} \|g - h\|_{\mathcal{H}_\sigma}\right).$$

Using this upper bound in Eq. (9) we get

$$\begin{aligned}
\|T_\sigma(g) - T_\sigma(h)\|_\infty &= \left\| \frac{\xi}{\alpha\beta} \right\|_\infty \\
&\leq \frac{\|\xi_1\|_\infty + \|\xi_2\|_\infty}{\alpha\beta} \\
&\overset{(iv)}{\leq} \|f\|_\infty \left(\frac{c_1 c_2}{c_1} \nu_\sigma^{-1} \|g-h\|_{\mathcal{H}_\sigma} + \frac{c_2 c_3}{c_1^2(1-\delta)} \nu_\sigma^{-1} \|g-h\|_{\mathcal{H}_\sigma} \right) \\
&\overset{(v)}{=} C\nu_\sigma^{-1} \|g-h\|_{\mathcal{H}_\sigma} \\
&\overset{(vi)}{\leq} C\nu_\sigma^{-1} \|g-h\|_\infty^{\frac{1}{2}},
\end{aligned}$$

where in (iv) we use Eq. (13), in (v) we use the fact that whenever $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$, we have $\|f\|_\infty < \infty$ and $C > 0$ is a constant depending only on c_1, c_2, c_3 and $\|f\|_\infty$. Additionally, (vi) holds through an application of Lemma A.1 to $g - h \in \mathcal{Q}_\sigma \subset \mathcal{D}_\sigma$. This confirms that T_σ is a contraction mapping. We use this to show that (b) vanishes as $\sigma \to 0$. Since $f_{\rho,\sigma}, \mathbf{0} \in \mathcal{Q}_\sigma$ and $f_{\rho,\sigma} - \mathbf{0} \in \mathcal{D}_\sigma$, we have that

$$\begin{aligned}
\|f_{\rho,\sigma} - f_\sigma\|_\infty &= \|T_\sigma(f_{\rho,\sigma}) - T_\sigma(\mathbf{0})\|_\infty \\
&\leq C\nu_\sigma^{-1} \|f_{\rho,\sigma} - \mathbf{0}\|_\infty^{\frac{1}{2}} \\
&= C\nu_\sigma^{-1} \|f_{\rho,\sigma}\|_\infty^{\frac{1}{2}}.
\end{aligned}$$

Using the triangle inequality $\|f_{\rho,\sigma}\|_\infty^{\frac{1}{2}} \leq \|f_{\rho,\sigma} - f_\sigma\|_\infty^{\frac{1}{2}} + \|f_\sigma\|_\infty^{\frac{1}{2}}$ we get

$$\begin{aligned}
\|f_{\rho,\sigma} - f_\sigma\|_\infty &\leq C\nu_\sigma^{-1} \left(\|f_{\rho,\sigma} - f_\sigma\|_\infty^{\frac{1}{2}} + \|f_\sigma\|_\infty^{\frac{1}{2}} \right) \\
&= C\nu_\sigma^{-1} \left(\|T_\sigma(f_{\rho,\sigma}) - T_\sigma(\mathbf{0})\|_\infty^{\frac{1}{2}} + \|f_\sigma\|_\infty^{\frac{1}{2}} \right) \\
&\leq C\nu_\sigma^{-1} \left(\left(C\nu_\sigma^{-1} \|f_{\rho,\sigma} - \mathbf{0}\|_\infty^{\frac{1}{2}} \right)^{\frac{1}{2}} + \|f_\sigma\|_\infty^{\frac{1}{2}} \right) \\
&= C^{\frac{3}{2}} \nu_\sigma^{-\frac{3}{2}} \|f_{\rho,\sigma}\|_\infty^{\frac{1}{4}} + C\nu_\sigma^{-1} \|f_\sigma\|_\infty^{\frac{1}{2}}, \qquad (16)
\end{aligned}$$

by using the contraction mapping twice. Observe that

$$\|f_{\rho,\sigma}\|_\infty \leq \nu_\sigma \|f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} \leq \delta\nu_\sigma^2,$$

where the first inequality follows from Lemma A.1 and the second inequality follows from the fact that $\|f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} \leq \delta\nu_\sigma$ since $f_{\rho,\sigma} \in \mathcal{Q}_\sigma$. Furthermore, $\|f_\sigma\|_\infty = \|\psi_\sigma * f\|_\infty \leq \|\psi_\sigma\|_\infty \|f\|_1 \leq \nu_\sigma$ from Young's inequality. By noting that $\nu_\sigma = \psi_\sigma(0) = \sigma^{-d}\psi(0)$, collecting these bounds back into Eq. (16) we get

$$\|f_{\rho,\sigma} - f_\sigma\|_\infty \leq C^{\frac{3}{2}} \delta^{\frac{1}{4}} \nu_\sigma^{-1} + C\nu_\sigma^{-\frac{1}{2}} \sqrt{\psi(0)}.$$

yielding that $\|f_{\rho,\sigma} - f_\sigma\|_\infty \to 0$ as $\sigma \to 0$, thereby verifying that (b) vanishes as $\sigma \to 0$. ■

6.4 Proof of Theorem 4.3

The proof proceeds in two steps: We first establish the uniform consistency for the robust KDE and then use the bottleneck stability to show consistency of the robust persistence diagrams in W_∞. From the stability theorem for persistence diagrams (Cohen-Steiner et al., 2007; Chazal et al., 2016), we have that $W_\infty\left(\mathsf{Dgm}\left(f^n_{\rho,\sigma}\right), \mathsf{Dgm}\left(f_{\rho,\sigma}\right)\right) \leq \left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty$. Thus, it suffices to show that $\left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty \xrightarrow{p} 0$ as $n \to \infty$. In order to prove the latter, we adapt the argmax consistency theorem (Van der Vaart, 2000, Theorem 5.7) for minimizers of a risk function.

Lemma 6.1 (Theorem 5.7, Van der Vaart (2000))**.** *Given a metric space $(\mathcal{G}, d)$, let $\mathcal{J}_n$ be random functions and $\mathcal{J}$ be a fixed function of $g \in \mathcal{G}$ such that for every $\epsilon > 0$,*

(1) $\inf_{g: d(g,g_0) \geq \epsilon} \mathcal{J}(g) > \mathcal{J}(g_0)$*, and*

(2) $\sup_{g \in \mathcal{G}} |\mathcal{J}_n(g) - \mathcal{J}(g)| \xrightarrow{p} 0$.

Then any sequence g_n satisfying $\mathcal{J}_n(g_n) < \mathcal{J}_n(g_0) + O_p(1)$ satisfies $d(g_n, g_0) \xrightarrow{p} 0$.

For $\mathcal{G} = \mathcal{H}_\sigma \cap \mathcal{D}_\sigma$, and $d(f^n_{\rho,\sigma}, f_{\rho,\sigma}) = \left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_\infty$, in order to establish uniform consistency of the robust KDE, as per Lemma 6.1, we need to verify that conditions (1) and (2) are satisfied.

Condition (1) follows from the strict convexity of $\mathcal{J}(g)$ in Proposition A.1. Specifically, Kim and Scott (2012) establish that assumptions $(\mathcal{A}1)-(\mathcal{A}3)$ guarantee the existence and uniqueness of $f_{\rho,\sigma} = \arg\inf_{g \in \mathcal{G}} \mathcal{J}(g)$. Then, for any $g \in \mathcal{G}$ such that $\|g - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} > \delta$, we have that $\mathcal{J}(g) > \mathcal{J}(f_{\rho,\sigma})$.

We now turn to verifying condition (2). Observe that $\sup_{g \in \mathcal{G}} |\mathcal{J}_n(g) - \mathcal{J}(g)|$ can be rewritten as the supremum of an empirical process, i.e.,

$$\sup_{g \in \mathcal{G}} |\mathcal{J}_n(g) - \mathcal{J}(g)| = \sup_{\ell_g \in \widetilde{\mathcal{F}}} |\mathbb{P}_n \ell_g - \mathbb{P}\ell_g| \doteq \|\mathbb{P}_n - \mathbb{P}\|_{\widetilde{\mathcal{F}}},$$

where $\widetilde{\mathcal{F}} = \{\ell_g : g \in \mathcal{G}\}$, and $\ell_g(\boldsymbol{x}) = \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)$. Verifying condition (2) reduces to showing that $\widetilde{\mathcal{F}}$ is a Glivenko-Cantelli class.

Define $\eta(\cdot) = \|\Phi_\sigma(\cdot) - g\|^2_{\mathcal{H}_\sigma}$ and let $\mathcal{F} = \{\eta_g : g \in \mathcal{G}\}$. For the continuous map $\xi : [0, \infty) \to [0, \infty)$ given by $\xi(t) = \rho(\sqrt{t})$, we have that

$$\xi \circ \mathcal{F} = \{\xi(f) : f \in \mathcal{F}\} = \{\xi \circ \eta_g(\cdot) : g \in \mathcal{G}\} = \left\{\rho(\|\Phi_\sigma(\cdot) - g\|_{\mathcal{H}_\sigma}) : g \in \mathcal{G}\right\} = \widetilde{\mathcal{F}}.$$

By the preservation theorem for Glivenko-Cantelli classes (Van Der Vaart and Wellner, 2000, Theorem 3), it holds that if $\mathcal{F}$ is a Glivenko-Cantelli class, then $\widetilde{\mathcal{F}}$ is also a Glivenko-Cantelli class. So verifying condition (2) reduces to verifying that $\mathcal{F}$ is a Glivenko-Cantelli class.

To this end, we first show that $F(\boldsymbol{x}_{1:n}) = F(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \sup_{g \in \mathcal{G}} |\mathbb{P}_n \eta_g - \mathbb{P}\eta_g| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ satisfies the self-bounded property for McDiarmid's inequality, i.e.,

$$\sup_{\boldsymbol{x}_i \neq \boldsymbol{x}'_i} \left|F(\boldsymbol{x}_{1:n}) - F(\boldsymbol{x}'_{1:n})\right| \leq \frac{1}{n} \sup_{\boldsymbol{x}_i, \boldsymbol{x}'_i} \sup_{g \in \mathcal{G}} \left(\|\Phi_\sigma(\boldsymbol{x}_i)\|^2_{\mathcal{H}_\sigma} + \left\|\Phi_\sigma(\boldsymbol{x}'_i)\right\|^2_{\mathcal{H}_\sigma} + 2\left|g(\boldsymbol{x}_i)\right| + 2\left|g(\boldsymbol{x}'_i)\right|\right).$$

Observe that $\|\Phi_\sigma(\boldsymbol{x})\|^2_{\mathcal{H}_\sigma} = K_\sigma(\boldsymbol{x},\boldsymbol{x}) \leq \|K_\sigma\|_\infty$ and $|g(\boldsymbol{x})| \leq \|g\|_\infty < \|K_\sigma\|_\infty$ by Lemma A.1. Thus, we have that

$$\sup_{\boldsymbol{x}_i \neq \boldsymbol{x}'_i} \left|F(\boldsymbol{x}_{1:n}) - F(\boldsymbol{x}'_{1:n})\right| \leq \frac{6\,\|K_\sigma\|_\infty}{n}.$$

From (Bartlett and Mendelson, 2002, Theorem 9), we have that with probability greater than $1-e^{-\delta}$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{3\delta\,\|K_\sigma\|_\infty}{n}}, \tag{17}$$

where $\mathfrak{R}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$ given by,

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}) &= \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,\|\Phi_\sigma(\boldsymbol{x}_i) - g\|^2_{\mathcal{H}_\sigma}\right|\right)\\
&\leq \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\left\{\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,\|\Phi_\sigma(\boldsymbol{x}_i)\|^2_{\mathcal{H}_\sigma}\right| + \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,\|g\|^2_{\mathcal{H}_\sigma}\right| + 2\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i)\right|\right\}\right)\\
&= ① + ② + ③.
\end{aligned}$$

Note that $\mathbb{E}_\epsilon\left(f(\epsilon_{1:n},\boldsymbol{x}_{1:n})\right) \doteq \mathbb{E}\left(f(\epsilon_{1:n},\boldsymbol{x}_{1:n})|\boldsymbol{x}_{1:n}\right)$ is the conditional expectation of the Rademacher random variables $\epsilon_1,\epsilon_2,\ldots,\epsilon_n$, keeping $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n$ fixed. First, we have that,

$$\begin{aligned}
① = \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,\|\Phi_\sigma(\boldsymbol{x}_i)\|^2_{\mathcal{H}_\sigma}\right|\right) &\overset{(i)}{=} \mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i K_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_i)\right|\\
&\overset{(ii)}{\leq} \sqrt{\mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i K_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_i)\right|^2}\\
&\leq \sqrt{\mathbb{E}_\epsilon\left(\frac{1}{n^2}\sum_{i,j}\epsilon_i\epsilon_j K_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_i)K_\sigma(\boldsymbol{x}_j,\boldsymbol{x}_j)\right)}\\
&\overset{(iii)}{=} \frac{1}{\sqrt{n}}\,\|K_\sigma\|_\infty,
\end{aligned}$$

where (i) follows from the absence of g inside the expectation, (ii) follows from Jensen's inequality and (iii) follows from the fact that $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for $i \neq j$. For the second term, we have

$$\begin{aligned}
② = \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,\|g\|^2_{\mathcal{H}_\sigma}\right|\right) = \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\|g\|^2_{\mathcal{H}_\sigma}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|\right)\\
\leq \sup_{g\in\mathcal{G}}\|g\|^2_{\mathcal{H}_\sigma}\sqrt{\mathbb{E}_\epsilon\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|^2},\\
\overset{(iv)}{\leq} \frac{1}{\sqrt{n}}\,\|K_\sigma\|_\infty,
\end{aligned}$$

where (iv) follows from the fact that $\|g\|_{\mathcal{H}_\sigma}^2 \leq \|K_\sigma\|_\infty$. Lastly, we have

$$
\begin{aligned}
③ = 2\mathbb{E}_\epsilon \left(\sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i g(\boldsymbol{x}_i) \right| \right) &\overset{(v)}{=} 2\mathbb{E}_\epsilon \left(\sup_{g\in\mathcal{G}} \left| \left\langle g, \frac{1}{n} \sum_{i=1}^n \epsilon_i K_\sigma(\cdot, \boldsymbol{x}_i) \right\rangle_{\mathcal{H}_\sigma} \right| \right) \\
&\overset{(vi)}{\leq} 2\mathbb{E}_\epsilon \left(\sup_{g\in\mathcal{G}} \|g\|_{\mathcal{H}_\sigma} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i K_\sigma(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}_\sigma} \right) \\
&= 2 \sup_{g\in\mathcal{G}} \|g\|_{\mathcal{H}_\sigma} \, \mathbb{E}_\epsilon \left(\sqrt{\frac{1}{n^2} \sum_{i,j} \epsilon_i \epsilon_j K_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j)} \right) \\
&\overset{(vii)}{\leq} 2 \frac{\|K_\sigma\|_\infty^{\frac{1}{2}}}{n} \sqrt{\mathbb{E}_\epsilon \left(\sum_{i,j} \epsilon_i \epsilon_j K_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) \right)} \\
&\overset{(viii)}{\leq} \frac{2}{\sqrt{n}} \|K_\sigma\|_\infty ,
\end{aligned}
$$

where (v) follows from the reproducing property, (vi) is obtained from Cauchy-Schwarz inequality, (vii) follows from Jensen's inequality, and (viii) follows from the fact that $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for $i \neq j$. Collecting these three inequalities, we have

$$
\mathfrak{R}_n(\mathcal{F}) = ① + ② + ③ \leq \frac{4}{\sqrt{n}} \|K_\sigma\|_\infty .
$$

Plugging this into Eq. (17), we have with probability greater than $1 - e^{-\delta}$,

$$
\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \frac{8 \|K_\sigma\|_\infty}{\sqrt{n}} + \sqrt{\frac{3\delta \|K_\sigma\|_\infty}{n}},
$$

which implies that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ as $n \to \infty$, implying that $\mathcal{F}$ is a Glivenko-Cantelli class. The result, therefore, follows from Lemma 6.1. ■

6.5 Proof of Theorem 4.4

For $g \in \mathcal{G}$ define the random fluctuation w.r.t. $f_{\rho,\sigma}$ as

$$
\Delta(\boldsymbol{X}, g) = \left(\ell_g(\boldsymbol{X}) - \ell_{f_{\rho,\sigma}}(\boldsymbol{X}) \right) - \left(\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}) \right).
$$

The fluctuation process is an empirical process defined as

$$
\begin{aligned}
\Delta_n(g) = \mathbb{P}_n \Delta(\boldsymbol{X}, g) &= \left(\mathcal{J}_n(g) - \mathcal{J}_n(f_{\rho,\sigma}) \right) - \left(\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}) \right), \\
&= \mathbb{P}_n \left(\ell_g - \ell_{f_{\rho,\sigma}} \right) - \mathbb{P} \left(\ell_g - \ell_{f_{\rho,\sigma}} \right).
\end{aligned}
$$

We first show that the behaviour of $\left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_{\mathcal{H}_\sigma}$ is controlled by the tail behaviour of the supremum of the fluctuation process. To this end, for $\delta > 0$, let

$$
\mathcal{G}_\delta = \left\{ g \in \mathcal{G} : \|g - f_{\rho,\sigma}\|_{\mathcal{H}_\sigma} \leq \delta \right\} = B_{\mathcal{H}_\sigma}(f_{\rho,\sigma}, \delta) \cap \mathcal{D}_\sigma.
$$

Suppose $f^n_{\rho,\sigma}$ is such that $\left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} > \delta$, then, for sufficiently small $\lambda \in (0,1)$ such that $g = \lambda f^n_{\rho,\sigma} + (1-\lambda) f_{\rho,\sigma} \in \mathcal{G}_\delta$, we have that

$$\begin{aligned}\mathcal{J}_n(g) - \mathcal{J}_n(f_{\rho,\sigma}) &\overset{(i)}{<} \lambda \mathcal{J}_n(f^n_{\rho,\sigma}) + (1-\lambda)\mathcal{J}_n(f_{\rho,\sigma}) - \mathcal{J}_n(f_{\rho,\sigma}) \\ &= \lambda \cdot \left(\mathcal{J}_n(f^n_{\rho,\sigma}) - \mathcal{J}_n(f_{\rho,\sigma})\right) \overset{(ii)}{\leq} 0, \end{aligned} \tag{18}$$

where (i) follows from the strict convexity of $\mathcal{J}_n$ (Proposition A.1), and (ii) follows from the fact that $f^n_{\rho,\sigma} = \arg\inf_{g\in\mathcal{G}} \mathcal{J}_n(g)$. From Proposition A.1, we also know that $\mathcal{J}$ is strongly convex such that

$$\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}) \geq \frac{\mu}{2} \left\|g - f_{\rho,\sigma}\right\|^2_{\mathcal{H}_\sigma}. \tag{19}$$

Combining equations (18) and (19) we have

$$\begin{aligned}\frac{\mu}{2} \left\|g - f_{\rho,\sigma}\right\|^2_{\mathcal{H}_\sigma} &\leq \mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}), \\ &= -\left\{(\mathcal{J}_n(g) - \mathcal{J}_n(f_{\rho,\sigma})) - (\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}))\right\} + (\mathcal{J}_n(g) - \mathcal{J}_n(f_{\rho,\sigma})) \\ &\leq -\Delta_n(g) \leq \sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)|. \end{aligned}$$

By taking the supremum of the left hand side in the above inequality over all $g \in \mathcal{G}_\delta$ we have

$$\sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)| \geq \frac{\mu}{2}\delta^2 \tag{20}$$

This implies that whenever $\left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} > \delta$ holds, then the condition in Eq. (20) holds. Therefore,

$$\mathbb{P}^{\otimes n}\left\{\boldsymbol{X}_{1:n} : \left\|f^n_{\rho,\sigma} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} > \delta\right\} \leq \mathbb{P}^{\otimes n}\left\{\boldsymbol{X}_{1:n} : \sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)| \geq \frac{\mu}{2}\delta^2\right\}. \tag{21}$$

We now study the behaviour of the r.h.s. in Eq. (21) using tools from empirical process theory. First, we show that $F(\boldsymbol{x}_{1:n}) = F(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)|$ satisfies the self-bounding property.

$$\begin{aligned}\sup_{\boldsymbol{x}_i \neq \boldsymbol{x}'_i} \left|F(\boldsymbol{x}_{1:n}) - F(\boldsymbol{x}'_{1:n})\right| &= \sup_{\boldsymbol{x}_i\neq\boldsymbol{x}'_i} \left|\sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)| - \sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)|\right|, \\ &\leq \sup_{\boldsymbol{x}_i\neq\boldsymbol{x}'_i} \sup_{g\in\mathcal{G}_\delta} \left|\Delta_n(g) - \Delta'_n(g)\right|, \\ &= \frac{1}{n} \sup_{\boldsymbol{x}_i\neq\boldsymbol{x}'_i} \sup_{g\in\mathcal{G}_\delta} \left|\left(\ell_g(\boldsymbol{x}_i) - \ell_{f_{\rho,\sigma}}(\boldsymbol{x}_i)\right) - \left(\ell_g(\boldsymbol{x}'_i) - \ell_{f_{\rho,\sigma}}(\boldsymbol{x}'_i)\right)\right|, \\ &\leq \frac{1}{n} \sup_{\boldsymbol{x}_i\neq\boldsymbol{x}'_i} \sup_{g\in\mathcal{G}_\delta} \left|\left(\ell_g(\boldsymbol{x}_i) - \ell_{f_{\rho,\sigma}}(\boldsymbol{x}_i)\right)\right| + \left|\left(\ell_g(\boldsymbol{x}'_i) - \ell_{f_{\rho,\sigma}}(\boldsymbol{x}'_i)\right)\right|, \\ &\overset{(i)}{\leq} \frac{1}{n} \sup_{g\in\mathcal{G}_\delta} 2M \left\|g - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} = \frac{2M\delta}{n}, \end{aligned}$$

where (i) follows from Proposition A.1 that ℓ_g is M-Lipschitz w.r.t. $\|\cdot\|_{\mathcal{H}_\sigma}$. Therefore, from McDiarmid's inequality (Vershynin, 2018, Theorem 2.9.1) we have

$$\mathbb{P}^{\otimes n}\left\{\boldsymbol{X}_{1:n} : \sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)| > \mathbb{E}\sup_{g\in\mathcal{G}_\delta} |\Delta_n(g)| + \epsilon\right\} \leq \exp\left(-\frac{n\epsilon^2}{2M^2\delta^2}\right). \tag{22}$$

Next, we find an upper bound for the expected supremum of the fluctuation process. In order to do so, we first show that $\Delta_n(g)$ has sub-Gaussian increments. For fixed $g, h \in \mathcal{G}$ we have that $\mathbb{E}\left(\Delta(\boldsymbol{X}, g) - \Delta(\boldsymbol{X}, h)\right) = 0$ and

$$\left|\Delta(\boldsymbol{X}, g) - \Delta(\boldsymbol{X}, h)\right| \leq \left|\ell_g(\boldsymbol{X}) - \ell_h(\boldsymbol{X})\right| - \left|\mathcal{J}(g) - \mathcal{J}(h)\right| \leq 2M \left\|g - h\right\|_{\mathcal{H}_\sigma}.$$

Since $\left|\Delta(\boldsymbol{X}, g) - \Delta(\boldsymbol{X}, h)\right|$ is bounded, it is, therefore, sub-Gaussian and from Vershynin (2018, Example 2.5.8), we have that the sub-Gaussian norm $\left\|\Delta(\boldsymbol{X}, g) - \Delta(\boldsymbol{X}, h)\right\|_{\psi_2} \leq 2cM \left\|g - h\right\|_{\mathcal{H}_\sigma}$ for $c > 1/\sqrt{\log 2}$. Consequently, the fluctuation process has sub-Gaussian increments with respect to the metric $\left\|g - h\right\|_{\mathcal{H}_\sigma}$, i.e.,

$$\left\|\Delta_n(g) - \Delta_n(h)\right\|_{\psi_2} \leq \frac{1}{n}\sqrt{\sum_{i=1}^{n} \left\|\Delta(\boldsymbol{X}_i, g) - \Delta(\boldsymbol{X}_i, h)\right\|_{\psi_2}^2} \leq \frac{M}{\sqrt{n}} \left\|g - h\right\|_{\mathcal{H}_\sigma}.$$

From the generalized entropy integral (Srebro et al., 2010, Lemma A.3), for a fixed constant $\gamma > 12/\sqrt{\log 2}$ we have

$$\mathbb{E} \sup_{g \in \mathcal{G}_\delta} \left|\Delta_n(g)\right| \leq \inf_{\alpha > 0} \left\{ 2\alpha + \frac{\gamma M}{\sqrt{n}} \int_\alpha^\delta \sqrt{\log \mathcal{N}\left(\mathcal{G}_\delta, \left\|\cdot\right\|_{\mathcal{H}_\sigma}, \epsilon\right)} d\epsilon \right\}, \tag{23}$$

where $\mathcal{N}\left(\mathcal{G}_\delta, d, \epsilon\right)$ is the ϵ-covering number of the class $\mathcal{G}_\delta$ with respect to metric d.

We now turn our attention to finding an upper bound for $\mathcal{N}\left(\mathcal{G}_\delta, d, \epsilon\right)$. Note that if $\mathcal{B}_{\mathcal{H}_\sigma}$ is a unit ball in the RKHS, then

$$\begin{aligned}
\log \mathcal{N}\left(\mathcal{G}_\delta, \left\|\cdot\right\|_{\mathcal{H}_\sigma}, \epsilon\right) &= \log \mathcal{N}\left(\mathcal{B}_{\mathcal{H}_\sigma} \cap \mathcal{D}_\sigma, \left\|\cdot\right\|_{\mathcal{H}_\sigma}, \frac{\epsilon}{\delta}\right) \\
&\overset{(i)}{\leq} \log \mathcal{N}\left(\mathcal{B}_{\mathcal{H}_\sigma} \cap \mathcal{D}_\sigma, \left\|\cdot\right\|_{\infty}, \left(\frac{\epsilon}{\delta}\right)^2\right) \\
&\leq \log \mathcal{N}\left(\mathcal{B}_{\mathcal{H}_\sigma}, \left\|\cdot\right\|_{\infty}, \left(\frac{\epsilon}{\delta}\right)^2\right),
\end{aligned}$$

where (i) follows from Lemma A.1 that $\left\|g - h\right\|_{\mathcal{H}_\sigma}^2 \leq \left\|g - h\right\|_{\infty}$. When the entropy numbers $e_n\left(\mathrm{id} : \mathcal{H}_\sigma \to L_\infty(\mathcal{X})\right)$ satisfy the assumption, from (Steinwart and Christmann, 2008, Lemma 6.21) we have

$$\log \mathcal{N}\left(\mathcal{B}_{\mathcal{H}_\sigma}, \left\|\cdot\right\|_{\infty}, \left(\frac{\epsilon}{\delta}\right)^2\right) \leq \left(\frac{a_\sigma \delta^2}{\epsilon^2}\right)^{2p}.$$

Plugging this into Eq. (23), we have that

$$\mathbb{E} \sup_{g \in \mathcal{G}_\delta} \left|\Delta_n(g)\right| \leq \inf_{\alpha > 0} \left\{ 2\alpha + \frac{\gamma M a_\sigma \delta^{2p}}{\sqrt{n}} \int_\alpha^\delta \epsilon^{-2p} d\epsilon \right\} = \inf_{\alpha > 0} T(\alpha),$$

where $T(\alpha)$ is given by

$$T(\alpha) = \begin{cases} 2\alpha + \gamma M \delta \sqrt{\frac{a_\sigma}{n}} \log\left(\frac{\delta}{\alpha}\right) & \text{if } p = \frac{1}{2}, \\ 2\alpha + \frac{\gamma M}{(1-2p)\sqrt{n}} \left(\delta - \delta^{2p} \alpha^{1-2p}\right) & \text{if } 0 < p \neq \frac{1}{2} < 1. \end{cases}$$

At the value α_0 where $T(\alpha_0) = \inf_{\alpha>0} T(\alpha)$, we have

$$T(\alpha_0) = \begin{cases} \gamma C a_\sigma^{\frac{1}{2}} \cdot \frac{M\delta \log(n)}{\sqrt{n}} & \text{if } p = \frac{1}{2}, \\ \\ \frac{\gamma a_\sigma^p}{(1-2p)} \cdot \frac{M\delta}{\sqrt{n}} - \frac{K p a_\sigma^{\frac{1}{2}}}{(1-2p)} \cdot \frac{M\delta}{n^{1/4p}} & \text{if } 0 < p \neq \frac{1}{2} < 1, \end{cases} \tag{24}$$

for some fixed constant $C > 3 - \log(9a)$. Observe that when $0 < p < \frac{1}{2}$, the last term of Eq. (24) is negative, and similarly when $\frac{1}{2} < p < 1$, the first term is negative. From this, we have that $T(\alpha_0) \leq M\delta\xi(n,p)$ where

$$\xi(n,p) = \begin{cases} \frac{\gamma a_\sigma^p}{(1-2p)} \cdot \frac{1}{\sqrt{n}} & \text{if } 0 < p < \frac{1}{2}, \\ \\ \gamma C a_\sigma^{\frac{1}{2}} \cdot \frac{\log(n)}{\sqrt{n}} & \text{if } p = \frac{1}{2}, \\ \\ \frac{\gamma p a_\sigma^{\frac{1}{2}}}{2p-1} \cdot \frac{1}{n^{1/4p}} & \text{if } \frac{1}{2} < p < 1. \end{cases}$$

Plugging this into Eq. (22), we have that with probability greater than $1 - e^{-t}$,

$$\sup_{g \in \mathcal{G}_\delta} |\Delta_n(g)| < M\delta\xi(n,p) + M\delta\sqrt{\frac{2t}{n}}. \tag{25}$$

From Eq. (21), this implies that

$$\mathbb{P}^{\otimes n}\left\{ \boldsymbol{X}_{1:n} : \left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_{\mathcal{H}_\sigma} > \delta \right\} \leq \mathbb{P}^{\otimes n}\left\{ \boldsymbol{X}_{1:n} : \sup_{g \in \mathcal{G}_\delta} \Delta_n(g) \geq \frac{\mu\delta^2}{2} \right\}.$$

Thus, in Eq. (25), by letting

$$\frac{\mu\delta^2}{2} = \left(M\delta\xi(n,p) + M\delta\sqrt{\frac{2t}{n}} \right),$$

we have that with probability greater than $1 - e^{-t}$,

$$\left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_{\mathcal{H}_\sigma} \leq \frac{2M}{\mu}\left(\xi(n,p) + \sqrt{\frac{2t}{n}} \right).$$

Observe that $\left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_\infty \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_{\mathcal{H}_\sigma}$. For $0 < \alpha < 1$, by choosing δ_n as

$$\delta_n = \frac{2M \|K_\sigma\|_\infty^{\frac{1}{2}}}{\mu}\left(\xi(n,p) + \sqrt{\frac{2\log(1/\alpha)}{n}} \right),$$

we have that

$$\mathbb{P}^{\otimes n}\left\{ \boldsymbol{X}_{1:n} : \left\| f_{\rho,\sigma}^n - f_{\rho,\sigma} \right\|_\infty \leq \delta_n \right\} > 1 - \alpha.$$

From the stability of persistence diagrams in Proposition 2.1, this implies that

$$\mathbb{P}^{\otimes n}\left\{ \boldsymbol{X}_{1:n} : W_\infty\left(\mathsf{Dgm}\left(f_{\rho,\sigma}^n \right), \mathsf{Dgm}\left(f_{\rho,\sigma} \right) \right) > \delta_n \right\} \leq \alpha,$$

yielding the desired result. ■

7 Conclusion & Discussion

In this paper, we proposed a statistically consistent robust persistent diagram using RKHS-based robust KDE as the filter function. By generalizing the notion of influence function to the space of persistence diagrams, we mathematically and empirically demonstrated the robustness of the proposed method to that of persistence diagrams induced by other filter functions such as KDE. Through numerical experiments, we demonstrated the advantage of using robust persistence diagrams in machine learning applications.

We would like to highlight that most of the theoretical results of this paper crucially hinge on the loss function being convex. As a future direction, we would like to generalize the current results to non-convex loss functions, which potentially yield more robust persistence diagrams. Another important direction we intend to explore is to develop robust persistence diagrams induced by other types of robust density estimators, which enables to understand the power and limitation of the proposed method.

Acknowledgements

KF is supported in part by JST CREST Grant Number JPMJCR15D3, Japan. SK is partially supported by JSPS KAKENHI Grant Number JP16H02792.

References

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence Images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1): 218–252, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Paul Bendich, Taras Galkovskyi, and John Harer. Improving homology estimates with random walks. *Inverse Problems*, 27(12):124002, 2011.

Paul Bendich, James S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *The Annals of Applied Statistics*, 10(1):198, 2016.

Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodriguez. A weighted k-Nearest Neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5: 204–237, 2011.

A Braides. Γ-convergence for Beginners. *Oxford Lecture Series in Mathematics and its Applications*, 22, 2002.

Rickard Brüel-Gabrielsson, Vignesh Ganapathi-Subramanian, Primoz Skraba, and Leonidas J Guibas. Topology-aware surface reconstruction for point clouds. *arXiv preprint arXiv:1811.12543*, 2018.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.

Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013.

Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *Journal of Machine Learning Research*, 18(1):5845–5884, 2017.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2573–2582, 2019.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

Gianni Dal Maso. *An Introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media, 2012.

Vincent Divol and Théo Lacombe. Understanding the topology and the geometry of the persistence diagram space via optimal partial transport. *arXiv preprint arXiv:1901.03048*, 2019.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2000.

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

Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, and Vidit Nanda. A topological measurement of protein compressibility. *Japan Journal of Industrial and Applied Mathematics*, 32(1):1–17, 2015.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust Statistics: The Approach Based on Influence Functions*, volume 196. John Wiley & Sons, 2011.

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

E. Hewitt and K.A. Ross. *Abstract Harmonic Analysis: Volume I*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1979. ISBN 0387941908.

Peter J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2004. ISBN 9780471650720.

JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13(Sep):2529–2565, 2012.

Longin Jan Latecki, Rolf Lakamper, and T Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 424–429. IEEE, 2000.

Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.

Rémi Peyre. Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

Jeff M Phillips, Bei Wang, and Yan Zheng. Geometric inference on kernel density estimates. In *31st International Symposium on Computational Geometry (SoCG 2015)*, 2015.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

Aad Van Der Vaart and Jon A Wellner. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High dimensional probability II*, pages 115–133. Springer, 2000.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Robert Vandermeulen and Clayton Scott. Consistency of robust kernel density estimators. In *Conference on Learning Theory*, pages 568–591, 2013.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Cédric Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.

Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5: 501–532, 2018.

Xin Xu, Jessi Cisewski-Kehe, Sheridan B Green, and Daisuke Nagai. Finding cosmic voids and filament loops using topological data analysis. *Astronomy and Computing*, 27:34–52, 2019.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

A Supplementary Results

In this section, we establish some results which play a key role in the proofs presented in Section 6.

A.1 Properties of the Risk Functional $\mathcal{J}(g)$

We establish some important properties of the risk functional, given by

$$\mathcal{J}(g) = \int_{\mathbb{R}^d} \ell_g(\boldsymbol{x})\; d\mathbb{P}(\boldsymbol{x}) = \int_{\mathbb{R}^d} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)\; d\mathbb{P}(\boldsymbol{x}).$$

The following result establishes that some important properties of the robust loss ρ carry forward to $\mathcal{J}(g)$. (i) The Lipschitz property of ρ is inherited by $\mathcal{J}(g)$, (ii) the convexity of ρ is strengthened to guarantee that $\mathcal{J}(g)$ is strictly convex, and (iii) $\mathcal{J}(g)$ is strongly convex with respect to the $\|\cdot\|_{\mathcal{H}_\sigma}$–norm around its minimizer.

Proposition A.1 (Convexity and Lipchitz properties of $\mathcal{J}$)**.** *Under assumptions* $(\mathcal{A}1) - (\mathcal{A}3)$*,*

(i) The risk functionals $\mathcal{J}(g)$ and $\mathcal{J}_n(g)$ are M-Lipschitz w.r.t. $\|\cdot\|_{\mathcal{H}_\sigma}$.

(ii) Furthermore, if ρ is convex, $\mathcal{J}(g)$ and $\mathcal{J}_n(g)$ are strictly convex.

(iii) Additionally, under assumption $(\mathcal{A}4)$*, for $f_{\rho,\sigma} = \arg\inf_{g\in\mathcal{G}} \mathcal{J}(g)$, the risk functional satisfies the strong convexity condition*

$$\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}) \geq \frac{\mu}{2}\, \|f_{\rho,\sigma} - g\|^2_{\mathcal{H}_\sigma}\,,$$

for $\mu = 2\min\left\{\varphi\left(2\,\|K_\sigma\|_\infty^{\frac{1}{2}}\right), \rho''\left(2\,\|K_\sigma\|_\infty^{\frac{1}{2}}\right)\right\}$.

Proof. **Lipschitz property.** Observe that,

$$\begin{aligned}
|\ell_{g_1}(\boldsymbol{x}) - \ell_{g_2}(\boldsymbol{x})| &= \left|\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma}\right) - \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}\right)\right| \\
&\leq M\left|\|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma} - \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}\right| \\
&\leq M\,\|g_1 - g_2\|_{\mathcal{H}_\sigma}\;,
\end{aligned}$$

where the first inequality follows from the fact that ρ is M-Lipschitz and the last inequality follows from reverse triangle inequality. This shows that the loss functions $\ell_g(\cdot)$ are M-Lipschitz with respect to g. For the risk functionals, we have that,

$$\begin{aligned}
|\mathcal{J}(g_1) - \mathcal{J}(g_2)| &= \left|\int_{\mathbb{R}^d} \left(\ell_{g_1}(\boldsymbol{x}) - \ell_{g_2}(\boldsymbol{x})\right) d\mathbb{P}(\boldsymbol{x})\right| \\
&\leq \int_{\mathbb{R}^d} \left|\ell_{g_1}(\boldsymbol{x}) - \ell_{g_2}(\boldsymbol{x})\right| d\mathbb{P}(\boldsymbol{x}) \\
&\leq M\,\|g_1 - g_2\|_{\mathcal{H}_\sigma}\,,
\end{aligned}$$

where the first inequality follows from Jensen's inequality. This verifies that $\mathcal{J}(g)$ is M-Lipchitz. The proof for $\mathcal{J}_n(g)$ is identical.

Strict Convexity. We begin by establishing that for translation invariant kernels $\|\Phi_\sigma(\boldsymbol{x}) - \cdot\|_{\mathcal{H}_\sigma}$ is strictly convex. Suppose $g_1, g_2 \in \mathcal{H}_\sigma \cap \mathcal{D}_\sigma$ and $\lambda \in (0,1)$, and let $g = (1-\lambda)g_1 + \lambda g_2$. Then

$$\begin{aligned}\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}^2 &= \|(1-\lambda)(\Phi_\sigma(\boldsymbol{x}) - g_1) + \lambda(\Phi_\sigma(\boldsymbol{x}) - g_2)\|_{\mathcal{H}_\sigma}^2 \\ &= (1-\lambda)^2 \|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma}^2 \\ &\quad + \lambda^2 \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}^2 + 2\lambda(1-\lambda)\Big\langle \Phi_\sigma(\boldsymbol{x}) - g_1, \Phi_\sigma(\boldsymbol{x}) - g_2 \Big\rangle_{\mathcal{H}_\sigma}. \end{aligned} \tag{A.1}$$

From Cauchy-Schwarz inequality, we know that

$$\Big\langle \Phi_\sigma(\boldsymbol{x}) - g_1, \Phi_\sigma(\boldsymbol{x}) - g_2 \Big\rangle_{\mathcal{H}_\sigma} \leq \|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma} \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}.$$

In the following, we argue that for translation invariant kernels,

$$\Big\langle \Phi_\sigma(\boldsymbol{x}) - g_1, \Phi_\sigma(\boldsymbol{x}) - g_2 \Big\rangle_{\mathcal{H}_\sigma} < \|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma} \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}, \tag{A.2}$$

for $g_1 \neq g_2$. On the contrary, suppose

$$\Big\langle \Phi_\sigma(\boldsymbol{x}) - g_1, \Phi_\sigma(\boldsymbol{x}) - g_2 \Big\rangle_{\mathcal{H}_\sigma} = \|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma} \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}$$

holds. Then this implies that there is a function $a(\boldsymbol{x})$, depending only on g_1 and g_2, such that $a(\boldsymbol{x}) \neq 0$ for $\boldsymbol{x} \in \mathbb{R}^d$ and

$$\Phi_\sigma(\boldsymbol{x}) - g_1 = a(\boldsymbol{x})\left(\Phi_\sigma(\boldsymbol{x}) - g_2\right).$$

Rearranging the terms this implies that

$$\Phi_\sigma(\boldsymbol{x}) = \frac{g_1 - a(\boldsymbol{x})g_2}{1 - a(\boldsymbol{x})} = (1 + b(\boldsymbol{x}))g_1 + b(\boldsymbol{x})g_2,$$

where $b(\boldsymbol{x}) = -a(\boldsymbol{x})/(1-a(\boldsymbol{x}))$ also does not vanish on $\boldsymbol{x} \in \mathbb{R}^d$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, from the reproducing property we have

$$\begin{aligned} K_\sigma(\boldsymbol{x}, \boldsymbol{y}) &= \Big\langle \Phi_\sigma(\boldsymbol{x}), \Phi_\sigma(\boldsymbol{y}) \Big\rangle_{\mathcal{H}_\sigma} \\ &= \Big\langle g_1 + b(\boldsymbol{x})(g_1 + g_2), g_1 + b(\boldsymbol{y})(g_1 + g_2) \Big\rangle_{\mathcal{H}_\sigma} \\ &= b(\boldsymbol{x})b(\boldsymbol{y}) \|g_1 + g_2\|_{\mathcal{H}_\sigma}^2 + (b(\boldsymbol{x}) + b(\boldsymbol{y})) \langle g_1, g_1 + g_2 \rangle_{\mathcal{H}_\sigma} + \|g_1\|_{\mathcal{H}_\sigma}^2. \end{aligned}$$

Note that because the kernel is translation invariant, i.e., $K_\sigma(\boldsymbol{x}, \boldsymbol{x}) = K_\sigma(\boldsymbol{y}, \boldsymbol{y}) = \sigma^{-d}\psi(0)$, this must imply that

$$\begin{aligned} 0 &= \left(b(\boldsymbol{x})^2 - b(\boldsymbol{y})^2\right) \|g_1 + g_2\|_{\mathcal{H}_\sigma}^2 + 2(b(\boldsymbol{x}) - b(\boldsymbol{y}))\langle g_1, g_1 + g_2 \rangle_{\mathcal{H}_\sigma} \\ &= (b(\boldsymbol{x}) - b(\boldsymbol{y})) \left((b(\boldsymbol{x}) + b(\boldsymbol{y})) \|g_1 + g_2\|_{\mathcal{H}_\sigma}^2 + 2\langle g_1, g_1 + g_2 \rangle_{\mathcal{H}_\sigma} \right). \end{aligned}$$

Since $b(\boldsymbol{x})$ and $b(\boldsymbol{y})$ are nonvanishing, the above equation is satisfied only when $b(\boldsymbol{x}) = b(\boldsymbol{y})$. This implies that $K_\sigma(\boldsymbol{x}, \boldsymbol{y})$ is constant for all $\boldsymbol{y}$, giving us a contradiction. Thus, we have that Eq. (A.2) holds. Plugging this back in Eq. (A.1) we get that for $\lambda \in (0,1)$ and $g = (1-\lambda)g_1 + \lambda g_2$,

$$\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma} < (1-\lambda) \|\Phi_\sigma(\boldsymbol{x}) - g_1\|_{\mathcal{H}_\sigma} + \lambda \|\Phi_\sigma(\boldsymbol{x}) - g_2\|_{\mathcal{H}_\sigma}.$$

Since, ρ is strictly increasing and convex, this implies that

$$\ell_g(\boldsymbol{x}) < (1-\lambda)\ell_{g_1}(\boldsymbol{x}) + \lambda\ell_{g_2}(\boldsymbol{x}).$$

The map $\ell_g(\cdot) \mapsto \mathbb{P}\ell_g$ is a linear operator, and ℓ_g is strictly convex in g, this implies that $\mathcal{J}(g)$ is also strictly convex in g. The same holds for $\mathcal{J}_n(g)$.

Strong Convexity around the minimizer. We now turn our attention to the strong convexity property. For this, we first show that $\mathcal{J}(g)$ is twice Gâteaux differentiable. Let $g, h \in \mathcal{G}$, then the second Gâteaux derivative of the loss $\ell_g(\boldsymbol{x}) = \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)$ at g in the direction h is given by,

$$\begin{aligned}
\delta^2\ell(\boldsymbol{x}, g; h) &= \frac{d^2}{d\alpha^2}\ell(\boldsymbol{x}, g+\alpha h)\Big|_{\alpha=0} \\
&= \frac{d^2}{d\alpha^2}\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g - \alpha h\|_{\mathcal{H}_\sigma}\right)\Big|_{\alpha=0} \\
&= \frac{d}{d\alpha}\left[\varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - g - \alpha h\|_{\mathcal{H}_\sigma}\right)\left(-\langle\Phi_\sigma(\boldsymbol{x}) - g, h\rangle_{\mathcal{H}_\sigma} + \alpha\|h\|^2_{\mathcal{H}_\sigma}\right)\right]\Big|_{\alpha=0} \\
&= \varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)\|h\|^2_{\mathcal{H}_\sigma} + \langle\Phi_\sigma(\boldsymbol{x}) - g, h\rangle^2_{\mathcal{H}_\sigma}\frac{\varphi'\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}} \\
&= \varphi\left(z(\boldsymbol{x}, g)\right)\|h\|^2_{\mathcal{H}_\sigma} + \|h\|^2_{\mathcal{H}_\sigma}\,\lambda(\boldsymbol{x}, g, h)z(\boldsymbol{x}, g)\varphi'\left(z(\boldsymbol{x}, g)\right), \qquad \text{(A.3)}
\end{aligned}$$

where for a fixed $g \in \mathcal{G}$, in the interest of brevity, we define $z(\boldsymbol{x}, g) = \|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}$ and

$$\lambda(\boldsymbol{x}, g, h) = \left\langle\frac{\Phi_\sigma(\boldsymbol{x}) - g}{\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}}, \frac{h}{\|h\|_{\mathcal{H}_\sigma}}\right\rangle^2_{\mathcal{H}_\sigma} \in [0, 1].$$

Observe that $z\varphi'(z) = \rho''(z) - \varphi(z)$, thus Eq. (A.3) becomes

$$\delta^2\ell(\boldsymbol{x}, g; h) = \|h\|^2_{\mathcal{H}_\sigma}\left((1-\lambda(\boldsymbol{x}, g, h))\,\varphi\left(z(\boldsymbol{x}, g)\right) + \lambda(\boldsymbol{x}, g, h)\rho''\left(z(\boldsymbol{x}, g)\right)\right).$$

From assumption $(\mathcal{A}4)$ we have that ρ'' and φ are nonincreasing, and

$$z(\boldsymbol{x}, g) = \|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma} \leq 2\,\|K_\sigma\|^{\frac{1}{2}}_\infty.$$

Thus, we have that

$$\delta^2\ell(\boldsymbol{x}, g; h) \geq c\,\|h\|^2_{\mathcal{H}_\sigma}, \qquad \text{(A.4)}$$

where

$$c = \min\left\{\varphi\left(2\,\|K_\sigma\|^{\frac{1}{2}}_\infty\right), \rho''\left(2\,\|K_\sigma\|^{\frac{1}{2}}_\infty\right)\right\}.$$

We also note that $\delta^2\ell(\boldsymbol{x}, g; h)$ is bounded above. To see this, note that from assumption $(\mathcal{A}4)$, ρ'' and φ are bounded and nonincreasing. Consequently, for $\lambda(\boldsymbol{x}, g, h) \in (0, 1)$ and

$$C = \max\left\{\rho''(0), \varphi(0)\right\} < \infty,$$

from Eq. (A.3) we have that

$$\delta^2\ell(\boldsymbol{x}, g; h) \leq C\,\|h\|^2_{\mathcal{H}_\sigma} < \infty.$$

The Gâteaux derivative of $\mathcal{J}(g)$ is, then, given by

$$\begin{aligned}\delta^2\mathcal{J}(g;h) &= \frac{d^2}{d\alpha^2}\mathcal{J}(g+\alpha h)\Big|_{\alpha=0} = \frac{d^2}{d\alpha^2}\int_{\mathbb{R}^d}\ell(\boldsymbol{x}, g+\alpha h)\; d\mathbb{P}(\boldsymbol{x})\Big|_{\alpha=0} \\ &= \int_{\mathbb{R}^d}\frac{d^2}{d\alpha^2}\ell(\boldsymbol{x}, g+\alpha h)\; d\mathbb{P}(\boldsymbol{x})\Big|_{\alpha=0} \\ &= \int_{\mathbb{R}^d}\delta^2\ell(\boldsymbol{x}, g; h)\; d\mathbb{P}(\boldsymbol{x}).\end{aligned}$$

The exchange of the derivative and integral in the second line follows from the dominated convergence theorem since $\left|\delta^2\ell(\boldsymbol{x}, g; h)\right|$ is bounded. This confirms the Gâteaux differentiability of $\mathcal{J}(g)$. From Eq. (A.4) we have

$$\delta^2\mathcal{J}(g;h) = \int_{\mathbb{R}^d}\delta^2\ell(\boldsymbol{x}, g; h)\; d\mathbb{P}(\boldsymbol{x}) \geq c\left\|h\right\|_{\mathcal{H}_\sigma}^2. \tag{A.5}$$

For $f_{\rho,\sigma} = \arg\inf_{g\in\mathcal{G}}\mathcal{J}(g)$ and $g \in \mathcal{G}$, we proceed to show the strong-convexity guarantee. Let $h = g - f_{\rho,\sigma}$. From the first-order Taylor approximation for $\mathcal{J}(g)$ we have,

$$\mathcal{J}(g) = \mathcal{J}(f_{\rho,\sigma}) + \delta\mathcal{J}(f_{\rho,\sigma}, h) + R_2(f_{\rho,\sigma}, h),$$

where the first Gâteaux derivative, $\delta\mathcal{J}(f_{\rho,\sigma}, h) = 0$ for all h since $f_{\rho,\sigma}$ is the unique minimizer of $\mathcal{J}(g)$ and the remainder term $R_2(f_{\rho,\sigma}, h)$ is given by

$$\begin{aligned}R_2(f_{\rho,\sigma}, h) &= \frac{1}{2}\int_0^1 (1-t)\delta^2\mathcal{J}(f_{\rho,\sigma}+th; h)\; dt \\ &\geq \frac{c}{2}\left\|h\right\|_{\mathcal{H}_\sigma}^2\int_0^1(1-t)dt = \frac{c}{4}\left\|h\right\|_{\mathcal{H}_\sigma}^2,\end{aligned}$$

where the inequality follows from Eq. (A.5). As a result, for any $g \in \mathcal{G}$ and $\mu = \frac{c}{2}$ we have that

$$\mathcal{J}(g) - \mathcal{J}(f_{\rho,\sigma}) \geq \frac{\mu}{2}\left\|g - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma}^2,$$

yielding the desired result. ■

We now turn to examining the behaviour of the risk functional $\mathcal{J}(g)$ w.r.t. the underlying probability measure $\mathbb{P}$. For $0 \leq \epsilon \leq 1$ and $\boldsymbol{x} \in \mathbb{R}^d$, let $\mathbb{P}_{\boldsymbol{x}}^\epsilon = (1-\epsilon)\mathbb{P} + \epsilon\delta_{\boldsymbol{x}}$ be a perturbation curve, as defined in Theorem 4.1. The risk functional associated with $\mathbb{P}_{\boldsymbol{x}}^\epsilon$ is given by

$$\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) = \mathbb{P}_{\boldsymbol{x}}^\epsilon\ell_g = (1-\epsilon)\mathcal{J}(g) + \epsilon\rho\left(\left\|\Phi_\sigma(\boldsymbol{x}) - g\right\|_{\mathcal{H}_\sigma}\right),$$

and $f_{\rho,\sigma}^{\epsilon,\boldsymbol{x}} = \inf_{g\in\mathcal{G}}\mathcal{J}_{\epsilon,\boldsymbol{x}}(g)$ is the minimizer. The convergence of $f_{\rho,\sigma}^{\epsilon,\boldsymbol{x}}$ to $f_{\rho,\sigma}$ can be studied by examining the convergence of $\mathcal{J}_{\epsilon,\boldsymbol{x}}$ to $\mathcal{J}$. Specifically, under conditions on $\mathcal{J}$ and $\mathcal{J}_{\epsilon,\boldsymbol{x}}$, it can be shown that $\left\|f_{\rho,\sigma}^{\epsilon,\boldsymbol{x}} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} \to 0$ as $\epsilon \to 0$. The machinery we use here uses the notion of Γ–convergence, which is defined as follows.

Definition A.1 (Γ convergence). *Given a functional $F : \mathcal{X} \to \mathbb{R}\cup\{\pm\infty\}$ and a sequence of functionals $\{F_n\}_{n\in\mathbb{N}}$, $F_n \xrightarrow{\Gamma} F$ as $n \to \infty$ when*

(i) $F(\boldsymbol{x}) \leq \liminf_{n\to\infty} F_n(\boldsymbol{x}_n)$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and every* $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ *such that* $d(\boldsymbol{x}_n, \boldsymbol{x}) \to 0$*;*

(ii) For every $\boldsymbol{x} \in \mathcal{X}$*, there exists* $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$*,* $d(\boldsymbol{x}_n, \boldsymbol{x}) \to 0$ *such that* $F(\boldsymbol{x}) \geq \limsup_{n\to\infty} F_n(\boldsymbol{x}_n)$*.*

The following result shows that the sequence of functionals $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ Γ–converges to $\mathcal{J}$.

Proposition A.2 (Γ–convergence of $\mathcal{J}_{\epsilon,\boldsymbol{x}}$ to $\mathcal{J}$)**.** *Under assumptions* $(\mathcal{A}1)$*–*$(\mathcal{A}3)$*,*

$$\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) \xrightarrow{\Gamma} \mathcal{J}(g) \quad \text{as} \quad \epsilon \to 0.$$

Proof. Let $g \in \mathcal{G}$ and $\{g_\epsilon\}_{\epsilon>0}$ be a sequence in $\mathcal{G}$ such that $\|g_\epsilon - g\|_{\mathcal{H}_\sigma} \to 0$ as $\epsilon \to 0$. In order to verify Γ–convergence we first show that the following holds

$$\lim_{\epsilon\to 0}\Big|\mathcal{J}_{\epsilon,\boldsymbol{x}}(g_\epsilon) - \mathcal{J}(g)\Big| = 0.$$

For $\epsilon > 0$, using the triangle inequality we have that

$$\begin{aligned}\Big|\mathcal{J}_{\epsilon,\boldsymbol{x}}(g_\epsilon) - \mathcal{J}(g)\Big| &\leq \Big|\mathcal{J}(g_\epsilon) - J(g)\Big| + \Big|\mathcal{J}_{\epsilon,\boldsymbol{x}}(g_\epsilon) - \mathcal{J}(g_\epsilon)\Big| \\ &\overset{(i)}{\leq} M\,\|g_\epsilon - g\|_{\mathcal{H}_\sigma} + \Big|\mathcal{J}_{\epsilon,\boldsymbol{x}}(g_\epsilon) - \mathcal{J}(g_\epsilon)\Big| \\ &\overset{(ii)}{\leq} M\,\|g_\epsilon - g\|_{\mathcal{H}_\sigma} + \epsilon \cdot \Big|\mathcal{J}(g) - \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)\Big|,\end{aligned}$$

where (i) uses the fact that $\mathcal{J}(g)$ is M–Lipschitz from Proposition A.1, and (ii) uses the fact that

$$\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) = (1-\epsilon)\mathcal{J}(g) + \epsilon\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right).$$

Since $\|g_\epsilon - g\|_{\mathcal{H}_\sigma} \to 0$ as $\epsilon \to 0$ we have

$$\lim_{\epsilon\to 0}\Big|\mathcal{J}_{\epsilon,\boldsymbol{x}}(g_\epsilon) - \mathcal{J}(g)\Big| \leq M \lim_{\epsilon\to 0}\|g_\epsilon - g\|_{\mathcal{H}_\sigma} + \lim_{\epsilon\to 0}\epsilon \cdot \Big|\mathcal{J}(g) - \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)\Big| = 0.$$

Since $\mathcal{J}_{\epsilon,\boldsymbol{x}}$ and $\mathcal{J}$ are continuous, using (Dal Maso, 2012, Remark 4.8) it follows that $\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) \xrightarrow{\Gamma} \mathcal{J}(g)$. ■

Now, we examine the coercivity of the sequence $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$.

Definition A.2 (Equi-coercivity)**.** *A sequence of functionals* $\{F_n\}_{n\in\mathbb{N}} : \mathcal{X} \to \mathbb{R}\cup\{\pm\infty\}$ *is said to be equi-coercive if for every* $t \in \mathbb{R}$*, there exists a compact set* $K_t \subseteq \mathcal{X}$ *such that* $\{\boldsymbol{x} \in \mathcal{X} : F_n \leq t\} \subseteq K_t$ *for every* $n \in \mathbb{N}$*.*

The following result shows that the sequence $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ is equi-coercive.

Proposition A.3 (Equi-coercivity of $\mathcal{J}_{\epsilon,\boldsymbol{x}}$)**.** *Under assumptions* $(\mathcal{A}1)$*–*$(\mathcal{A}3)$*, the sequence of functionals* $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ *is equi-coercive.*

Proof. For $0 < \epsilon < 1$, $\boldsymbol{x} \in \mathbb{R}^d$ and $g \in \mathcal{G}$, we have that

$$\mathcal{J}_{\epsilon,\boldsymbol{x}}(g) = (1-\epsilon)\mathcal{J}(g) + \epsilon\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right).$$

From (Dal Maso, 2012, Proposition 7.7) in order to show that the sequence of functionals $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ is equi-coercive, it suffices to show that there exists a lower semicontinuous, coercive functional $F : \mathcal{H}_\sigma \to \mathbb{R} \cup \{\pm\infty\}$ such that $F \leq \mathcal{J}_{\epsilon,\boldsymbol{x}}$ for every $\epsilon \geq 0$. To this end consider the functional

$$F(g) = \min\left\{\mathcal{J}(g), \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)\right\}.$$

As $\mathcal{J}_{\epsilon,\boldsymbol{x}}$ is a convex combination of $\mathcal{J}(\cdot)$ and $\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - \cdot\|_{\mathcal{H}_\sigma}\right)$, it implies that $F \leq \mathcal{J}_{\epsilon,\boldsymbol{x}}$ for every $\epsilon \geq 0$. Additionally, because $\mathcal{J}(\cdot)$ and $\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - \cdot\|_{\mathcal{H}_\sigma}\right)$ are both continuous, it follows that F is also continuous, and, therefore, lower semicontinuous.

We now verify that F is coercive. Since ρ is strictly increasing we have that

$$\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) \to \infty \qquad \text{as} \quad \|g\|_{\mathcal{H}_\sigma} \to \infty,$$

verifying that $\rho\left(\|\Phi_\sigma(\boldsymbol{x}) - \cdot\|_{\mathcal{H}_\sigma}\right)$ is coercive. Next, from the reverse triangle inequality we have that

$$\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma} \geq \left|\|\Phi_\sigma(\boldsymbol{x})\|_{\mathcal{H}_\sigma} - \|g\|_{\mathcal{H}_\sigma}\right| = \left|\sqrt{K_\sigma(\boldsymbol{x},\boldsymbol{x})} - \|g\|_{\mathcal{H}_\sigma}\right|.$$

Observe that $K_\sigma(\boldsymbol{x},\boldsymbol{x}) = \|K_\sigma\|_\infty$, and because ρ is strictly increasing we have

$$\rho\left(\left|\|K_\sigma\|_\infty^{\frac{1}{2}} - \|g\|_{\mathcal{H}_\sigma}\right|\right) \leq \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right).$$

Taking expectations on both sides w.r.t. $\mathbb{P}$,

$$\rho\left(\left|\|K_\sigma\|_\infty^{\frac{1}{2}} - \|g\|_{\mathcal{H}_\sigma}\right|\right) \leq \int_{\mathbb{R}^d} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) = \mathcal{J}(g).$$

Since

$$\rho\left(\left|\|K_\sigma\|_\infty^{\frac{1}{2}} - \|g\|_{\mathcal{H}_\sigma}\right|\right) \to \infty \qquad \text{as} \quad \|g\|_{\mathcal{H}_\sigma} \to \infty,$$

it implies that $\mathcal{J}(g)$ is coercive as well. It follows from this that F is coercive, and the sequence of functionals $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ is equi-coercive. ■

Propositions A.2 and A.3 together imply, from the fundamental theorem of Γ-convergence (Braides, 2002), that the sequence of minimizers associated with $\{\mathcal{J}_{\epsilon,\boldsymbol{x}}\}$ converge to the minimizer of $\mathcal{J}$, i.e.,

$$\left\|f_{\rho,\sigma}^{\epsilon,\boldsymbol{x}} - f_{\rho,\sigma}\right\|_{\mathcal{H}_\sigma} \to 0 \qquad \text{as} \quad \epsilon \to 0.$$

A.2 Some Additional Results

Next, we note an important property of the hypothesis class, $\mathcal{G} = \mathcal{H}_\sigma \cap \mathcal{D}_\sigma$. The elements of $\mathcal{G}$ can be shown to have their $\|\cdot\|_\infty$–norm related their $\|\cdot\|_{\mathcal{H}_\sigma}$–norm.

Lemma A.1 (Vandermeulen and Scott, 2013, Lemma 6 and Sriperumbudur, 2016, Proposition 5.1)**.** *For every $g \in \mathcal{H}_\sigma \cap \mathcal{D}_\sigma$,*

$$\|g\|_{\mathcal{H}_\sigma}^2 \leq \|g\|_\infty \leq \|K_\sigma\|_\infty^{\frac{1}{2}} \|g\|_{\mathcal{H}_\sigma}.$$

The following result, which is essentially the population analogue of Vandermeulen and Scott (2013, Lemma 7), guarantees that for small enough $\sigma > 0$, there exists $0 < \delta < 1$ such that $f_{\rho,\sigma}$ is contained in the RKHS ball $B_{\mathcal{H}_\sigma}(\mathbf{0}, \delta\nu_\sigma)$, where for brevity we denote $\nu_\sigma = \|K_\sigma\|_\infty^{1/2}$. We provide the proof for completeness, however, the proof uses exactly the same ideas from Vandermeulen and Scott (2013). For notational convenience, we also define $\psi_\sigma(\|\boldsymbol{x} - \boldsymbol{y}\|_2) = K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \sigma^{-d}\psi\left(\|\boldsymbol{x} - \boldsymbol{y}\|_2 / \sigma\right)$.

Lemma A.2. *Let $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$ and $f_{\rho,\sigma}$ be the robust KDE for $\sigma > 0$. For sufficiently small $\sigma > 0$, there exists $0 < \delta < 1$ such that $f_{\rho,\sigma} \in B(\mathbf{0}, \delta\nu_\sigma)$.*

Proof. For $\mathbb{P} \in \mathcal{M}(\mathbb{R}^d)$, and $\mathcal{G} = \mathcal{H}_\sigma \cap \mathcal{D}_\sigma$, consider the map $T_\sigma : \mathcal{G} \to \mathcal{G}$ given by

$$T_\sigma(g) = \int_{\mathbb{R}^d} \frac{\varphi\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right)}{\int_{\mathbb{R}^d} \varphi\left(\|\Phi_\sigma(\boldsymbol{y}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{y})} K_\sigma(\cdot, \boldsymbol{x})\; d\mathbb{P}(\boldsymbol{x}) = \int_{\mathbb{R}^d} K_\sigma(\cdot, \boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}),$$

for each $g \in \mathcal{G}$. Observe that $w_\sigma \in L_1(\mathbb{P})$ is a non-negative function such that

$$\int_{\mathbb{R}^d} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) = 1. \tag{A.6}$$

Let $S_\sigma = \mathrm{Im}(T_\sigma) \subset \mathcal{G}$. It follows from Vandermeulen and Scott (2013, Page 11) that the robust KDE, $f_{\rho,\sigma} = \arg\inf_{g\in\mathcal{G}} \mathcal{J}(g)$, is the fixed point of the map T_σ and therefore $f_{\rho,\sigma} \in S_\sigma$. For a small $\epsilon > 0$, from Vandermeulen and Scott (2013, Lemma 12; Corollary 13) there exist $r, s > 0$ such that $\mathbb{P}(B(\boldsymbol{x}, r)) \leq \epsilon$ and $\mathbb{P}(B(\boldsymbol{x}, r+s) \setminus B(\boldsymbol{x}, r)) \leq \epsilon$ for all $\boldsymbol{x} \in \mathbb{R}^d$. This implies that $\mathbb{P}(B(\boldsymbol{x}, r+s)^c) > 1 - 2\epsilon$. We point out that the constant ϵ chosen here is related to $\sqrt{9/10}$ used by Vandermeulen and Scott (2013) as $\sqrt{1-\epsilon} = \sqrt{9/10}$, which, as remarked by the authors, was chosen simply for convenience. Define the sets $B_\sigma = B_{\mathcal{H}_\sigma}(\mathbf{0}, \nu_\sigma\sqrt{1-\epsilon})$, and let

$$R_\sigma \doteq S_\sigma \cap B_\sigma^c.$$

In what follows we will show that $f_{\rho,\sigma}$ does not lie in R_σ. To this end, let $g = \arg\inf_{h\in R_\sigma} \mathcal{J}(h)$. It suffices to show that $\mathcal{J}(g) > \mathcal{J}(\mathbf{0}) > \mathcal{J}(f_{\rho,\sigma})$. Since $g \in R_\sigma$, it must follow that

$$(1-\epsilon)\nu_\sigma^2 < \|g\|_{\mathcal{H}_\sigma}^2 \leq \|g\|_\infty = g(\boldsymbol{z}), \tag{A.7}$$

for some $\boldsymbol{z} \in \mathbb{R}^d$, where the second inequality follows from Lemma A.1. Since $g \in S_\sigma$, there exists a non-negative function w_σ satisfying Eq. (A.6), such that $g = \int_{\mathbb{R}^d} w_\sigma(\boldsymbol{x}) K_\sigma(\cdot, \boldsymbol{x}) d\mathbb{P}(\boldsymbol{x})$. Therefore,

$$\begin{aligned}
(1-\epsilon)\nu_\sigma^2 \leq g(\boldsymbol{z}) &= \int_{\mathbb{R}^d} K_\sigma(\boldsymbol{z}, \boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \\
&= \int_{B(\boldsymbol{z},r)} K_\sigma(\boldsymbol{z}, \boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) + \int_{B(\boldsymbol{z},r)^c} K_\sigma(\boldsymbol{z}, \boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \\
&\overset{(i)}{\leq} \nu_\sigma^2 \int_{B(\boldsymbol{z},r)} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) + \psi_\sigma(r) \underbrace{\int_{B(\boldsymbol{z},r)^c} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x})}_{\leq 1} \\
&\overset{(ii)}{\leq} \nu_\sigma^2 \int_{B(\boldsymbol{z},r)} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) + \psi_\sigma(r),
\end{aligned} \tag{A.8}$$

where (i) follows from the fact that $\sup_{B(\boldsymbol{z},r)^c} K_\sigma(\boldsymbol{z},\boldsymbol{x}) = \psi_\sigma(r)$ and (ii) follows from Eq. (A.6). From Vandermeulen and Scott (2013, Lemma 7), there exists σ small enough such that $\psi_\sigma(r) < \frac{\epsilon}{2}\nu_\sigma^2$. Plugging this back in Eq. (A.8) we get

$$\int_{B(\boldsymbol{z},r)} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \geq \left(1 - \frac{3\epsilon}{2}\right). \tag{A.9}$$

Additionally,

$$\begin{aligned}
\sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} g(\boldsymbol{y}) &= \sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} \Big(\int_{B(\boldsymbol{z},r)} K_\sigma(\boldsymbol{y},\boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) + \int_{B(\boldsymbol{z},r)^c} K_\sigma(\boldsymbol{y},\boldsymbol{x}) w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \Big) \\
&\leq \sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} \sup_{\boldsymbol{x}\in B(\boldsymbol{z},r)} K_\sigma(\boldsymbol{y},\boldsymbol{x}) \int_{B(\boldsymbol{z},r)} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \\
&\quad + \sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} \sup_{\boldsymbol{x}\in B(\boldsymbol{z},r)} K_\sigma(\boldsymbol{y},\boldsymbol{x}) \int_{B(\boldsymbol{z},r)^c} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) \\
\sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} g(\boldsymbol{y}) &\leq \psi_\sigma(s) + \nu_\sigma^2 \int_{B(\boldsymbol{z},r)^c} w_\sigma(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}).
\end{aligned}$$

For a choice of $\tau > 0$, there is σ small enough satisfying $\psi_\sigma(s) \leq \tau$ such that from Eq. (A.9)

$$\sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} g(\boldsymbol{y}) \leq \tau + \frac{3\epsilon}{2}\nu_\sigma^2. \tag{A.10}$$

Then we have that

$$\begin{aligned}
\mathcal{J}(g) &= \int_{\mathbb{R}^d} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) \\
&= \int_{B(\boldsymbol{z},r+s)} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) + \int_{B(\boldsymbol{z},r+s)^c} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) \\
&\geq \int_{B(\boldsymbol{z},r+s)^c} \rho\left(\|\Phi_\sigma(\boldsymbol{x}) - g\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) \\
&= \int_{B(\boldsymbol{z},r+s)^c} \rho\left(\sqrt{\nu_\sigma^2 + \|g\|_{\mathcal{H}_\sigma}^2 - 2\langle g, \Phi_\sigma(\boldsymbol{x})\rangle_{\mathcal{H}_\sigma}}\right) d\mathbb{P}(\boldsymbol{x}) \\
&\geq \int_{B(\boldsymbol{z},r+s)^c} \rho\left(\sqrt{\nu_\sigma^2 + \|g\|_{\mathcal{H}_\sigma}^2 - 2\sup_{\boldsymbol{y}\in B(\boldsymbol{z},r+s)^c} g(\boldsymbol{y})}\right) d\mathbb{P}(\boldsymbol{x}).
\end{aligned}$$

Plugging in Equations (A.10) and (A.7) we get

$$\mathcal{J}(g) \geq (1-2\epsilon)\rho\left(\sqrt{(2-4\epsilon)\nu_\sigma^2 - 2\tau}\right).$$

Since ρ is assumed to be strictly convex, this implies that ρ' is strictly increasing. Additionally, from $(\mathcal{A}2)$ we have that ρ' is bounded. This implies that, for any $0 < \alpha < \|\rho'\|_\infty$, there is $\beta > 0$ such that $\rho'(z) > \|\rho'\|_\infty - \alpha$ for all $z > \beta$. Using Vandermeulen and Scott (2013, Eq. 11), we have

$$\begin{aligned}
\rho\left(\sqrt{(2-4\epsilon)\nu_\sigma^2 - 2\tau}\right) &= \int_0^{(2-4\epsilon)\nu_\sigma^2 - 2\tau} \rho'(z)dz \\
&\geq \int_\beta^{(2-4\epsilon)\nu_\sigma^2 - 2\tau} \rho'(z)dz \\
&\geq \int_\beta^{(2-4\epsilon)\nu_\sigma^2 - 2\tau} \left(\|\rho'\|_\infty - \alpha\right) dz \\
&\geq \left(\|\rho'\|_\infty - \alpha\right)\left(\sqrt{(2-4\epsilon)\nu_\sigma^2 - 2\tau} - \beta\right).
\end{aligned}$$

Without loss of generality, we can assume $\|\rho'\|_\infty = 1$. Choosing α, τ and σ small enough we obtain

$$\mathcal{J}(g) \geq \nu_\sigma.$$

Now we note that

$$\begin{aligned}
\mathcal{J}(\mathbf{0}) &= \int_{\mathbb{R}^d} \rho\left(\|\Phi_\sigma(\boldsymbol{x})\|_{\mathcal{H}_\sigma}\right) d\mathbb{P}(\boldsymbol{x}) \\
&= \rho\left(\nu_\sigma\right) \\
&= \rho(0) + \int_0^{\nu_\sigma} \rho'(z)dz \\
&\leq \rho(0) + \|\rho'\|_\infty \int_0^{\nu_\sigma} dz = \nu_\sigma.
\end{aligned}$$

Thus, we obtain that $\mathcal{J}(g) > \mathcal{J}(\mathbf{0})$. We have $g = \arg\inf_{h \in R_\sigma} \mathcal{J}(h)$ and $f_{\rho,\sigma} = \arg\inf_{h \in \mathcal{G}} \mathcal{J}(h)$, and, additionally we know that $f_{\rho,\sigma} \neq \mathbf{0}$. It follows that since $\mathcal{J}(f_{\rho,\sigma}) \leq \mathcal{J}(\mathbf{0}) < \mathcal{J}(g)$, then $f_{\rho,\sigma} \notin R_\sigma$ as $\sigma \to 0$. Taking $\delta = \sqrt{1-\epsilon}$, we get the desired result. ■

B Background on Persistent Homology

Given a set of a points $\mathbb{X}_n = \{\boldsymbol{x}_1 \dots \boldsymbol{x}_n\}$ in a metric space $(\mathcal{X}, d)$ their topology is encoded in a geometric object called a simplicial complex $\mathcal{K} \subseteq 2^{\mathbb{X}_n}$.

Definition B.1. (Hatcher, 2002). *A simplicial complex $\mathcal{K}$ is a collection of simplices $\langle\sigma\rangle$ i.e. points, lines, triangles, tetrahedra and its higher dimensional analogues, such that*

1. *$\forall \tau \preccurlyeq \sigma$, $\sigma \in \mathcal{K}$ we have $\tau \in \mathcal{K}$;*
2. *$\forall \sigma, \tau \in \mathcal{K}$, we have that $\sigma \cap \tau \preccurlyeq \sigma, \tau$ or $\sigma \cap \tau = \phi$.*

For a given spatial resolution $r > 0$, the simplicial complex for $\mathbb{X}_n$, given by $\mathcal{K}(\mathbb{X}_n, r)$, can be constructed in multiple ways. For example, the Vietoris-Rips complex is the simplicial complex

$$\mathcal{K}_r = \{\sigma \subseteq \mathbb{X}_n : \bigcap_{\boldsymbol{x} \in \sigma} B(\boldsymbol{x}, r) \neq \varnothing\},$$

and the Čech complex is given by

$$\mathcal{K}_r = \{\sigma \subseteq \mathbb{X}_n : \max_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \sigma} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq r\}.$$

More generally, if $\mathcal{K}$ is a simplicial complex constructed using an approximation of the space $\mathcal{X}$ (e.g., triangulation, surface mesh, grid, etc.), and $\phi : \mathcal{X} \to \mathbb{R}$ a filter function, ϕ induces the map $\phi : \mathcal{K} \to \mathbb{R}$. Then, $\mathcal{K}_r = \phi^{-1}([0, r])$ encodes the information in the sublevel set of ϕ at resolution r. Similarly, $\mathcal{K}^r$ encodes the information in the superlevel sets at resolution r.

For $0 \leq k \leq d$, the k^{th}-*homology* (Hatcher, 2002) of a simplicial complex $\mathcal{K}$, given by $H_k(\mathcal{K})$ is an algebraic object encoding its topology as a vector-space (over a fixed field). Using the Nerve lemma, $H_k(\mathcal{K}(\mathbb{X}_n, r))$ is isomorphic to the homology of its union of r-balls, $H_k(\bigcup_{i=1}^n B_r(\boldsymbol{x}_i))$. The ordered sequence $\{\mathcal{K}(\mathbb{X}_n, r)\}_{r>0}$ forms a *filtration*, encoding the evolution of topological features over a spectrum of resolutions. For $0 < r < s$, the simplicial complex $\mathcal{K}(\mathbb{X}_n, r)$ is a *sub-simplicial complex* of $\mathcal{K}(\mathbb{X}_n, s)$. Their homology groups are associated with the inclusion maps

$$\iota_r^s : H_k(\mathcal{K}(\mathbb{X}_n, r)) \hookrightarrow H_k(\mathcal{K}(\mathbb{X}_n, s)),$$

which in turn carry information on the number of non-trivial k-cycles. As the resolution r varies, the evolution of the topology is captured in the filtration. Roughly speaking, new cycles (e.g., connected components, loops, voids and higher order analogues) can appear or existing cycles can merge. Formally, a new k-cycle σ_k with homology class $[\alpha_k]$ is *born* at $b \in \mathbb{R}$ if $[\alpha_k] \notin \mathrm{Im}(\iota_{b-\epsilon,b}^k)$ for all $\epsilon > 0$ and $[\alpha_k] \in \mathrm{Im}(\iota_{b,b+\delta}^k)$ for some $\delta > 0$. The same k-cycle born at b dies at $d > b$ if $\iota_{b,d-\delta}^k([\alpha_k]) \notin \mathrm{Im}(\iota_{b-\epsilon,d-\delta}^k)$ and $\iota_{b,d}^k([\alpha_k]) \in \mathrm{Im}(\iota_{b-\epsilon,d}^k)$ for all $\epsilon > 0$ and $0 < \delta < d - b$. Persistent homology, $PH_*(\phi)$, is an algebraic module which tracks the persistence pairs (b, d) of births b and deaths d across the entire filtration. By collecting all persistence pairs (b, d), the persistent homology is represented as a persistence diagram

$$\mathsf{Dgm}(\mathcal{K}(\mathbb{X}_n)) \doteq \left\{(b, d) \in \mathbb{R}^2 : 0 \leq b < d \leq \infty\right\}.$$

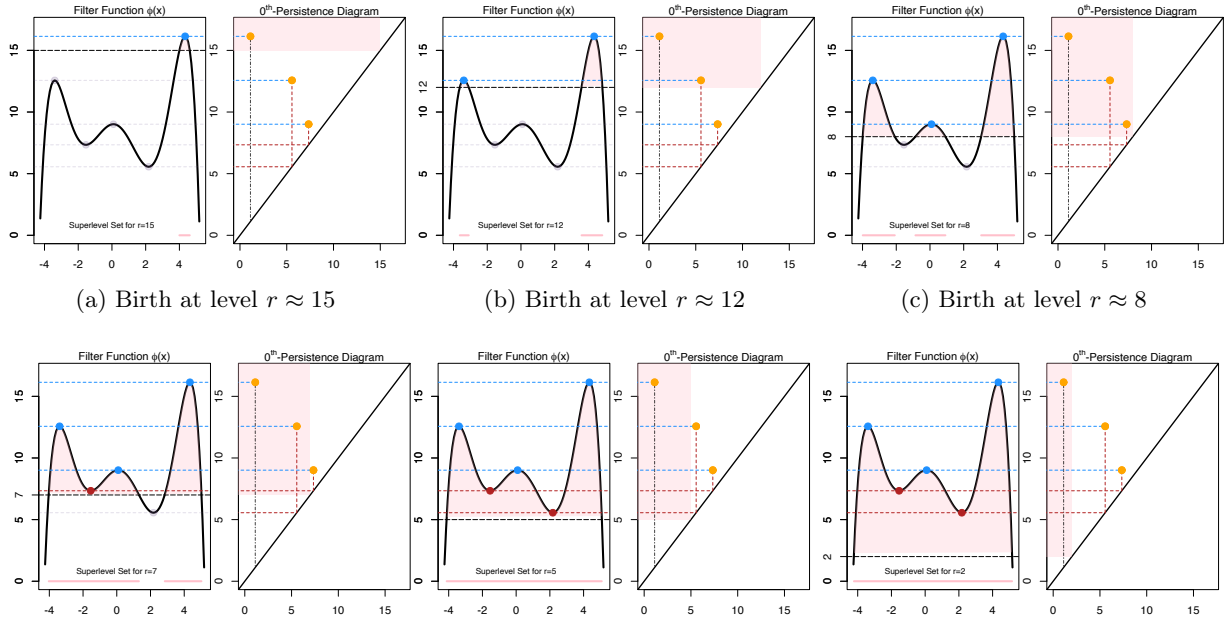
The persistence diagram is a multiset of points on the space $\Omega = \{(x, y) : 0 \leq x < y \leq \infty\}$, such that each point (x, y) in the persistence diagram corresponds to a distinct topological feature which existed in $\mathcal{K}(\mathbb{X}_n, r)$ for $x \leq r < y$. Given a persistence diagram $\mathbf{D}$ and $1 \leq p \leq \infty$ the *degree-p total persistence* of $\mathbf{D}$ is given by

$$\mathrm{pers}_p(\mathbf{D}) = \left(\sum_{(b,d) \in \mathbf{D}} |d - b|^p\right)^{\frac{1}{p}}.$$

The space of persistence diagrams, given by $\mathcal{D}_p = \left\{\mathbf{D} : \mathrm{pers}_p(\mathbf{D}) < \infty\right\}$, is endowed with the family of p-Wasserstein metrics W_p. Given two persistence diagrams $\mathbf{D}_1, \mathbf{D}_2 \in \mathcal{D}_p$, the p-Wasserstein distance is given by

$$W_p(\mathbf{D}_1, \mathbf{D}_2) \doteq \left(\inf_{\gamma \in \Gamma} \sum_{\boldsymbol{z} \in \mathbf{D}_1 \cup \Delta} \|\boldsymbol{z} - \gamma(\boldsymbol{z})\|_\infty^p\right)^{\frac{1}{p}}, \tag{B.1}$$

where $\Gamma = \{\gamma : \mathbf{D}_1 \cup \Delta \to \mathbf{D}_2 \cup \Delta\}$ is the set of all bijections from $\mathbf{D}_1$ to $\mathbf{D}_2$ including the diagonal $\Delta = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x = y \leq \infty\right\}$ with infinite multiplicity.



(a) Birth at level $r \approx 15$ (b) Birth at level $r \approx 12$ (c) Birth at level $r \approx 8$

(d) Death at level $r \approx 7$ (e) Death at level $r \approx 5$ (f) Component persists until $r = 0$

Figure 9: An example for the superlevel filtration of $\phi : \mathbb{R} \to \mathbb{R}$. (a) As the superlevel set enters $r \approx 15$, the first connected component is born, corresponding to the blue dot on the highest peak of ϕ. The superlevel set for $r = 15$ is depicted in pink below. This is recorded as a birth in the corresponding orange dot enclosed in the pink shaded region of the persistence diagram. (b) As the r enters $r \approx 12$, another connected component is born. This is recorded as the second orange dot in the shaded region of the persistence diagram. (c) Again, at $r \approx 8$, a third connected component is born at the lowest peak of ϕ. The three connected components in the superlevel set are shaded in pink below the function. The persistence diagram has three orange dots corresponding to these three connected components. (d) As r enters the first valley of ϕ, depicted by the red dot, two connected components merge (i.e., one of the existing connected components die). By convention, the most recent persistent feature is merged into the older one, i.e., the connected component from (c) merges into the one from (b), and thus, it dies at this resolution. In the persistence diagram, this is noted by the fact that the orange dot born in (c) dies at resolution $r \approx 7$. At this stage, there are only two orange dots in the pink shaded region of the persistence diagram, corresponding to the two pink connected components in the superlevel set of ϕ. (e) When r enters the second valley of ϕ, the connected component from (b) merges into the connected component from (a), and form a single connected component. The orange dot in the persistence diagram records the death of this feature. (f) The single connected component persists from then on, and eventually dies at $r = 0$.