

Geometric Inference on Kernel Density Estimates

Jeff M. Phillips*
jeffp@cs.utah.edu
University of Utah

Bei Wang†
beiwang@sci.utah.edu
University of Utah

Yan Zheng
yanzheng@cs.utah.edu
University of Utah

Abstract

We show that geometric inference of a point cloud can be calculated by examining its kernel density estimate with a Gaussian kernel. This allows one to consider kernel density estimates, which are robust to spatial noise, subsampling, and approximate computation in comparison to raw point sets. This is achieved by examining the sublevel sets of the *kernel distance*, which isomorphically map to superlevel sets of the kernel density estimate. We prove new properties about the kernel distance, demonstrating stability results and allowing it to inherit reconstruction results from recent advances in distance-based topological reconstruction. Moreover, we provide an algorithm to estimate its topology using weighted Vietoris-Rips complexes.

arXiv:1307.7760v3 [cs.CG] 26 Mar 2015

*Thanks to supported by NSF CCF-1350888, IIS-1251019, and ACI-1443046.

†Thanks to supported by INL 00115847 via DOE DE-AC07ID14517, DOE NETL DEEE0004449, DOE DEFC0206ER25781, DOE de-sc0007446, and NSF 0904631.

1 Introduction

Geometry and topology have become essential tools in modern data analysis: geometry to handle spatial noise and topology to identify the core structure. Topological data analysis (TDA) has found applications spanning protein structure analysis [32, 52] to heart modeling [41] to leaf science [60], and is the central tool of identifying quantities like connectedness, cyclic structure, and intersections at various scales. Yet it can suffer from spatial noise in data, particularly outliers.

When analyzing point cloud data, classically these approaches consider α -shapes [31], where each point is replaced with a ball of radius α , and the union of these balls is analyzed. More recently a distance function interpretation [11] has become more prevalent where the union of α -radius balls can be replaced by the sublevel set (at value α) of the Hausdorff distance to the point set. Moreover, the theory can be extended to other distance functions to the point sets, including the *distance-to-a-measure* [15] which is more robust to noise.

This has more recently led to statistical analysis of TDA. These results show not only robustness in the function reconstruction, but also in the topology it implies about the underlying dataset. This work often operates on persistence diagrams which summarize the persistence (difference in function values between appearance and disappearance) of all homological features in single diagram. A variety of work has developed metrics on these diagrams and probability distributions over them [55, 67], and robustness and confidence intervals on their landscapes [7, 39, 18] (summarizing again the most dominant persistent features [19]). Much of this work is independent of the function and data from which the diagram is generated, but it is now more clear than ever, it is most appropriate when the underlying function is robust to noise, e.g., the distance-to-a-measure [15].

A very recent addition to this progression is the new TDA package for R [38]; it includes built in functions to analyze point sets using Hausdorff distance, distance-to-a-measure, k -nearest neighbor density estimators, kernel density estimates, and kernel distance. The example in Figure 1 used this package to generate persistence diagrams. While, the stability of the Hausdorff distance is classic [11, 31], and the distance-to-a-measure [15] and k -nearest neighbor distances have been shown robust to various degrees [5], this paper is the first to analyze the stability of kernel density estimates and the kernel distance in the context of geometric inference. Some recent manuscripts show related results. Bobrowski *et al.* [6] consider kernels with finite support, and describe approximate confidence intervals on the superlevel sets, which recover approximate persistence diagrams. Chazal *et al.* [17] explore the robustness of the kernel distance in bootstrapping-based analysis.

In particular, we show that the kernel distance and kernel density estimates, using the Gaussian kernel, inherit some reconstruction properties of distance-to-a-measure, that these functions can also be approximately reconstructed using weighted (Vietoris-)Rips complexes [8], and that under certain regimes can infer homotopy of compact sets. Moreover, we show further robustness advantages of the kernel distance and kernel density estimates, including that they possess small coresets [57, 71] for persistence diagrams and inference.

1.1 Kernels, Kernel Density Estimates, and Kernel Distance

A *kernel* is a non-negative similarity measure $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$; more similar points have higher value. For any fixed $p \in \mathbb{R}^d$, a kernel $K(p, \cdot)$ can be normalized to be a probability distribution; that is $\int_{x \in \mathbb{R}^d} K(p, x) dx = 1$. For the purposes of this article we focus on the Gaussian kernel defined as $K(p, x) = \sigma^2 \exp(-\|p - x\|^2 / 2\sigma^2)$.¹

¹ $K(p, x)$ is normalized so that $K(x, x) = 1$ for $\sigma = 1$. The choice of coefficient σ^2 is not the standard normalization, but it is perfectly valid as it scales everything by a constant. It has the property that $\sigma^2 - K(p, x) \approx \|p - x\|^2 / 2$ for $\|p - x\|$ small.

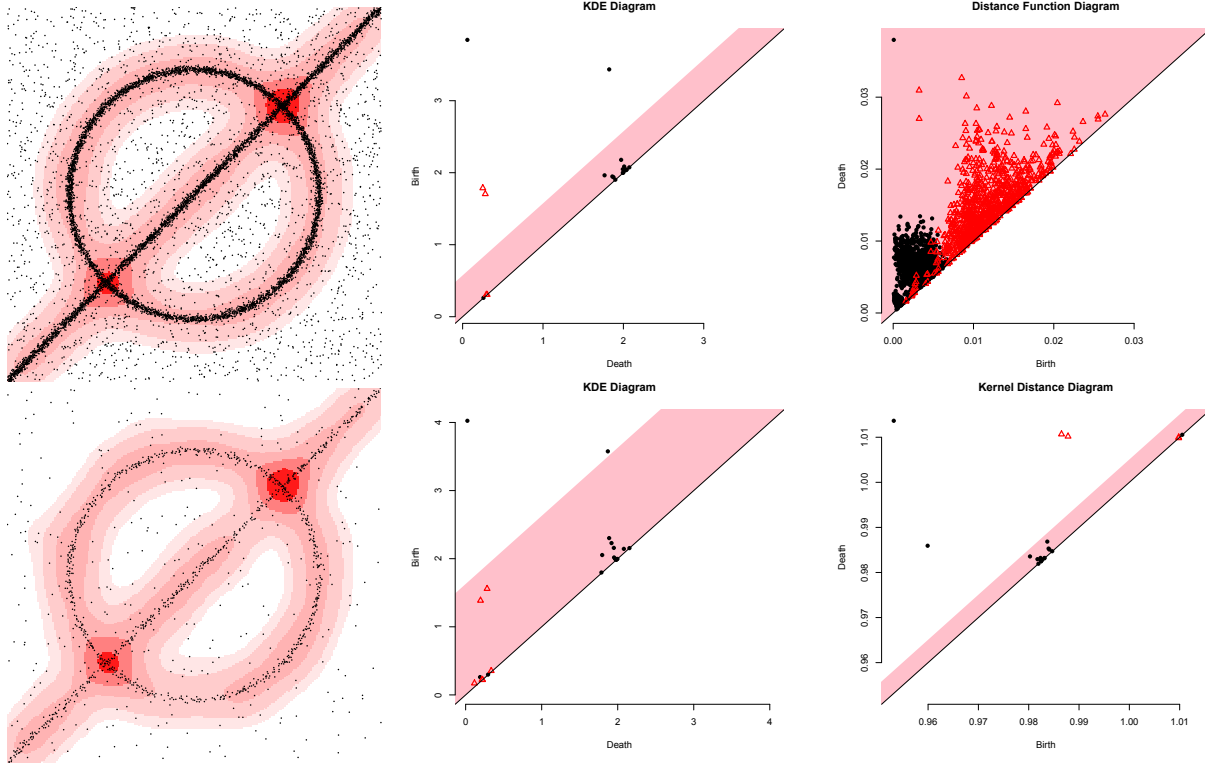


Figure 1: Example with 10,000 points in $[0, 1]^2$ generated on a circle or line with $N(0, 0.005)$ noise; 25% of points are uniform background noise. The generating function is reconstructed with KDE with $\sigma = 0.05$ (upper left), and its persistence diagram based on the superlevel set filtration is shown (upper middle). A coreset [71] of the same dataset with only 1,384 points (lower left) and persistence diagram (lower middle) are shown, again using KDE. This associated confidence interval contains the dimension 1 homology features (red triangles) suggesting they are noise; this is because it models data as iid – but the coreset data is not iid, it subsamples more intelligently. We also show persistence diagrams of the original data based on the sublevel set filtration of the standard distance function (upper right, with no useful features due to noise) and the kernel distance (lower right).

A *kernel density estimate* [65, 61, 26, 27] is a way to estimate a continuous distribution function over \mathbb{R}^d for a finite point set $P \subset \mathbb{R}^d$; they have been studied and applied in a variety of contexts, for instance, under subsampling [57, 71, 3], motion planning [59], multimodality [64, 33], and surveillance [37], road reconstruction [4]. Specifically,

$$\text{KDE}_P(x) = \frac{1}{|P|} \sum_{p \in P} K(p, x).$$

The *kernel distance* [47, 42, 48, 58] (also called *current distance* or *maximum mean discrepancy*) is a metric [56, 66] between two point sets P, Q (as long as the kernel used is characteristic [66], a slight restriction of being positive definite [2, 70], this includes the Gaussian and Laplace kernels). Define a similarity between the two point sets as

$$\kappa(P, Q) = \frac{1}{|P|} \frac{1}{|Q|} \sum_{p \in P} \sum_{q \in Q} K(p, q).$$

Then the kernel distance between two point sets is defined as

$$D_K(P, Q) = \sqrt{\kappa(P, P) + \kappa(Q, Q) - 2\kappa(P, Q)}.$$

When we let point set Q be a single point x , then $\kappa(P, x) = \text{KDE}_P(x)$.

Kernel density estimates can apply to any measure μ (on \mathbb{R}^d) as $\text{KDE}_\mu(x) = \int_{p \in \mathbb{R}^d} K(p, x) d\mu(p)$. The similarity between two measures is $\kappa(\mu, \nu) = \int_{(p,q) \in \mathbb{R}^d \times \mathbb{R}^d} K(p, q) dm_{\mu, \nu}(p, q)$, where $m_{\mu, \nu}$ is the product measure of μ and ν ($m_{\mu, \nu} := \mu \otimes \nu$), and then the kernel distance between two measures μ and ν is still a metric, defined as $D_K(\mu, \nu) = \sqrt{\kappa(\mu, \mu) + \kappa(\nu, \nu) - 2\kappa(\mu, \nu)}$. When the measure ν is a Dirac measure at x ($\nu(q) = 0$ for $x \neq q$, but integrates to 1), then $\kappa(\mu, x) = \text{KDE}_\mu(x)$. Given a finite point set $P \subset \mathbb{R}^d$, we can work with the empirical measure μ_P defined as $\mu_P = \frac{1}{|P|} \sum_{p \in P} \delta_p$, where δ_p is the Dirac measure on p , and $D_K(\mu_P, \mu_Q) = D_K(P, Q)$.

If K is positive definite, it is said to have the reproducing property [2, 70]. This implies that $K(p, x)$ is an inner product in some reproducing kernel Hilbert space (RKHS) \mathcal{H}_K . Specifically, there is a lifting map $\phi : \mathbb{R}^d \rightarrow \mathcal{H}_K$ so that $K(p, x) = \langle \phi(p), \phi(x) \rangle_{\mathcal{H}_K}$, and moreover the entire set P can be represented as $\Phi(P) = \sum_{p \in P} \phi(p)$, which is a single element of \mathcal{H}_K and has a norm $\|\Phi(P)\|_{\mathcal{H}_K} = \sqrt{\kappa(P, P)}$. A single point $x \in \mathbb{R}^d$ also has a norm $\|\phi(x)\|_{\mathcal{H}_K} = \sqrt{K(x, x)}$ in this space.

1.2 Geometric Inference and Distance to a Measure: A Review

Given an unknown compact set $S \subset \mathbb{R}^d$ and a finite point cloud $P \subset \mathbb{R}^d$ that comes from S under some process, geometric inference aims to recover topological and geometric properties of S from P . The offset-based (and more generally, the distance function-based) approach for geometric inference reconstructs a geometric and topological approximation of S by offsets from P (e.g. [13, 14, 15, 20, 21]).

Given a compact set $S \subset \mathbb{R}^d$, we can define a *distance function* f_S to S ; a common example is $f_S(x) = \inf_{y \in S} \|x - y\|$ (i.e. α -shapes). The offsets of S are the sublevel sets of f_S , denoted $(S)^r = f_S^{-1}([0, r])$. Now an approximation of S by another compact set $P \subset \mathbb{R}^d$ (e.g. a finite point cloud) can be quantified by the Hausdorff distance $d_H(S, P) := \|f_S - f_P\|_\infty = \inf_{x \in \mathbb{R}^d} |f_S(x) - f_P(x)|$ of their distance functions. The intuition behind the inference of topology is that if $d_H(S, P)$ is small, thus f_S and f_P are close, and subsequently, S , $(S)^r$ and $(P)^r$ carry the same topology for an appropriate scale r . In other words, to compare the topology of offsets $(S)^r$ and $(P)^r$, we require Hausdorff stability with respect to their distance functions f_S and f_P . An example of an offset-based topological inference result is formally stated as follows (as a particular version of the reconstruction Theorem 4.6 in [14]), where the *reach* of a compact set S , $\text{reach}(S)$, is defined as the minimum distance between S and its medial axis [54].

Theorem 1.1 (Reconstruction from f_P [14]). *Let $S, P \subset \mathbb{R}^d$ be compact sets such that $\text{reach}(S) > R$ and $\varepsilon := d_H(S, P) < R/17$. Then $(S)^\eta$ and $(P)^r$ are homotopy equivalent for sufficiently small η (e.g., $0 < \eta < R$) if $4\varepsilon \leq r < R - 3\varepsilon$.*

Here $\eta < R$ ensures that the topological properties of $(S)^\eta$ and $(S)^r$ are the same, and the ε parameter ensures $(S)^r$ and $(P)^r$ are close. Typically ε is tied to the density with which a point cloud P is sampled from S .

For function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^+$ to be *distance-like* it should satisfy the following properties:

- (D1) ϕ is 1-Lipschitz: For all $x, y \in \mathbb{R}^d$, $|\phi(x) - \phi(y)| \leq \|x - y\|$.
- (D2) ϕ^2 is 1-semiconcave: The map $x \in \mathbb{R}^d \mapsto (\phi(x))^2 - \|x\|^2$ is concave.
- (D3) ϕ is proper: $\phi(x)$ tends to the infimum of its domain (e.g., ∞) as x tends to infinity.

In addition to the Hausdorff stability property stated above, as explained in [15], f_S is distance-like. These three properties are paramount for geometric inference (e.g. [14, 53]). (D1) ensures that f_S is differentiable almost everywhere and the medial axis of S has zero d -volume [15]; and (D2) is a crucial technical tool, e.g., in proving the existence of the flow of the gradient of the distance function for topological inference [14].

Distance to a measure. Given a probability measure μ on \mathbb{R}^d and a parameter $m_0 > 0$ smaller than the total mass of μ , the *distance to a measure* $d_{\mu, m_0}^{\text{CCM}} : \mathbb{R}^d \rightarrow \mathbb{R}^+$ [15] is defined for any point $x \in \mathbb{R}^d$ as

$$d_{\mu, m_0}^{\text{CCM}}(x) = \left(\frac{1}{m_0} \int_{m=0}^{m_0} (\delta_{\mu, m}(x))^2 dm \right)^{1/2}, \quad \text{where } \delta_{\mu, m}(x) = \inf \{ r > 0 : \mu(\bar{B}_r(x)) \geq m \},$$

It has been shown in [15] that $d_{\mu, m_0}^{\text{CCM}}$ is a distance-like function (satisfying (D1), (D2), and (D3)), and:

- (M4) [Stability] For probability measures μ and ν on \mathbb{R}^d and $m_0 > 0$, then $\|d_{\mu, m_0}^{\text{CCM}} - d_{\nu, m_0}^{\text{CCM}}\|_{\infty} \leq \frac{1}{\sqrt{m_0}} W_2(\mu, \nu)$.

Here W_2 is the *Wasserstein distance* [69]: $W_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\pi(x, y) \right)^{1/2}$ between two measures, where $d\pi(x, y)$ measures the amount of mass transferred from location x to location y and $\pi \in \Pi(\mu, \nu)$ is a transference plan [69].

Given a point set P , the sublevel sets of $d_{\mu_P, m_0}^{\text{CCM}}$ can be described as the union of balls [45], and then one can algorithmically estimate the topology (e.g., persistence diagram) with weighted alpha-shapes [45] and weighted Rips complexes [8].

1.3 Our Results

We show how to estimate the topology (e.g., approximate persistence diagrams, infer homotopy of compact sets) using superlevel sets of the kernel density estimate of a point set P . We accomplish this by showing that a similar set of properties hold for the kernel distance with respect to a measure μ , (in place of distance to a measure $d_{\mu, m_0}^{\text{CCM}}$), defined as

$$d_{\mu}^K(x) = D_K(\mu, x) = \sqrt{\kappa(\mu, \mu) + \kappa(x, x) - 2\kappa(\mu, x)}.$$

This treats x as a probability measure represented by a Dirac mass at x . Specifically, we show d_{μ}^K is distance-like (it satisfies (D1), (D2), and (D3)), so it inherits reconstruction properties of $d_{\mu, m_0}^{\text{CCM}}$. Moreover, it is stable with respect to the kernel distance:

- (K4) [Stability] If μ and ν are two measures on \mathbb{R}^d , then $\|d_{\mu}^K - d_{\nu}^K\|_{\infty} \leq D_K(\mu, \nu)$.

In addition, we show how to construct these topological estimates for d_{μ}^K using weighted Rips complexes, following power distance machinery introduced in [8]. That is, a particular form of power distance permits a multiplicative approximation with the kernel distance.

We also describe further advantages of the kernel distance. (i) Its sublevel sets conveniently map to the superlevel sets of a kernel density estimate. (ii) It is Lipschitz with respect to the smoothing parameter σ when the input x is fixed. (iii) As σ tends to ∞ for any two probability measures μ, ν , the kernel distance is bounded by the Wasserstein distance: $\lim_{\sigma \rightarrow \infty} D_K(\mu, \nu) \leq W_2(\mu, \nu)$. (iv) It has a small coresets representation, which allows for sparse representation and efficient, scalable computation. In particular, an ε -kernel sample [48, 57, 71] Q of μ is a finite point set whose size only depends on $\varepsilon > 0$ and such that $\max_{x \in \mathbb{R}^d} |\text{KDE}_{\mu}(x) - \text{KDE}_{\mu_Q}(x)| = \max_{x \in \mathbb{R}^d} |\kappa(\mu, x) - \kappa(\mu_Q, x)| \leq \varepsilon$. These coresets preserve inference results and persistence diagrams.

2 Kernel Distance is Distance-Like

In this section we prove d_{μ}^K satisfies (D1), (D2), and (D3); hence it is distance-like. Recall we use the σ^2 -normalized Gaussian kernel $K_{\sigma}(p, x) = \sigma^2 \exp(-\|p - x\|^2 / 2\sigma^2)$. For ease of exposition, unless otherwise noted, we will assume σ is fixed and write K instead of K_{σ} .

2.1 Semiconcave Property for d_μ^K

Lemma 2.1 (D2). $(d_\mu^K)^2$ is 1-semiconcave: the map $x \mapsto (d_\mu^K(x))^2 - \|x\|^2$ is concave.

Proof. Let $T(x) = (d_\mu^K(x))^2 - \|x\|^2$. The proof will show that the second derivative of T along any direction is nonpositive. We can rewrite

$$\begin{aligned} T(x) &= \kappa(\mu, \mu) + \kappa(x, x) - 2\kappa(\mu, x) - \|x\|^2 \\ &= \kappa(\mu, \mu) + \kappa(x, x) - \int_{p \in \mathbb{R}^d} (2K(p, x) + \|x\|^2) d\mu(p). \end{aligned}$$

Note that both $\kappa(\mu, \mu)$ and $\kappa(x, x)$ are absolute constants, so we can ignore them in the second derivative. Furthermore, by setting $t(p, x) = -2K(p, x) - \|x\|^2$, the second derivative of $T(x)$ is nonpositive if the second derivative of $t(p, x)$ is nonpositive for all $p, x \in \mathbb{R}^d$. First note that the second derivative of $-\|x\|^2$ is a constant -2 in every direction. The second derivative of $K(p, x)$ is symmetric about p , so we can consider the second derivative along any vector $u = x - p$,

$$\frac{d^2}{du^2} t(p, x) = 2 \left(\frac{\|u\|^2}{\sigma^2} - 1 \right) \exp \left(-\frac{\|u\|^2}{2\sigma^2} \right) - 2.$$

This reaches its maximum value at $\|u\| = \|x - p\| = \sqrt{3}\sigma$ where it is $4 \exp(-3/2) - 2 \approx -1.1$; this follows setting the derivative of $s(y) = 2(y-1) \exp(-y/2) - 2$ to 0, ($\frac{d}{dy} s(y) = (1/2)(3-y) \exp(-y/2)$), substituting $y = \|u\|^2/\sigma^2$. \square

We also note in Appendix A that semiconcavity follows trivially in the RKHS \mathcal{H}_K .

2.2 Lipschitz Property for d_μ^K

We generalize a (folklore, see [15]) relation between semiconcave and Lipschitz functions and prove it for completeness. A function f is ℓ -semiconcave if the function $T(x) = (f(x))^2 - \ell\|x\|^2$ is concave.

Lemma 2.2. Consider a twice-differentiable function g and a parameter $\ell \geq 1$. If $(g(x))^2$ is ℓ -semiconcave, then $g(x)$ is ℓ -Lipschitz.

Proof. The proof is by contrapositive; we assume that $g(x)$ is not ℓ -Lipschitz and then show $(g(x))^2$ cannot be ℓ -semiconcave. By this assumption, then in some direction u , there is a point x' such that $(d/du)g(x') = c > \ell \geq 1$.

Now we examine $f(x) = (g(x))^2 - \ell\|x\|^2$ at $x = x'$, and specifically its second derivative in direction u .

$$\begin{aligned} \frac{d}{du} f(x)|_{x=x'} &= 2 \left(\frac{d}{du} g(x') \right) g(x') - 2\ell\|x'\| = 2c \cdot g(x') - 2\ell\|x'\| \\ \frac{d^2}{du^2} f(x)|_{x=x'} &= 2c \left(\frac{d}{du} g(x') \right) - 2\ell = 2c^2 - 2\ell = 2(c^2 - \ell) \end{aligned}$$

Since $c^2 > c > \ell \geq 1$, then $2(c^2 - \ell) > 0$ and $f(x)$ is not ℓ -semiconcave at x' . \square

We can now state the following lemma as a corollary of Lemma 2.2 and Lemma 2.1.

Lemma 2.3 (D1). d_μ^K is 1-Lipschitz on its input.

2.3 Properness of d_μ^K

Finally, for d_μ^K to be distance-like, we need to show it is proper when its range is restricted to be less than $c_\mu := \sqrt{\kappa(\mu, \mu) + \kappa(x, x)}$. Here, the value of c_μ depends on μ not on x since $\kappa(x, x) = K(x, x) = \sigma^2$. This is required for a distance-like version [15], Proposition 4.2) of the Isotopy Lemma ([44], Proposition 1.8).

Lemma 2.4 (D3). d_μ^K is proper.

We delay this technical proof to Appendix A. The main technical difficulty comes in mapping standard definitions and approaches for distance functions to our function d_μ^K with a restricted range.

Also by properness (see discussion in Appendix A), Lemma 2.4 also implies that d_μ^K is a closed map and its levelset at any value $a \in [0, c_\mu)$ is compact. This also means that the sublevel set of d_μ^K (for ranges $[0, a) \subset [0, c_\mu)$) is compact. Since the levelset (sublevel set) of d_μ^K corresponds to the levelset (superlevel set) of KDE_μ , we have the following corollary.

Corollary 2.1. *The superlevel sets of KDE_μ for all ranges with threshold $a > 0$, are compact.*

The result in [33] shows that given a measure μ_P defined by a point set P of size n , the KDE_{μ_P} has polynomial in n modes; hence the superlevel sets of KDE_{μ_P} are compact in this setting. The above corollary is a more general statement as it holds for any measure.

3 Power Distance using Kernel Distance

A *power distance* using d_μ^K is defined with a point set $P \subset \mathbb{R}^d$ and a metric $d(\cdot, \cdot)$ on \mathbb{R}^d ,

$$f_P(\mu, x) = \sqrt{\min_{p \in P} (d(p, x)^2 + d_\mu^K(p)^2)}.$$

A point $x \in \mathbb{R}^d$ takes the distance under $d(p, x)$ to the closest $p \in P$, plus a weight from $d_\mu^K(p)$; thus a sublevel set of $f_P(\mu, \cdot)$ is defined by a union of balls. We consider a particular choice of the distance $d(p, x) := D_K(p, x)$ which leads to a kernel version of power distance

$$f_P^K(\mu, x) = \sqrt{\min_{p \in P} (D_K(p, x)^2 + d_\mu^K(p)^2)}.$$

In Section 4.2 we use $f_P^K(\mu, x)$ to adapt the construction introduced in [8] to approximate the persistence diagram of the sublevel sets of d_μ^K , using a weighted Rips filtration of $f_P^K(\mu, x)$.

Given a measure μ , let $p_+ = \arg \max_{q \in \mathbb{R}^d} \kappa(\mu, q)$, and let $P_+ \subset \mathbb{R}^d$ be a point set that contains p_+ . We show below, in Theorem 3.3 and Theorem 3.2, that $\frac{1}{\sqrt{2}}d_\mu^K(x) \leq f_{P_+}^K(\mu, x) \leq \sqrt{14}d_\mu^K(x)$. However, constructing p_+ exactly seems quite difficult. We also attempt to use $p^* = \arg \min_{p \in P} \|p - x\|$ in place of p_+ (see Section C.1), but are not able to obtain useful bounds.

Now consider an empirical measure μ_P defined by a point set P . We show (in Theorem C.2 in Appendix C.2) how to construct a point \hat{p}_+ (that approximates p_+) such that $D_K(P, \hat{p}_+) \leq (1 + \delta)D_K(P, p_+)$ for any $\delta > 0$. For a point set P , the *median concentration* Λ_P is a radius such that no point $p \in P$ has more than half of the points of P within Λ_P , and the *spread* β_P is the ratio between the longest and shortest pairwise distances. The runtime is polynomial in n and $1/\delta$ assuming β_P is bounded, and that σ/Λ_P and d are constants.

Then we consider $\hat{P}_+ = P \cup \{\hat{p}_+\}$, where \hat{p}_+ is found with $\delta = 1/2$ in the above construction. Then we can provide the following multiplicative bound, proven in Theorem 3.4. The lower bound holds independent of the choice of P as shown in Theorem 3.2.

Theorem 3.1. For any point set $P \subset \mathbb{R}^d$ and point $x \in \mathbb{R}^d$, with empirical measure μ_P defined by P , then

$$\frac{1}{\sqrt{2}}d_{\mu_P}^K(x) \leq f_{\hat{P}_+}^K(\mu_P, x) \leq \sqrt{71}d_{\mu_P}^K(x).$$

3.1 Kernel Power Distance for a Measure μ

First consider the case for a kernel power distance $f_P^K(\mu, x)$ where μ is an arbitrary measure.

Theorem 3.2. For measure μ , point set $P \subset \mathbb{R}^d$, and $x \in \mathbb{R}^d$, $D_K(\mu, x) \leq \sqrt{2}f_P^K(\mu, x)$.

Proof. Let $p = \arg \min_{q \in P} (D_K(q, x)^2 + D_K(\mu, q)^2)$. Then we can use the triangle inequality and $(D_K(\mu, p) - D_K(p, x))^2 \geq 0$ to show

$$D_K(\mu, x)^2 \leq (D_K(\mu, p) + D_K(p, x))^2 \leq 2(D_K(\mu, p)^2 + D_K(p, x)^2) = 2f_P^K(\mu, x)^2. \quad \square$$

Lemma 3.1. For measure μ , point set $P \subset \mathbb{R}^d$, point $p \in P$, and point $x \in \mathbb{R}^d$ then $f_P^K(\mu, x)^2 \leq 2D_K(\mu, x)^2 + 3D_K(p, x)^2$.

Proof. Again, we can reach this result with the triangle inequality.

$$\begin{aligned} f_P^K(\mu, x)^2 &\leq D_K(\mu, p)^2 + D_K(p, x)^2 \\ &\leq (D_K(\mu, x) + D_K(p, x))^2 + D_K(p, x)^2 \\ &\leq 2D_K(\mu, x)^2 + 3D_K(p, x)^2. \end{aligned} \quad \square$$

Recall the definition of a point $p_+ = \arg \max_{q \in \mathbb{R}^d} \kappa(\mu, q)$.

Lemma 3.2. For any measure μ and point $x, p_+ \in \mathbb{R}^d$ we have $D_K(p_+, x) \leq 2D_K(\mu, x)$.

Proof. Since x is a point in \mathbb{R}^d , $\kappa(\mu, x) \leq \kappa(\mu, p_+)$ and thus $D_K(\mu, x) \geq D_K(\mu, p_+)$. Then by triangle inequality of D_K to see that $D_K(p_+, x) \leq D_K(\mu, x) + D_K(\mu, p_+) \leq 2D_K(\mu, x)$. \square

Theorem 3.3. For any measure μ in \mathbb{R}^d and any point $x \in \mathbb{R}^d$, using the point $p_+ = \arg \max_{q \in \mathbb{R}^d} \kappa(\mu, q)$ then $f_{\{p_+\}}^K(\mu, x) \leq \sqrt{14}D_K(\mu, x)$.

Proof. Combine Lemma 3.1 and Lemma 3.2 as

$$f_{\{p_+\}}^K(\mu, x)^2 \leq 2D_K(\mu, x)^2 + 3D_K(p_+, x)^2 \leq 2D_K(\mu, x)^2 + 3(4D_K(\mu, x)^2) = 14D_K(\mu, x)^2. \quad \square$$

We now need two properties of the point set P to reach our bound, namely, the spread β_P and the median concentration Λ_P . Typically $\log(\beta_P)$ is not too large, and it makes sense to choose σ so $\sigma/\Lambda_P \leq 1$, or at least $\sigma/\Lambda_P = O(1)$.

Theorem 3.4. Consider any point set $P \subset \mathbb{R}^d$ of size n , with measure μ_P , spread β_P , and median concentration Λ_P . We can construct a point set $\hat{P}_+ = P \cup \hat{p}_+$ in $O(n^2((\sigma/\Lambda_P\delta)^d + \log(\beta)))$ time such that for any point x , $f_{\hat{P}_+}^K(\mu_P, x) \leq \sqrt{71}D_K(\mu_P, x)$.

Proof. We use Theorem C.2 to find a point \hat{p}_+ such that $D_K(P, \hat{p}_+) \leq (3/2)D_K(P, p_+)$. Thus for any $x \in \mathbb{R}^d$, using the triangle inequality

$$\begin{aligned} D_K(\hat{p}_+, x) &\leq D_K(\hat{p}_+, p_+) + D_K(p_+, x) \leq D_K(\mu_P, \hat{p}_+) + D_K(\mu_P, p_+) + D_K(p_+, x) \\ &\leq (5/2)D_K(\mu_P, p_+) + D_K(p_+, x). \end{aligned}$$

Now combine this with Lemma 3.1 and Lemma 3.2 as

$$\begin{aligned}
f_{\hat{P}_+}^K(\mu_P, x)^2 &\leq 2D_K(\mu_P, x)^2 + 3D_K(\hat{p}_+, x)^2 \\
&\leq 2D_K(\mu_P, x)^2 + 3((5/2)D_K(\mu_P, x) + D_K(p_+, x))^2 \\
&\leq 2D_K(\mu_P, x)^2 + 3((25/4) + (5/2))D_K(\mu_P, x)^2 + (1 + 5/2)D_K(p_+, x)^2 \\
&= (113/4)D_K(\mu_P, x)^2 + (21/2)D_K(p_+, x)^2 \\
&\leq (113/4)D_K(\mu_P, x)^2 + (21/2)(4D_K(\mu_P, x)^2) < 71D_K(\mu_P, x)^2. \quad \square
\end{aligned}$$

4 Reconstruction and Topological Estimation using Kernel Distance

Now applying distance-like properties from Section 2 and the power distance properties of Section 3 we can apply known reconstruction results to the kernel distance.

4.1 Homotopy Equivalent Reconstruction using d_μ^K

We have shown that the kernel distance function d_μ^K is a distance-like function. Therefore the reconstruction theory for a distance-like function [15] (which is an extension of results for compact sets [14]) holds in the setting of d_μ^K . We state the following two corollaries for completeness, whose proofs follow from the proofs of Proposition 4.2 and Theorem 4.6 in [15]. Before their formal statement, we need some notation adapted from [15] to make these statements precise. Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^+$ be a distance-like function. A point $x \in \mathbb{R}^d$ is an α -critical point if $\phi^2(x+h) \leq \phi^2(x) + 2\alpha\|h\|\phi(x) + \|h\|^2$ with $\alpha \in [0, 1]$, $\forall h \in \mathbb{R}^d$. Let $(\phi)^r = \{x \in \mathbb{R}^d \mid \phi(x) \leq r\}$ denote the sublevel set of ϕ , and let $(\phi)^{[r_1, r_2]} = \{x \in \mathbb{R}^d \mid r_1 \leq \phi(x) \leq r_2\}$ denote all points at levels in the range $[r_1, r_2]$. For $\alpha \in [0, 1]$, the α -reach of ϕ is the maximum r such that $(\phi)^r$ has no α -critical point, denoted as $\text{reach}_\alpha(\phi)$. When $\alpha = 1$, reach_1 coincides with reach introduced in [40].

Theorem 4.1 (Isotopy lemma on d_μ^K). *Let $r_1 < r_2$ be two positive numbers such that d_μ^K has no critical points in $(d_\mu^K)^{[r_1, r_2]}$. Then all the sublevel sets $(d_\mu^K)^r$ are isotopic for $r \in [r_1, r_2]$.*

Theorem 4.2 (Reconstruction on d_μ^K). *Let d_μ^K and d_ν^K be two kernel distance functions such that $\|d_\mu^K - d_\nu^K\|_\infty \leq \varepsilon$. Suppose $\text{reach}_\alpha(d_\mu^K) \geq R$ for some $\alpha > 0$. Then $\forall r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$, and $\forall \eta \in (0, R)$, the sublevel sets $(d_\mu^K)^\eta$ and $(d_\nu^K)^r$ are homotopy equivalent for $\varepsilon \leq R/(5 + 4/\alpha^2)$.*

4.2 Constructing Topological Estimates using d_μ^K

In order to actually construct a topological estimate using the kernel distance d_μ^K , one needs to be able to compute quantities related to its sublevel sets, in particular, to compute the persistence diagram of the sub-level sets filtration of d_μ^K . Now we describe such tools needed for the kernel distance based on machinery recently developed by Buchet et al. [8], which shows how to approximate the persistent homology of distance-to-a-measure for any metric space via a power distance construction. Then using similar constructions, we can use the weighted Rips filtration to approximate the persistence diagram of the kernel distance.

To state our results, first we require some technical notions and assume basic knowledge on persistent homology (see [34, 35] for a readable background). Given a metric space \mathbb{X} with the distance $d_{\mathbb{X}}(\cdot, \cdot)$, a set $P \subseteq \mathbb{X}$ and a function $w : P \rightarrow \mathbb{R}$, the (general) *power distance* f associated with (P, w) is defined as $f(x) = \sqrt{\min_{p \in P} (d_{\mathbb{X}}(p, x)^2 + w(p)^2)}$. Now given the set (P, w) and its corresponding power distance f , one could use the weighted Rips filtration to approximate the persistence diagram of w , under certain

restrictive conditions proven in Appendix D.1. Consider the sublevel set of f , $f^{-1}((-\infty, \alpha])$. It is the union of balls centered at points $p \in P$ with radius $r_p(\alpha) = \sqrt{\alpha^2 - w(p)^2}$ for each p . The weighted Čech complex $C_\alpha(P, w)$ for parameter α is the union of simplices s such that $\bigcap_{p \in s} B(p, r_p(\alpha)) \neq \emptyset$. The weighted Rips complex $R_\alpha(P, w)$ for parameter α is the maximal complex whose 1-skeleton is the same as $C_\alpha(P, w)$. The corresponding weighted Rips filtration is denoted as $\{R_\alpha(P, w)\}$.

Setting $w := d_{\mu_P}^K$ and given point set \hat{P}_+ described in Section 3, consider the weighted Rips filtration $\{R_\alpha(\hat{P}_+, d_{\mu_P}^K)\}$ based on the kernel power distance, $f_{\hat{P}_+}^K$. We view the persistence diagrams on a logarithmic scale, that is, we change coordinates of points following the mapping $(x, y) \mapsto (\ln x, \ln y)$. d_B^{\ln} denotes the corresponding bottleneck distance between persistence diagrams. We now state a corollary of Theorem 3.1.

Corollary 4.1. *The weighted Rips filtration $\{R_\alpha(\hat{P}_+, d_{\mu_P}^K)\}$ can be used to approximate the persistence diagram of $d_{\mu_P}^K$ such that $d_B^{\ln}(\text{Dgm}(d_{\mu_P}^K), \text{Dgm}(\{R_\alpha(\hat{P}_+, d_{\mu_P}^K)\})) \leq \ln(2\sqrt{71})$.*

Proof. To prove that two persistence diagrams are close, one could prove that their filtration are interleaved [12], that is, two filtrations $\{U_\alpha\}$ and $\{V_\alpha\}$ are ε -interleaved if for any α , $U_\alpha \subseteq V_{\alpha+\varepsilon} \subseteq U_{\alpha+2\varepsilon}$.

First, Lemmas D.1 and D.2 prove that the persistence diagrams $\text{Dgm}(d_{\mu_P}^K)$ and $\text{Dgm}(\{R_\alpha(\hat{P}_+, d_{\mu_P}^K)\})$ are well-defined. Second, the results of Theorem 3.1 implies an $\sqrt{71}$ multiplicative interleaving. Therefore for any $\alpha \in \mathbb{R}$,

$$(d_{\mu_P}^K)^{-1}((-\infty, \alpha]) \subset (f_{\hat{P}_+}^K)^{-1}((-\infty, \sqrt{2}\alpha]) \subset (d_{\mu_P}^K)^{-1}((-\infty, \sqrt{71}\sqrt{2}\alpha]).$$

On a logarithmic scale (by taking the natural log of both sides), such interleaving becomes additive,

$$\ln d_{\mu_P}^K - \sqrt{2} \leq \ln f_{\hat{P}_+}^K \leq \ln d_{\mu_P}^K + \sqrt{71}.$$

Theorem 4 of [16] implies

$$d_B^{\ln}(\text{Dgm}(d_{\mu_P}^K), \text{Dgm}(f_{\hat{P}_+}^K)) \leq \sqrt{71}.$$

In addition, by the Persistent Nerve Lemma ([22], Theorem 6 of [62], an extension of the Nerve Theorem [46]), the sublevel sets filtration of $d_{\mu_P}^K$, which correspond to unions of balls of increasing radius, has the same persistent homology as the nerve filtration of these balls (which, by definition, is the Čech filtration). Finally, there exists a multiplicative interleaving between weighted Rips and Čech complexes (Proposition 31 of [16]), $C_\alpha \subseteq R_\alpha \subseteq C_{2\alpha}$. We then obtain the following bounds on persistence diagrams,

$$d_B^{\ln}(\text{Dgm}(f_{\hat{P}_+}^K), \text{Dgm}(\{R_\alpha(P_+, d_{\mu_P}^K)\})) \leq \ln(2).$$

We use triangle inequality to obtain the final result:

$$d_B^{\ln}(\text{Dgm}(d_{\mu_P}^K), \text{Dgm}(\{R_\alpha(P_+, d_{\mu_P}^K)\})) \leq \ln(2\sqrt{71}). \quad \square$$

Based on Corollary 4.1, we have an algorithm that approximates the persistent homology of the sublevel set filtration of $d_{\mu_P}^K$ by constructing the weighted Rips filtration corresponding to the kernel-based power distance and computing its persistent homology. For memory efficient computation, sparse (weighted) Rips filtrations could be adapted by considering simplices on subsamples at each scale [63, 16], although some restrictions on the space apply.

4.3 Distance to the Support of a Measure vs. Kernel Distance

Suppose μ is a uniform measure on a compact set S in \mathbb{R}^d . We now compare the kernel distance d_{μ}^K with the distance function f_S to the support S of μ . We show how d_{μ}^K approximates f_S , and thus allows one to infer geometric properties of S from samples from μ .

A generalized gradient and its corresponding flow associated with a distance function are described in [14] and later adapted for distance-like functions in [15]. Let $f_S : \mathbb{R}^d \rightarrow \mathbb{R}$ be a distance function associated with a compact set S of \mathbb{R}^d . It is not differentiable on the medial axis of S . A *generalized gradient function* $\nabla_S : \mathbb{R}^d \rightarrow \mathbb{R}^d$ coincides with the usual gradient of f_S where f_S is differentiable, and is defined everywhere and can be integrated into a continuous flow $\Phi^t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that points away from S . Let γ be an integral (flow) line. The following lemma shows that when close enough to S , that d_μ^K is strictly increasing along any γ . The proof is quite technical and is thus deferred to Appendix D.2.

Lemma 4.1. *Given any flow line γ associated with the generalized gradient function ∇_S , $d_\mu^K(x)$ is strictly monotonically increasing along γ for x sufficiently far away from the medial axis of S , for $\sigma \leq \frac{R}{6\Delta_G}$ and $f_S(x) \in (0.014R, 2\sigma)$. Here $B(\sigma/2)$ denotes a ball of radius $\sigma/2$, $G := \frac{\text{Vol}(B(\sigma/2))}{\text{Vol}(S)}$, $\Delta_G := \sqrt{12 + 3\ln(4/G)}$ and suppose $R := \min(\text{reach}(S), \text{reach}(\mathbb{R}^d \setminus S)) > 0$.*

The strict monotonicity of d_μ^K along the flow line under the conditions in Lemma 4.1 makes it possible to define a deformation retract of the sublevel sets of d_μ^K onto sublevel sets of f_S . Such a deformation retract defines a special case of homotopy equivalence between the sublevel sets of d_μ^K and sublevel sets of f_S . Consider a sufficiently large point set $P \in \mathbb{R}^d$ sampled from μ , and its induced measure μ_P . We can then also invoke Theorem 4.2 and a sampling bound (see Section 6 and Lemma B.2) to show homotopy equivalence between the sublevel sets of f_S and $d_{\mu_P}^K$.

Note that Lemma 4.1 uses somewhat restrictive conditions related to the reach of a compact set, however we believe such conditions could be further relaxed to be associated with the concept of μ -reach as described in [14].

5 Stability Properties for the Kernel Distance to a Measure

Lemma 5.1 (K4). *For two measures μ and ν on \mathbb{R}^d we have $\|d_\mu^K - d_\nu^K\|_\infty \leq D_K(\mu, \nu)$.*

Proof. Since $D_K(\cdot, \cdot)$ is a metric, then by triangle inequality, for any $x \in \mathbb{R}^d$ we have $D_K(\mu, x) \leq D_K(\mu, \nu) + D_K(\nu, x)$ and $D_K(\nu, x) \leq D_K(\nu, \mu) + D_K(\mu, x)$. Therefore for any $x \in \mathbb{R}^d$ we have $|D_K(\mu, x) - D_K(\nu, x)| \leq D_K(\mu, \nu)$, proving the claim. \square

Both the Wasserstein and kernel distance are *integral probability metrics* [66], so (M4) and (K4) are both interesting, but not easily comparable. We now attempt to reconcile this.

5.1 Comparing D_K to W_2

Lemma 5.2. *There is no Lipschitz constant γ such that for any two probability measures μ and ν we have $W_2(\mu, \nu) \leq \gamma D_K(\mu, \nu)$.*

Proof. Consider two measures μ and ν which are almost identical: the only difference is some mass of measure τ is moved from its location in μ a distance n in ν . The Wasserstein distance requires a transportation plan that moves this τ mass in ν back to where it was in μ with cost $\tau \cdot \Omega(n)$ in $W_2(\mu, \nu)$. On the other hand, $D_K(\mu, \nu) = \sqrt{\kappa(\mu, \mu) + \kappa(\nu, \nu) - 2\kappa(\mu, \nu)} \leq \sqrt{\sigma^2 + \sigma^2 - 2 \cdot 0} = \sqrt{2}\sigma$ is bounded. \square

We conjecture for any two probability measures μ and ν that $D_K(\mu, \nu) \leq W_2(\mu, \nu)$. This would show that d_μ^K is at least as stable as $d_{\mu, m_0}^{\text{CCM}}$ since a bound on $W_2(\mu, \nu)$ would also bound $D_K(\mu, \nu)$, but not vice versa. Alternatively, this can be viewed as d_μ^K is less discriminative than $d_{\mu, m_0}^{\text{CCM}}$; we view this as a positive in this setting, as it is mainly less discriminative towards outliers (far away points). Here we only show that this property for a special case and as $\sigma \rightarrow \infty$. To simplify notation, all integrals are assumed to be over the full domain \mathbb{R}^d .

Two Dirac masses. We first consider a special case when μ is a Dirac mass at a point p and ν is a Dirac mass at a point q . That is they are both single points. We can then write $D_K(\mu, \nu) = D_K(p, q)$. Figure 2 illustrates the result of this lemma.

Lemma 5.3. *For any points $p, q \in \mathbb{R}^d$ it always holds that $\|p - q\| \geq D_K(p, q)$. When $\|p - q\| \leq \sqrt{3}\sigma$ then $D_K(p, q) \geq \|p - q\|/2$.*

Proof. First expand $D_K(p, q)^2$ as

$$D_K(p, q)^2 = 2\sigma^2 - 2K(p, q) = 2\sigma^2 \left(1 - \exp\left(\frac{-\|p - q\|^2}{2\sigma^2}\right) \right).$$

Now using that $1 - t \leq e^{-t} \leq 1 - t + (1/2)t^2$ for $t \geq 0$

$$D_K(p, q)^2 = 2\sigma^2 \left(1 - \exp\left(\frac{-\|p - q\|^2}{2\sigma^2}\right) \right) \leq 2\sigma^2 \left(\frac{\|p - q\|^2}{2\sigma^2} \right) = \|p - q\|^2$$

and

$$\begin{aligned} D_K(p, q)^2 &= 2\sigma^2 \left(1 - \exp\left(\frac{-\|p - q\|^2}{2\sigma^2}\right) \right) \\ &\geq 2\sigma^2 \left(\frac{\|p - q\|^2}{2\sigma^2} - \frac{1}{2} \frac{\|p - q\|^4}{4\sigma^4} \right) \\ &= \frac{\|p - q\|^2}{4} \left(4 - \frac{\|p - q\|^2}{\sigma^2} \right) \\ &\geq \|p - q\|^2/4, \end{aligned}$$

where the last inequality holds when $\|p - q\|^2 \leq \sqrt{3}\sigma$. \square

One Dirac mass. Consider the case where one measure ν is a Dirac mass at point $x \in \mathbb{R}^d$.

Lemma 5.4. *Consider two probability measures μ and ν on \mathbb{R}^d where ν is represented by a Dirac mass at a point $x \in \mathbb{R}^d$. Then $d_{\mu}^K(x) = D_K(\mu, \nu) \leq W_2(\mu, \nu)$ for any $\sigma > 0$, where the equality only holds when μ is also a Dirac mass at x .*

Proof. Since both $W_2(\mu, \nu)$ and $D_K(\mu, \nu)$ are metrics and hence non-negative, we can instead consider their squared versions: $(W_2(\mu, \nu))^2 = \int_p \|p - x\|^2 \mu(p) dp$ and

$$\begin{aligned} (D_K(\mu, \nu))^2 &= K(x, x) + \int_{(p,q)} K(p, q) d\mathbf{m}_{\mu, \mu}(p, q) - 2 \int_p K(p, x) d\mu(p) \\ &= \sigma^2 \left(1 + \int_{(p,q)} \exp\left(\frac{-\|p - q\|^2}{2\sigma^2}\right) d\mathbf{m}_{\mu, \mu}(p, q) - 2 \int_p \exp\left(\frac{-\|p - x\|^2}{2\sigma^2}\right) d\mu(p) \right). \end{aligned}$$

Now use the bound $1 - t \leq e^{-t} \leq 1$ for $t \geq 0$ to approximate

$$\begin{aligned} (D_K(\mu, \nu))^2 &\leq \sigma^2 \left(1 + \int_{(p,q)} (1) d(\mathbf{m}_{\mu, \mu}(p, q)) - 2 \int_p \left(1 - \frac{\|p - x\|^2}{2\sigma^2} \right) d\mu(p) \right) \\ &= \int_p \|p - x\|^2 d\mu(p) = (W_2(\mu, \nu))^2. \end{aligned}$$

The inequality becomes an equality only when $\|p - x\| = 0$ for all $p \in P$, and since they are both metrics, this is the only location where they are both 0. \square

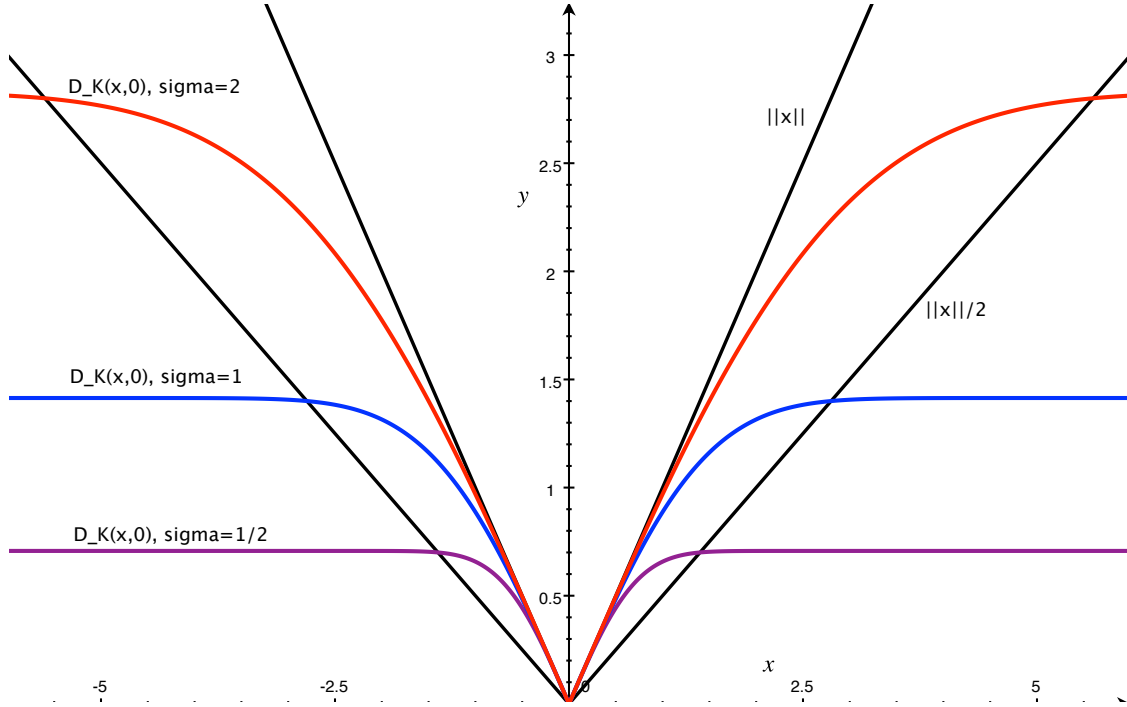


Figure 2: Showing that $\|x - 0\|/2 \leq D_K(x, 0) \leq \|x - 0\|$, where the second inequality holds for $\|x\| \leq \sqrt{3}\sigma$. The kernel distance $D_K(x, 0)$ is shown for $\sigma = \{1/2, 1, 2\}$ in purple, blue, and red, respectively.

General case. Next we show that if ν is not a unit Dirac, then this inequality holds in the limit as σ goes to infinity. The technical work is making precise how $\sigma^2 - K(p, x) \leq \|x - p\|^2/2$ and how this compares to bounds on $D_K(\mu, \nu)$ and $W_2(\mu, \nu)$.

For simpler exposition, we assume μ is a probability measure, that is $\int_p \mu(p) dp = 1$; otherwise we can normalize μ at the appropriate locations, and all of the results go through.

Lemma 5.5. For any $p, q \in \mathbb{R}^d$ we have $K(p, q) = \sigma^2 - \frac{\|p - q\|^2}{2} + \sum_{i=2}^{\infty} \frac{(-\|p - q\|^2)^i}{2^{i+1}\sigma^{2i-2}i!}$.

Proof. We use the Taylor expansion of $e^x = \sum_{i=0}^{\infty} x^i/i! = 1 + x + \sum_{i=2}^{\infty} x^i/i!$. Then it is easy to see

$$K(p, q) = \sigma^2 \exp\left(-\frac{\|p - q\|^2}{2\sigma^2}\right) = \sigma^2 - \frac{\|p - q\|^2}{2} + \sum_{i=2}^{\infty} \frac{(-\|p - q\|^2)^i}{2^i \sigma^{2i-2} i!}. \quad \square$$

This lemma illustrates why the choice of coefficient of σ^2 is convenient. Since then $\sigma^2 - K(p, q)$ acts like $\frac{1}{2}\|p - q\|^2$, and becomes closer as σ increases. Define $\bar{\mu} = \int_p p \cdot d\mu(p)$ to represent the mean point of measure μ ; $\text{Var}(\mu) = (\int_p \|p\|^2 d\mu(p)) - \|\bar{\mu}\|^2$ to represent the variance of the measure μ ; and $\Delta_{\mu, \nu} = \int_{(p, q)} \sum_{i=2}^{\infty} \frac{(-\|p - q\|^2)^i}{2^i \sigma^{2i-2} i!} d\mathfrak{m}_{\mu, \nu}(p, q)$.

Lemma 5.6. For any $x \in \mathbb{R}^d$ we have $\int_p \|p - x\|^2 d\mu(p) = \|\bar{\mu} - x\|^2 + \text{Var}(\mu)$.

Proof.

$$\begin{aligned}
\int_p \|p - x\|^2 d\mu(p) &= \int_p (\|p\|^2 + \|x\|^2 - 2\langle p, x \rangle) d\mu(p) \\
&= \int_p \|p\|^2 d\mu(p) + \|x\|^2 - 2 \int_p \langle p, x \rangle d\mu(p) \\
&= \left(\int_p \|p\|^2 d\mu(p) - \|\bar{\mu}\|^2 \right) + \|\bar{\mu}\|^2 + \|x\|^2 - 2\langle \bar{\mu}, x \rangle \\
&= \text{Var}(\mu) + \|\bar{\mu} - x\|^2. \quad \square
\end{aligned}$$

Lemma 5.7. For probability measures μ and ν on \mathbb{R}^d , $\kappa(\mu, \nu) = \sigma^2 - \frac{1}{2} (\|\bar{\mu} - \bar{\nu}\|^2 + \text{Var}(\mu) + \text{Var}(\nu)) + \Delta_{\mu, \nu}$.

Proof. We use Lemma 5.5 to expand

$$\begin{aligned}
\kappa(\mu, \nu) &= \int_{(p, q)} K(p, q) d\mathbf{m}_{\mu, \nu}(p, q) \\
&= \sigma^2 - \int_{(p, q)} \left(\frac{\|p - q\|^2}{2} - \sum_{i=2}^{\infty} \frac{(-\|p - q\|^2)^i}{2^{i+1} \sigma^{2i-2} i!} \right) d\mathbf{m}_{\mu, \nu}(p, q).
\end{aligned}$$

After shifting the $\Delta_{\mu, \nu}$ term outside, we can use Lemma 5.6 (twice) to rewrite

$$\begin{aligned}
\int_p \left(\int_q \|p - q\|^2 d\nu(q) \right) d\mu(p) &= \int_p (\|p - \bar{\nu}\|^2 + \text{Var}(\nu)) d\mu(p) \\
&= \|\bar{\mu} - \bar{\nu}\|^2 + \text{Var}(\mu) + \text{Var}(\nu). \quad \square
\end{aligned}$$

Theorem 5.1. For any two probability measures μ and ν defined on \mathbb{R}^d $\lim_{\sigma \rightarrow \infty} D_K(\mu, \nu) = \|\bar{\mu} - \bar{\nu}\|$.

Proof. First expand

$$\begin{aligned}
(D_K(\mu, \nu))^2 &= \kappa(\mu, \mu) + \kappa(\nu, \nu) - 2\kappa(\mu, \nu) \\
&= \left(\sigma^2 - \frac{1}{2} \|\bar{\mu} - \bar{\mu}\|^2 - \text{Var}(\mu) + \Delta_{\mu, \mu} \right) + \left(\sigma^2 - \frac{1}{2} \|\bar{\nu} - \bar{\nu}\|^2 - \text{Var}(\nu) + \Delta_{\nu, \nu} \right) \\
&\quad - 2 \left(\sigma^2 - \frac{1}{2} \|\bar{\mu} - \bar{\nu}\|^2 - \frac{1}{2} \text{Var}(\mu) - \frac{1}{2} \text{Var}(\nu) + \Delta_{\mu, \nu} \right) \\
&= \|\bar{\mu} - \bar{\nu}\|^2 + \Delta_{\mu, \mu} + \Delta_{\nu, \nu} - 2\Delta_{\mu, \nu}.
\end{aligned}$$

Finally we observe that since all terms of $\Delta_{\mu, \nu}$ are divided by σ^2 or larger powers of σ . Thus as σ increases $\Delta_{\mu, \nu}$ approaches 0 and $(D_K(\mu, \nu))^2$ approaches $\|\bar{\mu} - \bar{\nu}\|^2$, completing the proof. \square

Now we can relate $D_K(\mu, \nu)$ to $W_2(\mu, \nu)$ through $\|\bar{\mu} - \bar{\nu}\|$. The next result is a known lower bounds for the Earth movers distance [23][Theorem 7]. We reprove it in Appendix E for completeness.

Lemma 5.8. For any probability measures μ and ν defined on \mathbb{R}^d we have $\|\bar{\mu} - \bar{\nu}\| \leq W_2(\mu, \nu)$.

We can now combine these results to achieve the following theorem.

Theorem 5.2. For any two probability measures μ and ν defined on \mathbb{R}^d $\lim_{\sigma \rightarrow \infty} D_K(\mu, \nu) = \|\bar{\mu} - \bar{\nu}\|$ and $\|\bar{\mu} - \bar{\nu}\| \leq W_2(\mu, \nu)$. Thus $\lim_{\sigma \rightarrow \infty} D_K(\mu, \nu) \leq W_2(\mu, \nu)$.

5.2 Kernel Distance Stability with Respect to σ

We now explore the Lipschitz properties of d_μ^K with respect to the noise parameter σ . We argue any distance function that is robust to noise needs some parameter to address how many outliers to ignore or how far away a point is that is an outlier. For instance, this parameter in $d_{\mu, m_0}^{\text{CCM}}$ is m_0 which controls the amount of measure μ to be used in the distance.

Here we show that d_μ^K has a particularly nice property, that it is Lipschitz with respect to the choice of σ for any fixed x . The larger σ the more effect outliers have, and the smaller σ the less the data is smoothed and thus the closer the noise needs to be to the underlying object to effect the inference.

Lemma 5.9. *Let $h(\sigma, z) = \exp(-z^2/2\sigma^2)$. We can bound $h(\sigma, z) \leq 1$, $\frac{d}{d\sigma}h(\sigma, z) \leq (2/e)/\sigma$ and $\frac{d^2}{d\sigma^2}h(\sigma, z) \leq (18/e^3)/\sigma^2$ over any choice of $z > 0$.*

Proof. The first bound follows from $y = -z^2/2\sigma^2 \leq 0$ and $\exp(y) \leq 1$ for $y \leq 0$.

Next we define

$$w_1(\sigma, z) = \frac{d}{d\sigma}h(\sigma, z) = \frac{z^2}{\sigma^3} \exp\left(\frac{-z^2}{2\sigma^2}\right), \text{ and}$$

$$w_2(\sigma, z) = \frac{d^2}{d\sigma^2}h(\sigma, z) = \left(\frac{z^4}{\sigma^6} - \frac{3z^2}{\sigma^4}\right) \exp\left(\frac{-z^2}{2\sigma^2}\right).$$

Now to solve the first part, we differentiate w_1 with respect to z to find its maximum over all choices of z .

$$\frac{d}{dz}w_1(\sigma, z) = \left(\frac{2z}{\sigma^3} - \frac{z^3}{\sigma^5}\right) \exp\left(\frac{-z^2}{2\sigma^2}\right)$$

Where $(d/dz)w_1(\sigma, z) = 0$ at $z = 0$, $z = \sqrt{2}\sigma$ and as z approaches ∞ . Thus the maximum must occur at one of these values. Both $w_1(\sigma, 0) = 0$ and $\lim_{z \rightarrow \infty} w_1(\sigma, z) = 0$, while $w_1(\sigma, \sqrt{2}\sigma) = (2/e)/\sigma$, proving the first part.

To solve the second part, we perform the same approach on w_2 .

$$\begin{aligned} \frac{d}{dz}w_2(\sigma, z) &= \left(\frac{-z^5}{\sigma^8} + \frac{3z^3}{\sigma^6} + \frac{4z^3}{\sigma^6} - \frac{6z}{\sigma^4}\right) \exp\left(\frac{-z^2}{2\sigma^2}\right) \\ &= \frac{z}{\sigma^4} \left(\frac{-z^4}{\sigma^4} + \frac{7z^2}{\sigma^2} - 6\right) \exp\left(\frac{-z^2}{2\sigma^2}\right) \end{aligned}$$

Thus $(d/dz)w_2(\sigma, z) = 0$ at $z = \{0, \sigma, \sqrt{6}\sigma\}$ and as z goes to ∞ for $z \in [0, \infty)$. Both $w_2(\sigma, 0) = 0$ and $\lim_{z \rightarrow \infty} w_2(\sigma, z) = 0$. The minimum occurs at $w_2(\sigma, z = \sigma) = (-2/\sqrt{e})/\sigma^2$. The maximum occurs at $w_2(\sigma, z = \sqrt{6}\sigma) = (18/e^3)/\sigma^2$. \square

Theorem 5.3. *For any measure μ defined on \mathbb{R}^d and $x \in \mathbb{R}^d$, $d_\mu^K(x)$ is ℓ -Lipschitz with respect to σ , for $\ell = 18/e^3 + 8/e + 2 < 6$.*

Proof. Recall that $m_{\mu, \nu}$ is the product measure of any μ and ν , and define $M_{\mu, \nu}$ as $M_{\mu, \nu}(p, q) = m_{\mu, \mu}(p, q) + m_{\nu, \nu}(p, q) - 2m_{\mu, \nu}(p, q)$. It is now useful to define a function $f_x(\sigma)$ as

$$f_x(\sigma) = \int_{(p, q)} \exp\left(\frac{-\|p - q\|^2}{2\sigma^2}\right) dM_{\mu, \delta_x}(p, q).$$

So $d_\mu^K(x) = \sigma \sqrt{f_x(\sigma)}$ and we can write another function as

$$F(\sigma) = (d_\mu^K(x))^2 - \ell \|\sigma\|^2 = \sigma^2 f_x(\sigma) - \ell \sigma^2.$$

Now to prove $d_\mu^K(x)$ is ℓ -Lipschitz, we can show that $(d_\mu^K)^2$ is ℓ -semiconcave with respect to σ , and apply Lemma 2.2. This boils down to showing the second derivative of $F(\sigma)$ is always non-positive.

$$\frac{d}{d\sigma}F(\sigma) = 2\sigma f_x(\sigma) + \sigma^2 \frac{d}{d\sigma}f_x(\sigma) - 2\sigma\ell.$$

$$\frac{d^2}{d\sigma^2}F(\sigma) = \sigma^2 \frac{d^2}{d\sigma^2}f_x(\sigma) + 4\sigma \frac{d}{d\sigma}f_x(\sigma) + 2f_x(\sigma) - 2\ell.$$

First we note that since $\int_{(p,q)} c \cdot d\mathbf{m}_{\mu,\nu}(p, q) = c$ for any product distribution $\mathbf{m}_{\mu,\nu}$ of two distributions μ and ν , including when μ or ν is a Dirac mass, then

$$\int_{(p,q)} c \cdot d\mathbf{M}_{\mu,\delta_x}(p, q) = \int_{(p,q)} c \cdot d[\mathbf{m}_{\mu,\mu} + \mathbf{m}_{\delta_x,\delta_x} - 2\mathbf{m}_{\mu,\delta_x}](p, q) \leq 2c.$$

Thus since $\exp\left(\frac{-\|p-q\|^2}{2\sigma^2}\right)$ is in $[0, 1]$ for all choices of p, q , and $\sigma > 0$, then $0 \leq f_x(\sigma) \leq 2$ and $2f_x(\sigma) \leq 4$.

This bounds the third term in $\frac{d^2}{d\sigma^2}F(\sigma)$, we now need to use a similar approach to bound the first and second terms.

Let $h(\sigma, z) = \exp\left(\frac{-z^2}{2\sigma^2}\right)$, so we can apply Lemma 5.9.

$$4\sigma \frac{d}{d\sigma}f_x(\sigma) = 4\sigma \int_{(p,q)} \left(\frac{d}{d\sigma}h(\sigma, \|p-q\|)\right) d\mathbf{M}_{\mu,\delta_x}(p, q) \leq 4\sigma((2/e)/\sigma)2 = 16/e$$

$$\sigma^2 \frac{d^2}{d\sigma^2}f_x(\sigma) = \sigma^2 \int_{(p,q)} \left(\frac{d^2}{d\sigma^2}h(\sigma, \|p-q\|)\right) d\mathbf{M}_{\mu,\delta_x}(p, q) \leq \sigma^2((18/e^3)/\sigma^2)2 = 36/e^3$$

Then we complete the proof using the upper bound of each item of $\frac{d^2}{d\sigma^2}F(\sigma)$

$$\begin{aligned} \frac{d^2}{d\sigma^2}F(\sigma) &= \sigma^2 \frac{d^2}{d\sigma^2}f_x(\sigma) + 4\sigma \frac{d}{d\sigma}f_x(\sigma) + 2f_x(\sigma) - 2\ell \\ &\leq 36/e^3 + 16/e + 4 - 2(18/e^3 + 8/e + 2) = 0. \end{aligned} \quad \square$$

Lipschitz in m_0 for d_{μ,m_0}^{CCM} . We show that there is no Lipschitz property for d_{μ,m_0}^{CCM} , with respect to m_0 that is independent of the measure μ . Consider a measure μ_P for point set $P \subset \mathbb{R}$ consisting of two points at $a = 0$ and at $b = \Delta$. Now consider $d_{\mu_P,m_0}^{\text{CCM}}(a)$. When $m_0 \leq 1/2$ then $d_{\mu_P,m_0}^{\text{CCM}}(a) = 0$ is constant. But for $m_0 = 1/2 + \alpha$ for $\alpha > 0$, then $d_{\mu_P,m_0}^{\text{CCM}}(a) = \alpha\Delta/(1/2 + \alpha)$ and $\frac{d}{dm_0}d_{\mu_P,m_0}^{\text{CCM}}(a) = \frac{d}{d\alpha}d_{\mu_P,1/2+\alpha}^{\text{CCM}}(a) = \frac{(1/2+2\alpha)\Delta}{(1/2+\alpha)^2}$, which is maximized as α approaches 0 with an infimum of 2Δ . If $n-1$ points are at b and 1 point at a , then the infimum of the first derivative of m_0 is $n\Delta$. Hence for a measure μ_P defined by a point set, the infimum of $\frac{d}{dm_0}d_{\mu_P,m_0}^{\text{CCM}}(a)$ and, hence a lower bound on the Lipschitz constant is $n\Delta_P$ where $\Delta_P = \max_{p,p' \in P} \|p - p'\|$.

6 Algorithmic and Approximation Observations

Kernel coresets. The kernel distance is robust under random samples [48]. Specifically, if Q is a point set randomly chosen from μ of size $O((1/\varepsilon^2)(d + \log(1/\delta)))$ then $\|\text{KDE}_\mu - \text{KDE}_Q\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$. We call such a subset Q and ε -kernel sample of (μ, K) . Furthermore, it is also possible to construct ε -kernel samples Q with even smaller size of $|Q| = O(((1/\varepsilon)\sqrt{\log(1/\varepsilon\delta)})^{2d/(d+2)})$ [57]; in particular in \mathbb{R}^2 the required size is $|Q| = O((1/\varepsilon)\sqrt{\log(1/\varepsilon\delta)})$. Exploiting the above constructions,

recent work [71] builds a data structure to allow for efficient approximate evaluations of KDE_P where $|P| = 100,000,000$.

These constructions of Q also immediately imply that $\|(d_\mu^K)^2 - (d_Q^K)^2\|_\infty \leq 4\varepsilon$ since $(d_\mu^K(x))^2 = \kappa(\mu, \mu) + \kappa(x, x) - 2\text{KDE}_\mu(x)$, and both the first and third term incur at most 2ε error in converting to $\kappa(Q, Q)$ and $2\text{KDE}_Q(x)$, respectively (see Lemma B.1). Thus, an $(\varepsilon^2/4)$ -kernel sample Q of (μ, K) implies that $\|d_\mu^K - d_Q^K\|_\infty \leq \varepsilon$ (see Lemma B.2).

This implies algorithms for geometric inference on enormous noisy data sets. Moreover, if we assume an input point set Q is drawn iid from some underlying, but unknown distribution μ , we can bound approximations with respect to μ .

Corollary 6.1. *Consider a measure μ defined on \mathbb{R}^d , a kernel K , and a parameter $\varepsilon \leq R(5 + 4/\alpha^2)$. We can create a coresets Q of size $|Q| = O(((1/\varepsilon^2)\sqrt{\log(1/\varepsilon\delta)})^{2d/(d+2)})$ or randomly sample $|Q| = O((1/\varepsilon^4)(d + \log(1/\delta)))$ points so, with probability at least $1 - \delta$, any sublevel set $(d_\mu^K)^\eta$ is homotopy equivalent to $(d_Q^K)^r$ for $r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$ and $\eta \in (0, R)$.*

Proof. Those bounds are obtained by constructing an $(\varepsilon^2/4)$ -kernel sample [48, 57], which guarantees $\|d_\mu^K - d_Q^K\|_\infty \leq \varepsilon$ via Lemma B.2. Then since $\varepsilon \leq R/(5 + 4/\alpha^2)$, with Theorem 4.2 any sublevel set $(d_\mu^K)^\eta$ is homotopy equivalent to $(d_Q^K)^r$. \square

Stability of persistence diagrams. Furthermore, the stability results on persistence diagrams [24] hold for kernel density estimates and kernel distance of μ and Q (where Q is a coresets of μ with the same size bounds as above). If $\|f - g\|_\infty \leq \varepsilon$, then $d_B(\text{Dgm}(f), \text{Dgm}(g)) \leq \varepsilon$, where d_B is the bottleneck distance between persistence diagrams. Combined with the coresets results above, this immediately implies the following corollaries.

Corollary 6.2. *Consider a measure μ defined on \mathbb{R}^d and a kernel K . We can create a core set Q of size $|Q| = O(((1/\varepsilon)\sqrt{\log(1/\varepsilon\delta)})^{2d/(d+2)})$ or randomly sample $|Q| = O((1/\varepsilon^2)(d + \log(1/\delta)))$ points which will have the following properties with probability at least $1 - \delta$.*

- $d_B(\text{Dgm}(\text{KDE}_\mu), \text{Dgm}(\text{KDE}_Q)) \leq \varepsilon$.
- $d_B(\text{Dgm}((d_\mu^K)^2), \text{Dgm}((d_Q^K)^2)) \leq \varepsilon$.

Corollary 6.3. *Consider a measure μ defined on \mathbb{R}^d and a kernel K . We can create a coresets Q of size $|Q| = O(((1/\varepsilon^2)\sqrt{\log(1/\varepsilon\delta)})^{2d/(d+2)})$ or randomly sample $|Q| = O((1/\varepsilon^4)(d + \log(1/\delta)))$ points which will have the following property with probability at least $1 - \delta$.*

- $d_B(\text{Dgm}(d_\mu^K), \text{Dgm}(d_Q^K)) \leq \varepsilon$.

Another bound was independently derived to show an upper bound on the size of a random sample Q such that $d_B(\text{Dgm}(\text{KDE}_{\mu_P}), \text{Dgm}(\text{KDE}_Q)) \leq \varepsilon$ in [3]; this can, as above, also be translated into bounds for $\text{Dgm}((d_\mu^K)^2)$ and $\text{Dgm}(d_Q^K)$. This result assumes $P \subset [-C, C]^d$ and is parametrized by a bandwidth parameter h that retains that $\int_{x \in \mathbb{R}^d} K_h(x, p) dx = 1$ for all p using that $K_1(\|x - p\|) = K(x, p)$ and $K_h(\|x - p\|) = \frac{1}{h^d} K_1(\|x - p\|^2/h)$. This ensures that $K(\cdot, p)$ is $(1/h^d)$ -Lipschitz and that $K(x, x) = \Theta(1/h^d)$ for any x . Then their bound requires $|Q| = O(\frac{d}{\varepsilon^2 h^d} \log(\frac{Cd}{\varepsilon \delta h}))$ random samples.

To compare directly against the random sampling result we derive from Joshi *et al.* [48], for kernel $K_h(x, p)$ then $\|\text{KDE}_{\mu_P} - \text{KDE}_Q\|_\infty \leq \varepsilon K_h(x, x) = \varepsilon/h^d$. Hence, our analysis requires $|Q| = O((1/\varepsilon^2 h^{2d})(d + \log(1/\delta)))$, and is an improvement when $h = \Omega(1)$ or C is not known or bounded, as well as in some other cases as a function of ε, h, δ , and d .

7 Discussion

We mention here a few other interesting aspects of our results and observations about topological inference using the kernel distance. They are related to how the noise parameter σ affects the idea of scale, and a few more experiments, including with alternate kernels.

7.1 Noise and Scale

Much of geometric and topological reconstruction grew out of the desire to understand shapes at various scales. A common mechanism is offset based; e.g., α -shapes [31] represent the scale of a shape with the α parameter controlling the offsets of a point cloud. There are two parameters with the kernel distance: r controls the offset through the sublevel set of the function, and σ controls the noise. We argue that any function which is robust to noise must have a parameter that controls the noise (e.g. σ for d_μ^K and m_0 for $d_{\mu, m_0}^{\text{CCM}}$). Here σ clearly defines some sense of scale in the setting of density estimation [65] and has a geometrical interpretation, while m_0 represents a fraction of the measure and is hard to interpret geometrically, as illustrated by the lack of a Lipschitz property for $d_{\mu, m_0}^{\text{CCM}}$ with respect to m_0 .

There are several experiments below, in Section 7.2, from which several insights can be drawn. One observation is that even though there are two parameters r and σ that control the scale, the interesting values typically have r very close to σ . Thus, we recommend to first set σ to control the scale at which the data is studied, and then explore the effect of varying r for values near σ . Moreover, not much structure seems to be missed by not exploring the space of both parameters; Figure 3 shows that fixing one (of r and σ) and varying the other can provide very similar superlevel sets. However, it is possible instead to fix r and explore the persistent topological features in the data [36, 34] (those less affected by smoothing) by varying σ . On the other hand, it remains a challenging problem to study two parameter persistent homology [10, 9] under the setting of kernel distance (or kernel density estimate).

7.2 Experiments

We consider measures μ_P defined by a point set $P \subset \mathbb{R}^2$. To experimentally visualize the structures of the superlevel sets of kernel density estimates, or equivalently sublevel sets of the kernel distance, we do the simplest thing and just evaluate $d_{\mu_P}^K$ at every grid point on a sufficiently dense grid.

Grid approximation. Due to the 1-Lipschitz property of the kernel distance, well chosen grid points have several nice properties. We consider the functions up to some resolution parameter $\varepsilon > 0$, consistent with the parameter used to create a coresets approximation Q . Now specifically, consider an axis-aligned grid $G_{\varepsilon, d}$ with edge length $\varepsilon/2\sqrt{d}$ so no point $x \in \mathbb{R}^d$ is further than ε from some grid point $g \in G_{\varepsilon, d}$. Since $K(x, y) \leq \varepsilon$ when $\|x - y\| \geq 2\sigma^2 \ln(\sigma^2/\varepsilon) = \delta_{\varepsilon, \sigma}$, we only need to consider grid points $g \in G_{\varepsilon, d}$ which are within $\delta_{\varepsilon, \sigma}$ of some point $p \in P$ (or $q \in Q$, of coresets Q of P) [48, 71]. This is at most $(2\sqrt{d}/\varepsilon)^d (2\delta_{\varepsilon, \sigma})^d = O((\sigma^2 \log(\varepsilon/d)/\varepsilon)^d)$ grid points total for d a fixed constant. Furthermore, due to the 1-Lipschitz property of d_P^K , when considering a specific level set at r

- a point x such that $d_P^K(x) \leq r - \varepsilon$ is no further than ε from some $g \in G$ such that $d_P^K(g) \leq r$, and
- every ball $B_\varepsilon(x)$ centered at some point $x \in \mathbb{R}^d$ of radius ε so that all $y \in B_\varepsilon(x)$ has $d_P^K(y) \leq r$ has some representative point $g \in G_{\varepsilon, d}$ such that $g \in B_\varepsilon(x)$, and hence $d_P^K(g) \leq r$.

Thus “deep” regions and spatially thick features are preserved, however thin passageways or layers that are near the threshold r , even if they do not correspond to a critical point, may erroneously become disconnected, causing phantom components or other topological features. However, due to the Lipschitz property, these can be different from r by at most ε , so the errors will have small deviation in persistence.

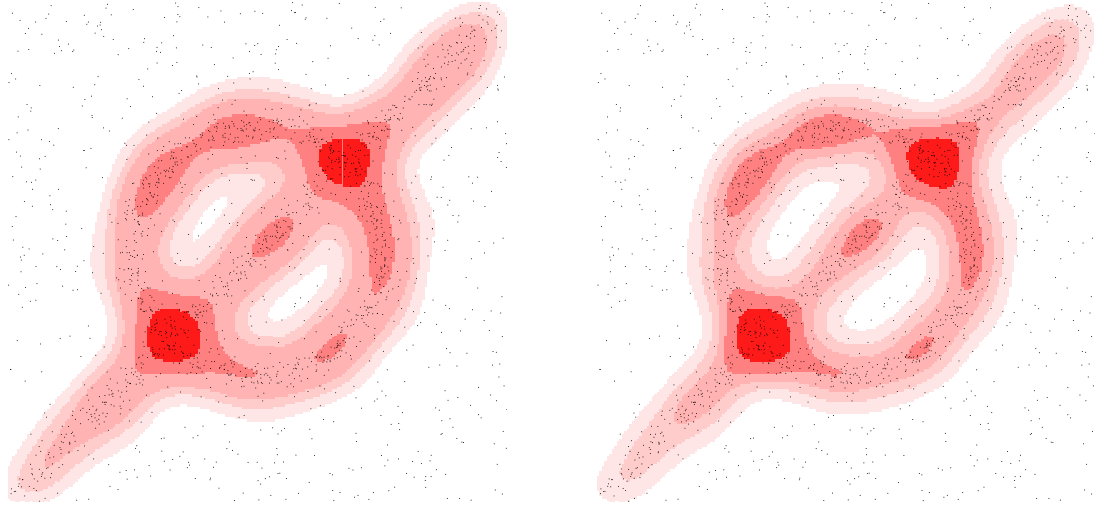


Figure 3: Sublevel sets for the kernel distance while varying the isolevel γ , for fixed σ (left) and for fixed isolevel γ but variable σ (right), with Gaussian kernel. The variable values of σ and γ are chosen to make the plots similar.

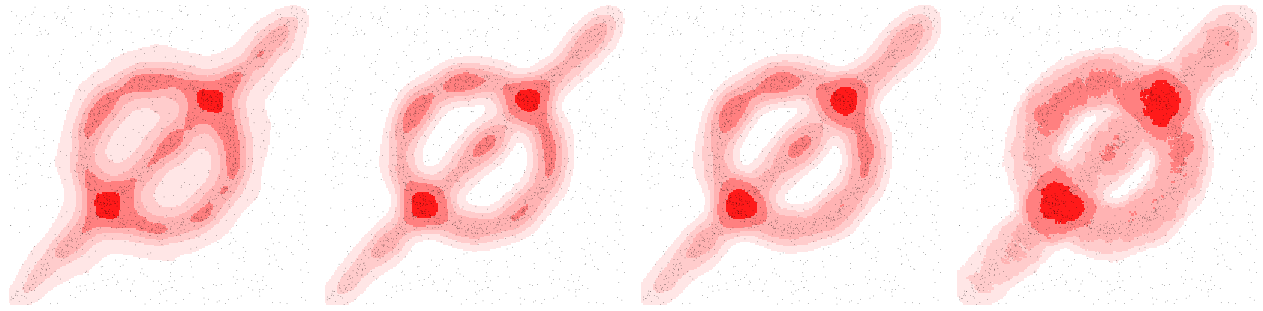


Figure 4: Alternate kernel density estimates for the same dataset as Figure 3. From left to right, they use the Laplace, triangle, Epanechnikov, and the ball kernel.

Varying parameter r or σ . We demonstrate the geometric inference on a synthetic dataset in $[0, 1]^2$ where 900 points are chosen near a circle centered at $(0.5, 0.5)$ with radius 0.25 or along a line segment from $(0, 0)$ to $(1, 1)$. Each point has Gaussian noise added with standard deviation 0.01. The remaining 1100 points are chosen uniformly from $[0, 1]^2$. We use a Gaussian kernel with $\sigma = 0.05$. Figure 3 shows (left) various sublevel sets $\gamma \in \Gamma$ for the kernel distance at a fixed $\sigma = 0.05$ and (right) various superlevel sets for a fixed $\gamma = 0.04853$, but various values of $\sigma \in \Sigma$, where

$$\Gamma = [0.05005, 0.04979, 0.04954, 0.04904, 0.04853] \text{ and}$$

$$\Sigma = [0.0485, 0.0489, 0.0492, 0.0495, 0.05].$$

This choice of Γ and Σ were made to highlight how similar the isolevels can be.

Alternative kernels. We can choose kernels other than the Gaussian kernel in the kernel density estimate, for instance

- the Laplace kernel $K(p, x) = \exp(-2\|x - y\|/\sigma)$,

- the triangle kernel $K(p, x) = \max\{0, 1 - \|x - y\|/\sigma\}$,
- the Epanechnikov kernel $K(p, x) = \max\{0, 1 - \|x - y\|^2/\sigma^2\}$, or
- the ball kernel ($K(p, x) = \{1 \text{ if } \|p - x\| \leq \sigma; \text{ o.w. } 0\}$).

Figure 4 chooses parameters to make them comparable to the Figure 3(left). Of these only the Laplace kernel is *characteristic* [66] making the corresponding version of the kernel distance a metric. Investigating which of the above reconstruction theorems hold when using the Laplace or other kernels is an interesting question for future work.

Additionally, normal vector information and even k -forms can be used in the definition of a kernel [42, 68, 30, 29, 43, 48]; this variant is known as the *current distance*. In some cases it retains its metric properties and has been shown to be very useful for shape alignment in conjunction with medical imaging.

7.3 Open Questions

This work shows it is possible to prove formal reconstruction results using kernel density estimates and the kernel distance. But it also opens many interesting questions.

- For what other types of kernels can we show reconstruction bounds? The Laplace and triangle kernels are natural choices. For both the coresets results match those of the Gaussian kernel. The kernel distance under the Laplace kernel is also a metric, but is not known to be for the triangle kernel. Yet, the triangle kernel would be interesting since it has bounded support, and may lend itself to easier computation.
- The power distance construction in Section 3 requires a point \hat{p}_+ , which approximates the point with minimum kernel distance. This is intuitively because it is possible to construct a point set P (say points lying on a circle with no points inside) such that the point $p_+ \in \mathbb{R}^d$ which minimizes the kernel distance and maximizes the kernel density estimate is far from any point in the point set. For one, can \hat{p}_+ be constructed efficiently without dependence on β_P or Λ_P/σ ?

But more interestingly, can we generally approximate the persistence diagram without creating a simplicial complex on a subset of the input points? We do describe some bounds on using a grid-based technique in Section 7.2, but this is also unsatisfying since it essentially requires a low-dimensional Euclidean space.

- Since d_μ^K is Lipschitz in x and σ , it may make sense to understand the simultaneous stability of both variables. What is the best way to understand persistence over both parameters?
- We provided some initial bound comparing the kernel distance under the Gaussian kernel and the Wasserstein 2 distance. Can we show that under our choice of normalization that $D_K(\mu, \nu) \leq W_2(\mu, \nu)$, unconstrained? More generally, how does the kernel distance under other kernels compare with other forms of Wasserstein and other distances on measures?

Acknowledgements The authors thank Don Sheehy, Frédéric Chazal and the rest of the Geometrica group at INRIA-Saclay for enlightening discussions on geometric and topological reconstruction. We also thank Don Sheehy for personal communications regarding the power distance constructions, and Yusu Wang for ideas towards Lemma 4.1. Finally, we are also indebted to the anonymous reviewers for many detailed suggestions leading to improvements in results and presentation.

References

- [1] Pankaj K. Agarwal, Sariel Har-Peled, Hiam Kaplan, and Micha Sharir. Union of random minkowski sums and network vulnerability analysis. In *Proceedings of 29th Symposium on Computational Geometry*, 2013.
- [2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.
- [3] Sivaraman Balakrishnan, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical inference for persistent homology. Technical report, ArXiv:1303.7117, March 2013.
- [4] James Biagioni and Jakob Eriksson. Map inference in the face of noise and disparity. In *ACM SIGSPATIAL GIS*, 2012.
- [5] 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.
- [6] Omer Bobrowski, Sayan Mukherjee, and Jonathan E. Taylor. Topological consistency via kernel estimation. Technical report, arXiv:1407.5272, 2014.
- [7] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 2014.
- [8] Mickael Buchet, Frederic Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust topological data analysis on metric spaces. arXiv:1306.0039, 2013.
- [9] Gunnar Carlsson, Gurjeet Singh, and Afra Zomorodian. Computing multidimensional persistence. *Algorithms and Computation: Lecture Notes in Computer Science*, 5878:730–739, 2009.
- [10] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Proc. 23rd Ann. Symp. Computational Geometry*, pages 184–193, 2007.
- [11] Frédéric Chazal and David Cohen-Steiner. Geometric inference. *Tessellations in the Sciences*, 2012.
- [12] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings 25th Annual Symposium on Computational Geometry*, pages 237–246, 2009.
- [13] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. Normal cone approximation and offset shape isotopy. *Computational Geometry: Theory and Applications*, 42:566–581, 2009.
- [14] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. *Discrete Computational Geometry*, 41(3):461–479, 2009.
- [15] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.
- [16] Frederic Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. arXiv:1207.3674, 2013.

- [17] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance-to-a-measure and kernel distance. Technical report, arXiv:1412.7197, 2014.
- [18] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems*, 20:96–105, 2013.
- [19] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes. In *Proceedings Symposium on Computational Geometry*, 2014.
- [20] Frédéric Chazal and André Lieutier. Weak feature size and persistent homology: computing homology of solids in \mathbb{R}^n from noisy data samples. *Proceedings 21st Annual Symposium on Computational Geometry*, pages 255–262, 2005.
- [21] Frédéric Chazal and André Lieutier. Topology guaranteeing manifold reconstruction using distance function to noisy data. *Proceedings 22nd Annual Symposium on Computational Geometry*, pages 112–118, 2006.
- [22] Frédéric Chazal and Steve Oudot. Towards persistence-based reconstruction in euclidean spaces. *Proceedings 24th Annual Symposium on Computational Geometry*, pages 232–241, 2008.
- [23] Scott Cohen. Finding color and shape patterns in images. Technical report, Stanford University: CS-TR-99-1620, 1999.
- [24] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37:103–120, 2007.
- [25] Hal Daumé. From zero to reproducing kernel hilbert spaces in twelve pages or less. <http://pub.hal3.name/daume04rkhs.ps>, 2004.
- [26] Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, 1984.
- [27] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer-Verlag, 2001.
- [28] Tamal K. Dey. *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis*. Cambridge University Press, 2007.
- [29] Stanley Durrleman, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Measuring brain variability via sulcal lines registration: A diffeomorphic approach. In *10th International Conference on Medical Image Computing and Computer Assisted Intervention*, 2007.
- [30] Stanley Durrleman, Xavier Pennec, Alain Trouvé, and Nicholas Ayache. Sparse approximation of currents for statistics on curves and surfaces. In *11th International Conference on Medical Image Computing and Computer Assisted Intervention*, 2008.
- [31] Herbert Edelsbrunner. The union of balls and its dual shape. *Proceedings 9th Annual Symposium on Computational Geometry*, pages 218–231, 1993.
- [32] Herbert Edelsbrunner, Michael Facello, Ping Fu, and Jie Liang. Measuring proteins and voids in proteins. In *Proceedings 28th Annual Hawaii International Conference on Systems Science*, 1995.

- [33] Herbert Edelsbrunner, Brittany Terese Fasy, and Günter Rote. Add isotropic Gaussian kernels at own risk: More and more resilient modes in higher dimensions. *Proceedings 28th Annual Symposium on Computational Geometry*, pages 91–100, 2012.
- [34] Herbert Edelsbrunner and John Harer. Persistent homology - a survey. *Contemporary Mathematics*, 453:257–282, 2008.
- [35] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, USA, 2010.
- [36] Herbert Edelsbrunner, David Letscher, and Afra J. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.
- [37] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proc. IEEE*, 90:1151–1163, 2002.
- [38] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the R package TDA. Technical report, arXiv:1411.1830, 2014.
- [39] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Statistical inference for persistent homology: Confidence sets for persistence diagrams. In *The Annals of Statistics*, volume 42, pages 2301–2339, 2014.
- [40] H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93:418–491, 1959.
- [41] Mingchen Gao, Chao Chen, Shaoting Zhang, Zhen Qian, Dimitris Metaxas, and Leon Axel. Segmenting the papillary muscles and the trabeculae from high resolution cardiac CT through restoration of topological handles. In *Proceedings International Conference on Information Processing in Medical Imaging*, 2013.
- [42] Joan Glaunès. *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l’anatomie numérique*. PhD thesis, Université Paris 13, 2005.
- [43] Joan Glaunès and Sarang Joshi. Template estimation from unlabeled point set data and surfaces for computational anatomy. In *International Workshop on Mathematical Foundations of Computational Anatomy*, 2006.
- [44] Karsten Grove. Critical point theory for distance functions. *Proceedings of Symposia in Pure Mathematics*, 54:357–385, 1993.
- [45] Leonidas Guibas, Quentin Mérigot, and Dmitriy Morozov. Witnessed k -distance. *Proceedings 27th Annual Symposium on Computational Geometry*, pages 57–64, 2011.
- [46] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.
- [47] Matrial Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *Proceedings 10th International Workshop on Artificial Intelligence and Statistics*, 2005.
- [48] Sarang Joshi, Raj Varma Kommaraju, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. *Proceedings 27th Annual Symposium on Computational Geometry*, 2011.

- [49] A. N. Kolmogorov, S. V. Fomin, and R. A. Silverman. *Introductory Real Analysis*. Dover Publications, 1975.
- [50] John M. Lee. *Introduction to topological manifolds*. Springer-Verlag, 2000.
- [51] John M. Lee. *Introduction to smooth manifolds*. Springer, 2003.
- [52] Jie Liang, Herbert Edelsbrunner, Ping Fu, Pamidighantam V. Sudhakar, and Shankar Subramanian. Analytic shape computation of macromolecules: I. molecular area and volume through alpha shape. *Proteins: Structure, Function, and Genetics*, 33:1–17, 1998.
- [53] André Lieutier. Any open bounded subset of \mathbb{R}^n has the same homotopy type as its medial axis. *Computer-Aided Design*, 36:1029–1046, 2004.
- [54] Quentin Mérigot. *Geometric structure detection in point clouds*. PhD thesis, Université de Nice Sophia-Antipolis, 2010.
- [55] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12), 2011.
- [56] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
- [57] Jeff M. Phillips. eps-samples for kernels. *Proceedings 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.
- [58] Jeff M. Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. arXiv:1103.1625, March 2011.
- [59] Florian T. Pokorny, Carl Henrik, Hedvig Kjellström, and Danica Kragic. Persistent homology for learning densities with bounded support. In *Neural Information Processing Systems*, 2012.
- [60] Charles A. Price, Olga Symonova, Yuriy Mileyko, Troy Hilley, and Joshua W. Weitz. Leaf gui: Segmenting and analyzing the structure of leaf veins and areoles. *Plant Physiology*, 155:236–245, 2011.
- [61] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.
- [62] Donald R. Sheehy. A multicover nerve for geometric inference. *Canadian Conference in Computational Geometry*, 2012.
- [63] Donald R. Sheehy. Linear-size approximations to the Vietoris-Rips filtration. *Discrete & Computational Geometry*, 49:778–796, 2013.
- [64] Bernard W. Silverman. Using kernel density estimates to investigate multimodality. *J. R. Statistical Society B*, 43:97–99, 1981.
- [65] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.
- [66] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
- [67] Kathryn Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 2014.

- [68] Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Proceedings Information Processing in Medical Imaging*, volume 19, pages 381–92, 2005.
- [69] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.
- [70] Grace Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomization GACV. In *Advances in Kernel Methods – Support Vector Learning*, pages 69–88. Bernhard Schölkopf and Alexander J. Smola and Christopher J. C. Burges and Rosanna Soentpiet, 1999.
- [71] Yan Zheng, Jeffrey Jests, Jeff M. Phillips, and Feifei Li. Quality and efficiency in kernel density estimates for large data. In *Proceedings ACM Conference on the Management of Data (SIGMOD)*, 2012.

A Details on Distance-Like Properties of Kernel Distance

We provide further details on distance-like properties of the kernel distance.

A.1 Semiconcave Properties of Kernel Distance

We also note that semiconcavity follows quite naturally and simply in the RKHS \mathcal{H}_K for d_μ^K .

Lemma A.1. $(d_\mu^K)^2$ is 1-semiconcave in \mathcal{H}_K : the map $x \mapsto (d_\mu^K(x))^2 - \|\phi(x)\|_{\mathcal{H}_K}^2$ is concave.

Proof. We can write

$$(d_\mu^K(x))^2 = (D_K(\mu, x))^2 = \kappa(\mu, \mu) + \kappa(x, x) - 2\kappa(\mu, x) = \|\Phi(\mu)\|_{\mathcal{H}_K}^2 + \|\phi(x)\|_{\mathcal{H}_K}^2 - 2\|\Phi(\mu) - \phi(x)\|_{\mathcal{H}_K}^2.$$

Now

$$(d_\mu^K(x))^2 - \|\phi(x)\|_{\mathcal{H}_K}^2 = \|\Phi(\mu)\|_{\mathcal{H}_K}^2 - 2\|\Phi(\mu) - \phi(x)\|_{\mathcal{H}_K}^2.$$

Since the above is twice-differentiable, we only need to show that its twice-differential is non-positive. By definition, for a fixed μ , $\Phi(\mu)$ and $\|\Phi(\mu)\|_{\mathcal{H}_K}^2$ are both constant. Suppose $\Phi(\mu) = c_1$ and $\|\Phi(\mu)\|_{\mathcal{H}_K}^2 = c_2$, we have $(d_\mu^K(x))^2 - \|\phi(x)\|_{\mathcal{H}_K}^2 = c_2 - \|c_1 - \phi(x)\|_{\mathcal{H}_K}^2$. Since the RKHS \mathcal{H}_K is a vector space with well-defined norm $\|\cdot\|_{\mathcal{H}_K}$, the above is a concave parabolic function. \square

However, this semiconcavity in \mathcal{H}_K is not that useful. For unit weight elements $x, y \in \mathbb{R}^d$, an element s_α such that $\phi(s_\alpha) = \alpha\phi(y) + (1 - \alpha)\phi(x)$ is a weighted point set with a point at x with weight $(1 - \alpha)$ and another at y with weight α . Lemma A.1 only implies that $(d_K(s_\alpha))^2 - \|\phi(s_\alpha)\|_{\mathcal{H}_K}^2 \leq \alpha((d_K(x))^2 - \|\phi(x)\|_{\mathcal{H}_K}^2) + (1 - \alpha)((d_K(y))^2 - \|\phi(y)\|_{\mathcal{H}_K}^2)$.

A.2 Kernel Distance is Proper

We use two more general, but equivalent definitions of a proper map. Definition (i): A continuous map $f : \mathbb{X} \rightarrow \mathbb{Y}$ between two topological spaces is *proper* if and only if the inverse image of every compact subset in \mathbb{Y} is compact in \mathbb{X} ([50], page 84; [51], page 45). Definition (ii): a continuous map $f : \mathbb{X} \rightarrow \mathbb{Y}$ between two topological manifolds is proper if and only if for every sequence $\{p_i\}$ in \mathbb{X} that escapes to infinity, $\{f(p_i)\}$ escapes to infinity in \mathbb{Y} ([51], Proposition 2.17). Here, for a topological space \mathbb{X} , a sequence $\{p_i\}$ in \mathbb{X} *escapes to infinity* if for every compact set $G \subset \mathbb{X}$, there are at most finitely many values of i for which $p_i \in G$ ([51], page 46).

Lemma A.2 (Lemma 2.4). d_μ^K is proper.

Proof. To prove that d_μ^K is proper, we prove the following two claims: (a) A continuous function $f : \mathbb{R}^d \rightarrow [0, c)$ (where c is a constant) is proper, if for any sequence $\{x_i\}$ in \mathbb{R}^d that escapes to infinity, the sequence $\{f(x_i)\}$ tends to c (approaches c in the limit); (b) Let $f := d_\mu^K$ and one needs to show that for any sequences $\{x_i\}$ that escapes to infinity, the sequence $\{f(x_i)\}$ tends to c_μ ; or equivalently, $\kappa(\mu, x_i)$ tends to 0.

We prove claim (a) by proving its contrapositive. If a continuous function $f : \mathbb{R}^d \rightarrow [0, c)$ is not proper, then there exists a sequence $\{x_i\}$ in \mathbb{R}^d that escapes to infinity, such that the sequence $\{f(x_i)\}$ does not tend to c . Suppose f is not proper, this implies that there exists a constant $b < c$ such that $f^{-1}[0, b]$ is not compact (based on properness definition (i)) and therefore either not closed or unbounded. We first show that $A := f^{-1}[0, b]$ is closed. We make use of the following theorem ([49], page 88, Theorem 10'): A mapping f of a topological space \mathbb{X} into a topological space \mathbb{Y} is continuous if and only if the pre-image $f^{-1}(F)$ of every closed set $F \subset \mathbb{Y}$ is closed in \mathbb{X} . Since f is continuous, it implies that the pre-image of every closed set $[a, b] \subset \mathbb{R}$ is closed in \mathbb{R}^d . Therefore, A is closed, therefore it must be unbounded.

Since every unbounded sequence contains a monotone subsequence that has either $+\infty$ or $-\infty$ as a limit, therefore A contains a subsequence $S := \{x_i\}$ that tends to an infinite limit. In addition, as elements in S escapes to infinity, $\{f(x_i)\}$ tends to b and does not tend to c . Therefore (a) holds by contraposition.

To prove claim (b), we need to show that for any sequence $\{x_i\}$ that escapes to infinity, $\kappa(\mu, x_i)$ tends to 0. For each x_i , define a radius $r_i = \|x_i - 0\|/2$ and define a ball B_i that is centered at the origin 0 and has radius r_i . As x_i goes to infinity, r_i increases until for any fixed arbitrary $\varepsilon > 0$, we have $\int_{p \in B_i} \mu(p) dp \geq 1 - \varepsilon/2\sigma^2$ and thus $\int_{p \in \mathbb{R}^d \setminus B_i} d\mu(p) \leq \varepsilon/2\sigma^2$. Furthermore, let $p_i = \arg \min_{p \in B} \|p - x_i\|$, so $\|x_i - p_i\| = r_i$. Thus also as x_i goes to infinity, r_i increases until for any $\varepsilon > 0$ we have $K(p_i, x_i) \leq \varepsilon/2$. We now decompose $\kappa(\mu, x_i) = \int_{p \in B_i} K(p, x_i) d\mu(p) + \int_{q \in \mathbb{R}^d \setminus B_i} K(q, x_i) d\mu(q)$. Thus for any $\varepsilon > 0$, as x_i goes to infinity, the first term is at most $\varepsilon/2$ since all $K(p, x_i) \leq K(p_i, x_i) \leq \varepsilon/2$ and the second term is at most $\varepsilon/2$ since $K(q, x) \leq \sigma^2$ and $\int_{q \in \mathbb{R}^d \setminus B_i} \mu(q) dq \leq \varepsilon/2\sigma^2$. Since these results hold for all ε , as x_i goes to infinity and ε goes to 0, $\kappa(\mu, x_i)$ goes to 0.

Combine (a) with (b) and the fact that d_μ^K is a continuous (in fact, Lipschitz) function, we obtained the properness result. \square

B ε -Approximation of the Kernel Distance

Here we make explicit the way that an ε -kernel sample approximated the kernel distance. Recall that if Q is an ε -kernel sample of μ , then $\|\text{KDE}_\mu - \text{KDE}_{\mu_Q}\| = \max_{x \in \mathbb{R}^d} |\kappa(\mu, x) - \kappa(\mu_Q, x)| \leq \varepsilon$.

Lemma B.1. *If Q is an ε -kernel sample of μ , then $\|(d_\mu^K)^2 - (d_{\mu_Q}^K)^2\|_\infty \leq 4\varepsilon$.*

Proof. First expand $D_K(\mu, x)^2 = \kappa(x, x) + \kappa(\mu, \mu) - 2\kappa(\mu, x) = \sigma^2 + \kappa(\mu, \mu) - 2\kappa(\mu, x)$. Replacing μ with μ_Q , the first term is unaffected. The second term is bounded,

$$\begin{aligned} \kappa(\mu, \mu) &= \int_{(p,q)} K(p, q) d\mathbf{m}_{\mu, \mu}(p, q) = \int_p \left(\int_q K(p, q) d\mu(q) \right) d\mu(p) \\ &= \int_p \text{KDE}_\mu(p) d\mu(p) \leq \int_p (\text{KDE}_{\mu_Q}(p) + \varepsilon) d\mu(p) \\ &= \int_p \text{KDE}_{\mu_Q}(p) d\mu(p) + \varepsilon = \int_p \left(\int_q K(p, q) d\mu_Q(q) \right) d\mu(p) + \varepsilon \\ &= \kappa(\mu_Q, \mu) + \varepsilon \\ &\leq \kappa(\mu_Q, \mu_Q) + 2\varepsilon. \end{aligned}$$

Similar results hold by switching μ_Q with μ in the above inequality, that is, $\kappa(\mu_Q, \mu_Q) \leq \kappa(\mu, \mu) + 2\varepsilon$. And for the third term we have similar inequality, $|2\kappa(\mu, x) - 2\kappa(\mu_Q, x)| \leq 2\varepsilon$. Combining all three terms, we have the desired result: $|D_K(\mu, x)^2 - D_K(\mu_Q, x)^2| \leq 4\varepsilon$. \square

Lemma B.2. *If Q is an $(\varepsilon^2/4)$ -kernel sample of μ , then $\|d_\mu^K - d_{\mu_Q}^K\|_\infty \leq \varepsilon$.*

Proof. By Lemma B.1 this condition on Q implies that $\|(d_\mu^K)^2 - (d_{\mu_Q}^K)^2\|_\infty \leq \varepsilon^2$. We then use a basic fact for values $\varepsilon \geq 0$ and $\gamma \geq 0$.

- $\sqrt{\gamma^2 + \varepsilon^2} \leq \gamma + \varepsilon$. This follows since $(\gamma + \varepsilon)^2 = \gamma^2 + \varepsilon^2 + 2\gamma\varepsilon \geq \gamma^2 + \varepsilon^2$.

We now prove the main result as an upper and lower bound using for any $x \in \mathbb{R}^d$. We first use $\gamma = d_\mu^K(x) \geq 0$ and expand $d_{\mu_Q}^K(x)$ to obtain

$$d_{\mu_Q}^K(x) = \sqrt{(d_{\mu_Q}^K(x))^2} \leq \sqrt{(d_\mu^K(x))^2 + \varepsilon^2} \leq d_\mu^K(x) + \varepsilon.$$

Now we use $\gamma = d_{\mu_Q}^K(x) \geq 0$ and expand $d_{\mu}^K(x)$ to obtain

$$d_{\mu}^K(x) = \sqrt{(d_{\mu}^K(x))^2} \leq \sqrt{(d_{\mu_Q}^K(x))^2 + \varepsilon^2} \leq d_{\mu_Q}^K(x) + \varepsilon.$$

Hence for any $x \in \mathbb{R}^d$ we have $d_{\mu}^K(x) - \varepsilon \leq d_{\mu_Q}^K(x) \leq d_{\mu}^K(x) + \varepsilon$. \square

C Power Distance Constructions

Recall we want to consider the following *power distance* using d_{μ}^K (as weight) for a measure μ associated with a subset $P \subset \mathbb{R}^d$ and metric $d(\cdot, \cdot)$ on \mathbb{R}^d ,

$$f_P(\mu, x) = \sqrt{\min_{p \in P} (d(p, x)^2 + d_{\mu}^K(p)^2)}.$$

We consider a particular choice of the distance metric $d(p, x) = D_K(p, x)$ which leads to a kernel version of the power distance

$$f_P^K(\mu, x) = \sqrt{\min_{p \in P} (D_K(p, x)^2 + d_{\mu}^K(p)^2)}.$$

Recall that $d_{\mu}^K(x) = D_K(\mu, x)$. In this section, we will always use the notation $D_K(\mu, \nu)$, and when μ or ν are points (e.g. μ is a Dirac mass at p and ν is a Dirac mass at q), then we will just write $D_K(p, q)$. This will be especially helpful when we apply the triangle inequality in several places.

C.1 Kernel Power Distance on Point Set P

Given a set P defining a measure of interest μ_P , it is of interest to consider if $f_P^K(\mu_P, x)$ is multiplicatively bounded by $D_K(\mu_P, x)$. Theorem 3.2 shows that the lower bound holds. In this section we try to provide a multiplicative approximation upper bound.

Let $p^* = \arg \min_{p \in P} \|p - x\|$. We can start with Lemma 3.1 which reduces the problem finding a multiplicative upper bound for $D_K(p^*, x)$ in terms of $D_K(\mu_P, x)$. However, we are not able to provide very useful bounds, and they require more advanced techniques than the previous section. In particular, they will only apply for points $x \in \mathbb{R}^d$ when $D_K(\mu_P, x)$ is large enough; hence not well-approximating the minima of d_{μ}^K .

For simplicity, we write $d_P^K(\cdot) = D_K(\mu_P, \cdot)$ as $D_K(P, \cdot)$.

The difficult case is when $D_K(P, x)$ is very small, and hence $\kappa(P, P)$ is very small. So we start by developing tools to upper bound $\kappa(P, P)$ using $\hat{p} = \arg \min_{p \in P} D_K(P, p)$, a point which only provides a worse approximation than p^* .

We first provide a general result in a Hilbert space (a refinement of a vector space [25]), and then next apply it to our setting in the RKHS.

Lemma C.1. *Consider a set $V = \{v_1, \dots, v_n\}$ of vectors in a Hilbert space endowed with norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. Let each v_i have norm $\|v_i\| = \eta$. Consider weights $W = \{w_1, \dots, w_n\}$ such that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. Let $r = \sum_{i=1}^n w_i v_i$. Let $\hat{v} = \arg \min_{v_i \in V} \|v_i - r\|$. Then*

$$\|r\|^2 \leq \eta^2 - \|r - \hat{v}\|^2.$$

Proof. Recall elementary properties of inner product space: $\|x\|^2 = \langle x, x \rangle$, $\langle ax, y \rangle = a \langle x, y \rangle$, $\langle x - y, x - y \rangle = \langle x, x \rangle + \langle y, y \rangle - 2 \langle x, y \rangle$. By definition of \hat{v} , for any $v_i \in V$,

$$\|v_i - r\|^2 \geq \|\hat{v} - r\|^2 \Rightarrow \langle v_i, v_i \rangle + \langle r, r \rangle - 2 \langle v_i, r \rangle \geq \langle \hat{v}, \hat{v} \rangle + \langle r, r \rangle - 2 \langle \hat{v}, r \rangle \Rightarrow \langle v_i, r \rangle \leq \langle \hat{v}, r \rangle.$$

We can decompose r (based on linearity of an inner product space) as

$$\|r\|^2 = \langle r, r \rangle = \sum_{i=1}^n w_i \langle v_i, r \rangle \leq \sum_{i=1}^n w_i \langle \hat{v}, r \rangle = \langle \hat{v}, r \rangle = \frac{1}{2}(\|r\|^2 + \|\hat{v}\|^2 - \|\hat{v} - r\|^2).$$

The last inequality holds by $\|\hat{v} - r\|^2 = \|r\|^2 + \|\hat{v}\|^2 - 2\langle \hat{v}, r \rangle$. Then since $\|\hat{v}\| = \eta$ we can solve for $\|r\|^2$ as

$$\|r\|^2 \leq \eta^2 - \|\hat{v} - r\|^2. \quad \square$$

Lemma C.2. *Let $\hat{p} = \arg \min_{p \in P} D_K(P, p)$, then $\kappa(P, P) \leq \sigma^2 - D_K(P, \hat{p})^2$.*

Proof. Let $\phi_K : \mathbb{R}^d \rightarrow \mathcal{H}_K$ map points in \mathbb{R}^d to the reproducing kernel Hilbert space (RKHS) \mathcal{H}_K defined by kernel K . This space has norm $\|P\|_{\mathcal{H}_K} = \sqrt{\kappa(P, P)}$ defined on a set of points P and inner product $\kappa(P, P)$. Let $\Phi_K(P) = \frac{1}{|P|} \sum_{p \in P} \phi_K(p)$ be the representation of a set of points P in \mathcal{H}_K . Note that $D_K(P, Q) = \|\Phi_K(P) - \Phi_K(Q)\|_{\mathcal{H}_K}$. We can now apply Lemma C.1 to $\{\phi_K(p)\}_{p \in P}$ with weights $w(p) = 1/|P|$ and $r = \Phi_K(P)$, and norm $\eta = \sigma$. Hence $\kappa(P, P) = \|P\|_{\mathcal{H}_K}^2 \leq \sigma^2 - D_K(P, \hat{p})^2$. \square

Lemma C.3. *For any $s > 0$ and any x , then $\sqrt{s^2 - x} \leq s - x/2s$.*

Proof. We expand the square of the desired result

$$(s^2 - x) \leq (s - x/2s)^2 = s^2 - x + x^2/4s^2.$$

After subtracting $(s^2 - x)$ from both sides, it is equivalent to $0 \leq x^2/4s^2$. This holds since x^2 and s are always nonnegative. \square

Lemma C.4. $D_K(P, x) \geq D_K(p^*, x)^2/C_\sigma$ for $C_\sigma = 2\sigma + 2$.

Proof. Refer to Figure 5 for geometric intuition in this proof. Let ν_0 be a measure that is $\nu_0(p) = 0$ for all $p \in \mathbb{R}^d$; thus it has a norm $\kappa(\nu_0, \nu_0) = 0$. We can measure the distance from ν_0 to x and P , noting that $D_K(\nu_0, x) = \sqrt{\kappa(x, x)} = \sigma$ and $D_K(\nu_0, P) = \sqrt{\kappa(P, P)}$. Thus by triangle inequality, Lemma C.2, and Lemma C.3,

$$\begin{aligned} D_K(P, x) &\geq D_K(\nu_0, x) - D_K(\nu_0, P) \\ &= \sigma - \sqrt{\kappa(P, P)} \\ &\geq \sigma - \sqrt{\sigma^2 - D_K(P, \hat{p})^2} \\ &\geq D_K(P, \hat{p})^2/2\sigma. \end{aligned}$$

We now assume that $D_K(P, x) < D_K(p^*, x)/C_\sigma$ and show this is not possible. First observe that $D_K(P, \hat{p}) + D_K(P, x) \geq D_K(\hat{p}, x) \geq D_K(p^*, x)$. These expressions imply that $D_K(P, \hat{p}) \geq D_K(p^*, x) - D_K(P, x) \geq (1 - 1/C_\sigma)D_K(p^*, x)$, and thus

$$D_K(P, x) \geq \frac{1}{2\sigma} D_K(P, \hat{p})^2 \geq \frac{1}{2\sigma} \left(1 - \frac{1}{C_\sigma}\right)^2 D_K(p^*, x)^2 \geq \frac{1}{C_\sigma} D_K(p^*, x)^2,$$

a contradiction. The last steps follows by setting

$$\frac{1}{2\sigma} \left(1 - \frac{1}{C_\sigma}\right)^2 \geq \frac{1}{C_\sigma} \Rightarrow C_\sigma^2 - (2 + 2\sigma)C_\sigma + 1 \geq 0$$

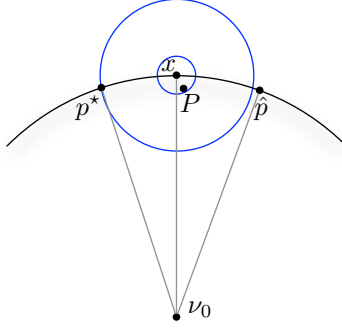


Figure 5: Illustration of x , p^* , \hat{p} , ν_0 , and P as vectors in a RKHS. Note we have omitted the ϕ_K and Φ_K maps to unclutter the notation.

and solving for C_σ ,

$$C_\sigma \geq \frac{(2 + 2\sigma) + \sqrt{(2 + 2\sigma)^2 - 4}}{2} = 1 + \sigma + \sqrt{\sigma^2 + 2\sigma} = 1 + \sigma + \sqrt{(\sigma + 1)^2 - 1}.$$

Since $C_\sigma = 2\sigma + 2 > 1 + \sigma + \sqrt{(\sigma + 1)^2 - 1}$, so we have $\frac{1}{2\sigma} \left(1 - \frac{1}{C_\sigma}\right)^2 \geq \frac{1}{C_\sigma}$. \square

Recall that an ε -kernel sample P of μ satisfies $\max_{x \in \mathbb{R}^d} |\kappa(\mu, x) - \kappa(\mu_P, x)| \leq \varepsilon$.

Theorem C.1. *If $D_K(P, x) \geq 1$ then $f_P^K(P, x) \leq \sqrt{6\sigma + 8}D_K(P, x)$. If P is an $(\varepsilon/4)$ -kernel sample of μ then $f_P^K(\mu, x) \leq \sqrt{6\sigma + 8}(D_K(\mu, x) + \varepsilon)$.*

Proof. We combine Lemma C.4 with Lemma 3.1 to achieve

$$f_P^K(P, x)^2 \leq 2D_K(P, x)^2 + 3D_K(p^*, x)^2 \leq 2D_K(P, x)^2 + 3(2\sigma + 2)D_K(P, x).$$

Aside: Note that the first $D_K(P, x)$ is squared and the second is not. If $D_K(P, x) \geq \alpha$ then $D_K(P, x) \leq (1/\alpha)D_K(P, x)^2$ we have

$$f_P^K(P, x)^2 \leq (2 + (6 + 6\sigma)/\alpha)D_K(P, x)^2.$$

Let $\alpha = 1$. We have

$$f_P^K(P, x)^2 \leq (6\sigma + 8)D_K(P, x)^2.$$

Since $D_K(P, x) \leq D_K(\mu, x) + \varepsilon$, via Lemma B.1. We obtain,

$$f_P^K(\mu, x) \leq \sqrt{6\sigma + 8}(D_K(\mu, x) + \varepsilon). \quad \square$$

C.2 Approximating the Minimum Kernel Distance Point

The goal in this section is to find a point that approximately minimizes the kernel distance to a point set P . We assume here P contains n points and describes a measure made of n Dirac mass at each $p \in P$ with weight $1/n$ (this is the empirical measure μ_P defined in Section 1.1). Let $p_+ = \arg \min_{q \in \mathbb{R}^d} D_K(\mu_P, q) = \arg \max_{q \in \mathbb{R}^d} \kappa(\mu_P, q)$. Since $D_K(\mu_P, q) = D_K(P, q)$, for simplicity in notation, we work with point set P instead of μ_P for the remaining of this section. That is, we define $p_+ = \arg \min_{q \in \mathbb{R}^d} D_K(P, q) = \arg \max_{q \in \mathbb{R}^d} \kappa(P, q)$. Note that p_+ is chosen over all of \mathbb{R}^d , as the bound in Theorem C.1 is not sufficient when choosing a point from P . In particular, for any $\delta > 0$, we want a point \hat{p}_+ such that $D_K(P, \hat{p}_+) \leq (1 + \delta)D_K(P, p_+)$.

Note that Agarwal *et al.* [1] provide an algorithm that with high probability finds a point \hat{q} such that $\kappa(P, \hat{q}) \geq (1 - \delta)\kappa(P, p_+)$ in time $O((1/\delta^4)n \log n)$. However this point \hat{q} is *not* sufficient for our purpose (that is, \hat{q} does not satisfy the condition $D_K(P, \hat{q}_+) \leq (1 + \delta)D_K(P, p_+)$), since \hat{q} yields

$$D_K(P, \hat{q})^2 \leq \sigma^2 + \kappa(P, P) - 2(1 - \delta)\kappa(P, p_+) \not\leq (1 + \delta)(\sigma^2 + \kappa(P, P) - 2\kappa(P, p_+)) = (1 - \delta)D_K(P, p_+)^2,$$

since in general it is not true that $4\kappa(P, p_+) \leq \sigma^2 + \kappa(P, P)$, as would be required.

First we need some structural properties. For each point $x \in \mathbb{R}^d$, define a radius $r_x = \arg \sup_{r>0} \{|B_r(x) \cap P| \leq n/2\}$, where $B_r(x)$ is a ball of radius r centered at x . In other words, it is the largest radius such that at most half of points in P are within $B_r(x)$. Let \hat{p}_2 be the point in P such that $\|p_+ - \hat{p}_2\| = r_{p_+}$. In other words, \hat{p}_2 is a point such that no more than $n/2$ points in P satisfy $\|p_+ - p\| \geq \|p_+ - \hat{p}_2\|$. Finally it is useful to define $r_{x,K}$ which is $r_{x,K} = D_K(x, p)$ where $\|x - p\| = r_x$; in particular $r_{p_+,K} = D_K(p_+, \hat{p}_2)$.

We now need to lower bound $D_K(P, p_+)$ in terms of $D_K(P, \hat{p}_2)$. Lemma C.4 already provides a bound in terms of the closest point for any $x \in \mathbb{R}^d$. We follow a similar construction here.

Lemma C.5. *Consider a set $V = \{v_1, \dots, v_n\}$ of vectors in a Hilbert space endowed with norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. Let each v_i have norm $\|v_i\| = \eta$. Consider weights $W = \{w_1, \dots, w_n\}$ such that $1/2 \geq w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. Let $r = \sum_{i=1}^n w_i v_i$. Define a partition of V with V_1 and V_2 such that V_2 is the smallest set such that $\sum_{v_i \in V_2} w_i \geq 1/2$, and for all $v_1 \in V_1$ and $v_2 \in V_2$ we have $\|r - v_1\| < \|r - v_2\|$. Let $\hat{v}_2 = \arg \min_{v_i \in V_2} \|v_i - r\|$. Then*

$$\|r\|^2 \leq \eta^2 - \frac{\|r - \hat{v}_2\|^2}{2}.$$

Proof. For ease of notation, we assume that $\langle v_i, r \rangle > \langle v_{i+1}, r \rangle$ for all i , and let $\{v_1, \dots, v_k\} = V_1$. Let $\hat{v}_1 = \arg \min_{v_i \in V_1} \|v_i - r\| = \arg \min_{v_i \in V} \|v_i - r\|$. Let \hat{v} be a norm η vector that has $\langle \hat{v}, r \rangle = (\langle \hat{v}_1, r \rangle + \langle \hat{v}_2, r \rangle)/2$. Since $\sum_{v_i \in V_2} w_i \geq 1/2$ and $\sum_{v_i \in V_1} w_i \leq 1/2$, let $\sum_{v_i \in V_2} w_i = 1/2 + \delta$ and $\sum_{v_i \in V_1} w_i = 1/2 - \delta$ (for $0 \leq \delta \leq 1/2$). By definition, we also have $\langle \hat{v}_1, r \rangle \geq \langle \hat{v}_2, r \rangle$. We can decompose r as

$$\begin{aligned} \|r\|^2 &= \langle r, r \rangle = \sum_{i=1}^n w_i \langle v_i, r \rangle = \sum_{i=1}^k w_i \langle v_i, r \rangle + \sum_{i=k+1}^n w_i \langle v_i, r \rangle \\ &\leq \sum_{i=1}^k w_i \langle \hat{v}_1, r \rangle + \sum_{i=k+1}^n w_i \langle \hat{v}_2, r \rangle = \left(\sum_{i=1}^k w_i \right) \langle \hat{v}_1, r \rangle + \left(\sum_{i=k+1}^n w_i \right) \langle \hat{v}_2, r \rangle \\ &= (1/2 - \delta) \langle \hat{v}_1, r \rangle + (1/2 + \delta) \langle \hat{v}_2, r \rangle = (1/2) (\langle \hat{v}_1, r \rangle + \langle \hat{v}_2, r \rangle) + \delta (\langle \hat{v}_2, r \rangle - \langle \hat{v}_1, r \rangle) \\ &\leq (\langle \hat{v}_1, r \rangle + \langle \hat{v}_2, r \rangle)/2 = \langle \hat{v}, r \rangle \\ &= \frac{1}{2} (\|r\|^2 + \|\hat{v}\|^2 - \|\hat{v} - r\|^2). \end{aligned}$$

The last inequality holds by $\|\hat{v} - r\|^2 = \|r\|^2 + \|\hat{v}\|^2 - 2\langle \hat{v}, r \rangle$. Then since $\|\hat{v}\| = \eta$ we can solve for $\|r\|^2$ as

$$\|r\|^2 \leq \eta^2 - \|\hat{v} - r\|^2 = \eta^2 - (\|\hat{v}_2 - r\|^2 + \|\hat{v}_1 - r\|^2)/2 \leq \eta^2 - \|\hat{v}_2 - r\|^2/2. \quad \square$$

Lemma C.6. *Using \hat{p}_2 as defined above, then $\kappa(P, P) \leq \sigma^2 - D_K(P, \hat{p}_2)^2/2$.*

Proof. Let $\phi_K : \mathbb{R}^d \rightarrow \mathcal{H}_K$ map points in \mathbb{R}^d to the reproducing kernel Hilbert space (RKHS) \mathcal{H}_K defined by kernel K . This space has norm $\|P\|_{\mathcal{H}_K} = \sqrt{\kappa(P, P)}$ defined on a set of points P and inner product $\kappa(P, P)$. Let $\Phi_K(P) = \frac{1}{|P|} \sum_{p \in P} \phi_K(p)$ be the representation of a set of points P in \mathcal{H}_K . Note that $D_K(P, Q) = \|\Phi_K(P) - \Phi_K(Q)\|_{\mathcal{H}_K}$. We can now apply Lemma C.5 to $\{\phi_K(p)\}_{p \in P}$ with weights $w(p) = 1/|P|$ and $r = \Phi_K(P)$, and norm $\eta = \sigma$. Finally note that we can use $\phi_K(\hat{p}_2) = \hat{v}_2$ since V_2 represents the set of points which are further or equal to P than is \hat{p}_2 . In addition, by the property of RKHS, $\|\Phi_K(P) - \phi_K(\hat{p}_2)\| = D_K(P, \hat{p}_2)$. Hence $\kappa(P, P) = \|P\|_{\mathcal{H}_K}^2 \leq \sigma^2 - D_K(P, \hat{p}_2)^2/2$. \square

Lemma C.7. $D_K(P, p_+) \geq D_K(p_+, \hat{p}_2)^2 / (4\sigma)$.

Proof. Refer to Figure 5 for geometric intuition in this proof. Let ν_0 be a measure that is $\nu_0(p) = 0$ for all $p \in \mathbb{R}^d$; thus it has a norm $\kappa(\nu_0, \nu_0) = 0$. We can measure the distance from ν_0 to p_+ and P , noting that $D_K(\nu_0, x) = \sqrt{\kappa(x, x)} = \sigma$ and $D_K(P, \nu_0) = \sqrt{\kappa(P, P)}$. Thus by triangle inequality, Lemma C.6, and Lemma C.3

$$\begin{aligned} D_K(P, p_+) &\geq D_K(\nu_0, p_+) - D_K(P, \nu_0) \\ &= \sigma - \sqrt{\kappa(P, P)} \\ &\geq \sigma - \sqrt{\sigma^2 - D_K(P, \hat{p}_2)^2 / 2} \\ &\geq D_K(P, \hat{p}_2)^2 / 4\sigma. \end{aligned} \quad \square$$

Now we place a net \mathcal{N} on \mathbb{R}^d ; specifically, it is a set of points such that for some $q \in \mathcal{N}$ that $\|q - p_+\| \leq \delta D_K(p_+, \hat{p}_2)^2 / 4\sigma \leq \delta D_K(P, p_+)$ (we refer to this inequality as the *net condition*, therefore, \mathcal{N} is a set of points such that some points in it satisfy the net condition). Since $D_K(P, \cdot)$ is 1-Lipschitz, we have $D_K(P, p_+) - D_K(P, q) \leq \|q - p_+\|$. This ensures that some point $q \in \mathcal{N}$ satisfies $D_K(P, q) \leq (1 + \delta)D_K(P, p_+)$, and can serve as \hat{p}_+ . In other words, \mathcal{N} is guaranteed to contain some point q that can serve as p_+ .

Note that p_+ must be in $\text{CH}(P)$, the convex hull of P . Otherwise, moving to the closest point on $\text{CH}(P)$ decreases the distance to all points, and thus increases $\kappa(P, p_+)$, which cannot happen by definition of p_+ . Let Δ be the diameter of P (the distance between the two furthest points). Clearly for some $p \in P$ we must have $\|p_+ - p\| \leq \Delta$.

Also note that $p_+ := \arg \max_{q \in \mathbb{R}^d} \kappa(P, q)$ must be within a distance $R_\sigma = \sigma \sqrt{2 \ln(n)}$ to some $p \in P$, otherwise for $p^* = \arg \min_{p \in P} \|p_+ - p\|$, we can bound $\kappa(P, p_+) \leq K(p^*, p_+) \leq \sigma^2/n = K(p^*, p^*)/n \leq \kappa(P, p^*)$, which means p_+ is not a maximum. The first inequality is by definition of p^* , the second by assuming $\|p_+ - p^*\| \geq \sigma \sqrt{2 \ln(n)}$.

Let $B_R(p)$ be the ball centered at p with radius $R = \min(R_\sigma, \Delta)$. Let $R_p = \min(R, r_p/2)$. So p_+ must be in $\bigcup_{p \in P} B_{R_p}(p)$. We describe a net \mathcal{N}_p construction for one ball $B_{R_p}(p)$; that is for any x such that $p \in P$ is the closest point to x , then some point $q \in \mathcal{N}_p$ satisfies $\|q - x\| \leq \delta(r_{x,K})^2 / 4\sigma$. Thus if this point $x = p_+$, the correct property holds, and we can use the corresponding q as \hat{p}_+ . Then $\mathcal{N} = \bigcup_{p \in P} \mathcal{N}_p$, and is at most n times the size of \mathcal{N}_p . Let k_p be the smallest integer k such that $r_p/2 \geq R/2^k$. The net \mathcal{N}_p will be composed of $\mathcal{N}_p = \bigcup_{i=0}^{k_p} \mathcal{N}_i = \mathcal{N}_0 \cup \mathcal{N}'_p$, where $\mathcal{N}'_p = \bigcup_{i=1}^{k_p} \mathcal{N}_i$.

Before we proceed with the construction, we need an assumption: That $\Lambda_P = \min_{p \in P} r_p$ is a bounded quantity, it is not too small. That is, no point has *more* than half the points within an absolute radius Λ_P . We call Λ_P the *median concentration*.

Lemma C.8. *A net \mathcal{N}_0 can be constructed of size $O((\sigma/\delta\Lambda_P)^d + \log^{d/2}(n))$ so that all points $x \in B_{R_p}(p)$ satisfy $\|q - x\| \leq \delta(r_{x,K})^2 / 4\sigma$ for some $q \in \mathcal{N}_0$.*

If $x = p_+$, then such a point satisfies the net condition, that is there is a point $q \in \mathcal{N}_0$ such that $\|q - x\| = \|q - p_+\| \leq \delta(r_{p_+,K})^2 / (4\sigma) = \delta D_K(p_+, \hat{p}_2) / (4\sigma) \leq \delta D_K(P, p_+)$.

Proof. For all points $x \in B_{R_p} \subset B_{r_p/2}(p)$, they must have $r_x \geq r_p/2$, otherwise $B_{r_p/2}(x)$ is completely inside $B_{r_p}(p)$, and cannot have enough points. Within $B_{R_p}(p)$ we place the net \mathcal{N}_0 so that all points $x \in B_{R_p}(p)$ satisfy $\|x - q\| \leq \min(\delta r_p^2 / 32\sigma, \sqrt{3}\sigma)$ for some $q \in \mathcal{N}_0$. Now $\delta r_p^2 / 32\sigma \leq \delta r_x^2 / 8\sigma$, and since $\|x - y\|^2 / 2 \leq D_K(x, y)^2$ (for $\|x - y\| \leq \sqrt{3}\sigma$, via Lemma 5.3), thus the net ensures if $p_+ \in B_{R_p}(p)$, then some $q \in \mathcal{N}_0$ is sufficiently close to p_+ .

Since $B_{R_p}(p)$ fits in a squared box of side length $\min(2R_\sigma, r_p)$, then we can describe \mathcal{N}_0 as an axis-aligned grid with g points along each axis. We define two cases to bound g . When $\delta r_p^2/32\sigma < \sqrt{3}\sigma$ then we can set

$$g = \frac{R_p}{\delta r_p^2/(32\sigma\sqrt{d})} \leq \frac{32\sigma\sqrt{d}}{\delta r_p} = O(\sigma/\delta r_p) = O(\sigma/\delta\Lambda_P)$$

Otherwise,

$$g = \frac{R_p}{\sqrt{3}\sigma/\sqrt{d}} \leq \frac{\sigma\sqrt{2\ln(n)}}{\sqrt{3}\sigma/\sqrt{d}} = \sqrt{2d\ln(n)/3} = O(\sqrt{\log(n)}).$$

Then we need $|\mathcal{N}_0| = O(g^d) = O((\sigma/\delta\Lambda_P)^2 + \ln^{d/2}(n))$. \square

When $r_p/2 < R$ we still need to handle the case for $x \in A_p$ where the annulus $A_p = B_R(p) \setminus B_{r_p/2}(p)$. For a point $x \in A_p$ if $p = \min_{p' \in P} \|x - p'\|$ then $r_x \geq \|x - p\|$. We only worry about the net \mathcal{N}'_p on A_p for these points where p is the closest point, the others will be handled by another $\mathcal{N}_{p'}$ for $p' \in P$ and $p' \neq p$.

Recall k_p is the smallest integer k such that $r_p/2 \geq R/2^k$.

Lemma C.9. *A net \mathcal{N}'_p can be constructed of size $O(k_p + (\sigma/\delta\Lambda_P)^d + \log^{d/2}(n))$ so that all points $x \in A_p$ where $p = \arg \min_{p' \in P} \|x - p'\|$, satisfy $\|q - x\| \leq \delta(r_{x,K})^2/4\sigma$ for some $q \in \mathcal{N}'_p$.*

If $x = p_+$, then such a point satisfies the net condition, that is there is a point $q \in \mathcal{N}'_p$ such that $\|q - x\| = \|q - p_+\| \leq \delta(r_{p_+,K})^2/(4\sigma) = \delta D_K(p_+, \hat{p}_2)/(4\sigma) \leq \delta D_K(P, p_+)$.

Proof. We now consider the k_p annuli $\{A_1, \dots, A_{k_p}\}$ which cover A_p . Each $A_i = \{x \in \mathbb{R}^d \mid R/2^{i-1} \geq \|p - x\| > R/2^i\}$ has volume $O((R/2^{i-1})^d)$. For any $x \in A_i$ we have $r_x \geq \|x - p\| \geq R/2^i$, so the Euclidean distance to the nearest $q \in \mathcal{N}_i$ can be at most $\min(\sqrt{3}\sigma, \delta(R/2^i)^2/8\sigma)$. Thus we can cover A_i with a net \mathcal{N}_i of size t_i based on two cases again. If $\delta(R/2^i)^2/8\sigma < \sqrt{3}\sigma$ then

$$t_i = O\left(1 + \left(\frac{R}{2^i} / \left(\frac{\delta}{\sigma} \left(\frac{R}{2^i}\right)^2\right)\right)^d\right) = O\left(1 + \left(\frac{2^i \sigma}{R \delta}\right)^d\right) = O(1) + O\left(\left(\frac{\sigma}{\delta R}\right)^d (2^d)^i\right).$$

Otherwise

$$t_i = O\left(1 + \left(\left(\frac{R}{2^i}\right) / \sqrt{3}\sigma\right)^d\right) = O\left(1 + \left(\frac{R_\sigma = \sigma\sqrt{\log(n)}}{2^i \sigma}\right)^d\right) = O(1) + O\left(\frac{\log^{d/2}(n)}{(2^d)^i}\right).$$

Since $R/2^{k_p} \geq r_p/2 \geq \Lambda_P/2$, then the total size of \mathcal{N}'_p , the union of all of these nets, is $\sum_{i=1}^{k_p} t_i \leq O(k_p) + 2t_{k_p} + 2t_1 = O(k_p + (\sigma/\delta\Lambda_P)^d + \log^{d/2}(n))$. In the first case t_{k_p} dominates the cost and in the second case it is t_1 . \square

Thus the total size of \mathcal{N}_p is $O((\sigma/\delta\Lambda_P)^d + \log^{d/2}(n) + k_p)$ where $k_p \leq \log(R/r_p) + 2$. It just remains to bound k_p . Given that no more than $n/2$ points are collocated on the same spot (which already holds by Λ_P being a bounded quantity), then for all $p \in P$, $r_p \geq \min_{q \neq q' \in P} \|q - q'\|$. The value $\beta_P = \Delta / \min_{q \neq q' \in P} \|q - q'\|$ is known as the *spread* of a point set, and it is common to assume it is an absolute bounded quantity related to the precision of coordinates, where $\log(\beta_P)$ is not too large. Thus we can bound $k_p = O(\log(\beta_P))$.

Theorem C.2. *Consider a point set $P \subset \mathbb{R}^d$ with n points, spread β_P , and median concentration Λ_P . For any $\delta > 0$, in time $O(n^2((\sigma/\delta\Lambda_P)^d + \log^{d/2}(n) + \log(\beta_P)))$ we can find a point \hat{p}_+ such that $D_K(P, \hat{p}_+) \leq (1 + \delta)D_K(P, p_+)$.*

Proof. Using Lemma C.8 and Lemma C.9 we can build a net \mathcal{N} of size $O(n((\sigma/\delta\Lambda_P)^d + \log^{d/2}(n) + \log(\beta_P)))$ such that some $q \in \mathcal{N}$ satisfies $\|q - p_+\| \leq \delta D_K(q, p_+)^2/4\sigma \leq \delta D_K(P, p_+)$. Lemma C.7 ensures that this q satisfies $D_K(P, q) \leq (1 + \delta)D_K(P, p_+)$ since $D_K(P, \cdot)$ is 1-Lipschitz.

We can find such a q and set it as p_+ by evaluating $\kappa(P, q)$ for all $q \in \mathcal{N}$ and taking the one with largest value. This takes $O(n)$ for each $q \in \mathcal{N}$. \square

We claim that in many realistic settings $\sigma/\Lambda_P = O(1)$. In such a case the algorithm runs in $O(n^2(1/\delta^d + \log^{d/2} n + \log(\beta_P)))$ time. If $\sigma/\Lambda_P = o(1)$, then *over* half of the measure described by P will essentially behave as a single point. In many settings P is drawn uniformly from a compact set S , so then choosing σ so that more than half of S has negligible diameter compared to σ will cause that data to be over smoothed. In fact, the definition of Λ_P can be modified so that this radius never contains more than any τn points for any constant $\tau < 1$, and the bounds do not change asymptotically.

D Details on Reconstruction Properties of Kernel Distance

In this section we provide the full proof for some statements from Section 4.

D.1 Topological Estimates using Kernel Power Distance

For persistence diagrams of sublevel sets filtration of d_μ^K and the weighted Rips filtration $\{R_\alpha(P, d_\mu^K)\}$ to be well-defined, we need the technical condition (proved in Lemma D.1 and D.2) that they are q -tame. Recall a filtration F is q -tame if for any $\alpha < \beta$, the homomorphism between $H(F_\alpha)$ and $H(F_\beta)$ induced by the canonical inclusion has finite rank [12, 16].

Lemma D.1. *The sublevel sets filtration of d_μ^K is q -tame.*

Proof. The proof resembles the proof of q -tameness for distance to measure sublevel sets filtration (Proposition 12, [8]). We have shown that d_μ^K is 1-Lipschitz and proper. Its properness property implies that any sublevel set $A := (d_\mu^K)^{-1}([0, \alpha])$ (for $\alpha < c_\mu$) is compact. Since \mathbb{R}^d is triangulable (i.e. homeomorphic to a locally finite simplicial complex), there exists a homeomorphism h from \mathbb{R}^d to a locally finite simplicial complex C . For any $\alpha > 0$, since A is compact, we consider the restriction of C to a finite simplicial complex C_α that contains $h(A)$. The function $(d_\mu^K \circ h^{-1})|_{C_\alpha}$ is continuous on C_α , therefore its sublevel set filtration is q -tame based on Theorem 2.22 of [16], which states that the sublevel sets filtration of a continuous function (defined on a realization of a finite simplicial complex) is q -tame. Extending the above construction to any α , the sublevel sets filtration of $d_\mu^K \circ h^{-1}$ is therefore q -tame. As homology is preserved by homeomorphisms h , this implies that the sublevel sets filtration of d_μ^K is q -tame. \square

Setting $\mu = \mu_P$, Lemma D.1 implies that the sublevel sets filtration of $d_{\mu_P}^K$ is also q -tame.

Lemma D.2. *The weighted Rips filtration $\{R_\alpha(P, d_\mu^K)\}$ is q -tame for compact subset $P \subset \mathbb{R}^d$.*

Proof. Since P is compact subset of \mathbb{R}^d , $\text{Dgm}(\{R_\alpha(P, d_\mu^K)\})$ is q -tame based on Proposition 32 of [16], which states that the weighted Rips filtration with respect to a compact subset P in metric space and its corresponding weight function is q -tame. \square

Setting $P = \hat{P}_+$, $\mu = \mu_P$, Lemma D.2 implies that the weighted Rips filtration $\{R_\alpha(\hat{P}_+, d_{\mu_P}^K)\}$ is well-defined.

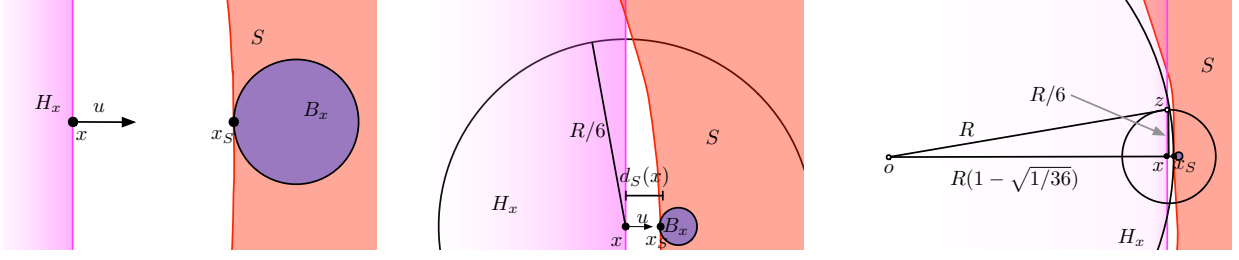


Figure 6: Illustrations of the geometric inference of S from d_μ^K at three scales.

D.2 Inference of Compact Set S with the Kernel Distance

Suppose μ is a uniform measure on a compact set S in \mathbb{R}^d . We now compare the kernel distance d_μ^K with the distance function f_S to the support S of μ . We show how d_μ^K approximates f_S , and thus allows one to infer geometric properties of S from samples from μ .

For a point $x \in \mathbb{R}^d$, the distance function f_S measures the minimum distance between x and any point in S , $f_S(x) = \inf_{y \in S} \|x - y\|$. The point x_S that realizes the minimum in the definition of $f_S(x)$ is the *orthogonal projection* of x on S . The location of the points $x \in \mathbb{R}^d$ that have more than one projection on S is the *medial axis* of S [54], denoted as $M(S)$. Since $M(S)$ resides in the unbounded component $\mathbb{R}^d \setminus S$, it is referred to as the *outer medial axis* similar to the concept found in [28]. The *reach* of S is the minimum distance between a point in S and a point in its medial axis, denoted as $\text{reach}(S)$. Similarly, one could define the medial axis of $\mathbb{R}^d \setminus S$ (i.e. the *inner medial axis* which resides in the interior of S) following definitions in [53], and denote its associated reach as $\text{reach}(\mathbb{R}^d \setminus S)$. The concepts of reach associated with the inner and outer medial axis of S capture curvature information of the compact set.

Recall that a generalized gradient and its corresponding flow to a distance function are described in [14] and later adapted for distance-like functions in [15]. Let $f_S : \mathbb{R}^d \rightarrow \mathbb{R}$ be a distance function associated with a compact set S of \mathbb{R}^d . It is not differentiable on the medial axis of S . It is possible to define a *generalized gradient function* $\nabla_S : \mathbb{R}^d \rightarrow \mathbb{R}^d$ coincides with the usual gradient of f_S where f_S is differentiable, and is defined everywhere and can be integrated into a continuous flow $\Phi^t : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Such a flow points away from S , towards local maxima of f_S (that belong to the medial axis of S) [54]. The integral (flow) line γ of this flow starting at point in \mathbb{R}^d can be parameterized by arc length, $\gamma : [a, b] \rightarrow \mathbb{R}^d$, and we have $f_S(\gamma(b)) = f_S(\gamma(a)) + \int_a^b \|\nabla_S(\gamma(t))\| dt$.

Lemma D.3 (Lemma 4.1). *Given any flow line γ associated with the generalized gradient function ∇_S , $d_\mu^K(x)$ is strictly monotonically increasing along γ for x sufficiently far away from the medial axis of S , for $\sigma \leq \frac{R}{6\Delta_G}$ and $f_S(x) \in (0.014R, 2\sigma)$. Here $B(\sigma/2)$ denotes a ball of radius $\sigma/2$, $G := \frac{\text{Vol}(B(\sigma/2))}{\text{Vol}(S)}$, $\Delta_G := \sqrt{12 + 3 \ln(4/G)}$ and suppose $R := \min(\text{reach}(S), \text{reach}(\mathbb{R}^d \setminus S)) > 0$.*

Proof. Since $d_\mu^K(x)$ is always positive, and $d_\mu^K(x) = \sqrt{c_\mu - 2\text{KDE}_\mu(x)}$ where c_μ is a constant that depends only on μ , K , and σ , then it is sufficient to show that $\text{KDE}_\mu(x)$ is strictly monotonically decreasing along γ .

Let u be the negative of the direction of the flow line γ at x (i.e u is a unit vector that points towards S). We show that $\text{KDE}_\mu(x)$ is strictly monotonically increasing along u . Informally, we will observe that all parts of S that are “close” to x are in the direction u , and that these parts dominate the gradient of $\text{KDE}_\mu(x)$ along u . We now make this more formal by describing two quantities, B_x and A_x , illustrated in Figure 6.

For a point $x \in \mathbb{R}^d$, let $x_S = \arg \min_{x' \in S} \|x' - x\|$; since x is not on the medial axis of S , x_S is uniquely defined and u points in the direction of $(x_S - x)/f_S(x)$. First, we claim that there exists a ball B_x of radius $\sigma/2$ incident to x_S that is completely contained in S . This holds since $\sigma/2 \leq \frac{R}{6\Delta_G} < R \leq \text{reach}(\mathbb{R}^d \setminus S)$. In addition, since $f_S(x) < 2\sigma$, no part in B_x is further than 3σ from x . Second, we claim that no part of S

within $\Delta_G \cdot \sigma$ ($\leq R/6$) of x (this includes B_x) is in the halfspace H_x with boundary passing through x and outward normal defined by u . To see this, let o be the center of a ball with radius R that is incident to x_S but not in S , refer to such a ball as B_o . This implies that points o , x and x_S are colinear. Then a ball centered at x with radius $R/6$ should intersect S outside of B_o , and in the worst case, on the boundary of H_x . This holds as long as $\|x - x_S\| \geq 0.014R \geq (1 - \sqrt{35/36})R$; see Figure 6. Define $A_x = \{y \in S \mid \|x - y\| > \Delta_G \cdot \sigma\}$.

Now we examine the contributions to the directional derivative of $\text{KDE}_\mu(x)$ along the direction of u from points in B_x and A_x , respectively. Such a directional derivative is denoted as $D_u \text{KDE}_\mu(x)$. Recall $\text{KDE}_\mu(x) = \int_{y \in S} K(x, y) d\mu(y)$ and μ is a uniform measure on S , $D_u \text{KDE}_\mu(x) = \frac{1}{\text{Vol}(S)} \int_{y \in S} D_u K(x, y)$. For any point $y \in \mathbb{R}^d$, we define $g(y) := D_u K(x, y) = \exp(-\|x - y\|^2/2\sigma^2) \langle y - x, u \rangle$. Therefore $D_u \text{KDE}_\mu(x) = \frac{1}{\text{Vol}(S)} \int_{y \in S} g(y)$.

We now examine the contribution to $D_u \text{KDE}_\mu(x)$ from points in B_x , $\frac{1}{\text{Vol}(S)} \int_{y \in B_x} g(y)$. First, for all points $y \in B_x$, since $\|x - y\| \leq 3\sigma$, we have $\exp(-\|x - y\|^2/2\sigma^2) \geq \exp(-9/2)$. Second, at least half of points $y \in B_x$ (that covers half the volume of B_x) is at least $\sigma/2$ away from x_S , and correspondingly for these points $\langle y - x, u \rangle \geq \sigma/2$. We have $\int_{y \in B_x} g(y) \geq \frac{1}{2} \text{Vol}(B_x) \cdot \exp(-9/2) \cdot \sigma/2$. Given $\text{Vol}(B_x) = G \cdot \text{Vol}(S)$, we have $\frac{1}{\text{Vol}(S)} \int_{y \in B_x} g(y) \geq \frac{1}{4} G \cdot \exp(-9/2) \cdot \sigma$. Denote $B = \frac{1}{4} G \cdot \exp(-9/2) \cdot \sigma$.

We now examine the contribution to $D_u \text{KDE}_\mu(x)$ from points in A_x , $\frac{1}{\text{Vol}(S)} \int_{y \in A_x} g(y)$. For any point $y \in \mathbb{R}^d$ (including $y \in A_x$), $\langle y - x, u \rangle \leq \|x - y\|$. Let $\phi_y = \|x - y\|/\sigma$ so we have $g(y) \leq \exp(-\phi_y^2/2) \phi_y \sigma$. Since this bound on $g(y)$ is maximized at $\phi_y = 1$, under the condition $\phi_y \geq \Delta_G \geq \sqrt{12} > 1$, we can set $\phi_y = \Delta_G$ to achieve the bound $g(y) \leq \exp(-\Delta_G^2/2) \cdot \Delta_G \sigma$ for $\|x - y\| \geq \Delta_G \cdot \sigma$ (that is, for all $y \in A_x$). Now we have $\int_{y \in A_x} g(y) \leq \text{Vol}(S) \exp(-\Delta_G^2/2) \cdot \Delta_G \sigma$, leading to $\frac{1}{\text{Vol}(S)} \int_{y \in A_x} g(y) \leq \exp(-\Delta_G^2/2) \cdot \Delta_G \sigma$. Denote $A = \exp(-\Delta_G^2/2) \cdot \Delta_G \sigma$.

Since only the points $y \in A_x$ could possibly reside in H_x and thus can cause $g(y)$ to be negative, we just need to show that $B > A$. This can be confirmed by plugging in $\Delta_G = \sqrt{12 + 3 \ln(4/G)}$, and using some algebraic manipulation. \square

E Lower Bound on Wasserstein Distance

We note the next result is a known lower bounds for the Earth movers distance [23][Theorem 7]. We reprove it here for completeness.

Lemma E.1 (Lemma. 5.8). *For any probability measures μ and ν defined on \mathbb{R}^d we have $\|\bar{\mu} - \bar{\nu}\| \leq W_2(\mu, \nu)$.*

Proof. Let $\pi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ describes the optimal transportation plan from μ to ν . Also let $u_{\mu, \nu} = \frac{(\bar{\mu} - \bar{\nu})}{\|\bar{\mu} - \bar{\nu}\|}$ be the unit vector from $\bar{\mu}$ to $\bar{\nu}$. Then we can expand

$$\begin{aligned} (W_2(\mu, \nu))^2 &= \int_{(p, q)} \|p - q\|^2 d\pi(p, q) \geq \int_{(p, q)} (\langle (p - q), u_{\mu, \nu} \rangle)^2 d\pi(p, q) \\ &\geq \|\bar{\mu} - \bar{\nu}\|^2. \end{aligned}$$

The first inequality follows since $\langle (p - q), u_{\mu, \nu} \rangle$ is the length of a projection and thus must be at most $\|p - q\|$. The second inequality follows since that projection describes the squared length of mass $\pi(p, q)$ along the direction between the two centers $\bar{\mu}$ and $\bar{\nu}$, and the total sum of squared length of unit mass moved is exactly $\|\bar{\mu} - \bar{\nu}\|^2$. Note the left-hand-side of the second inequality could be larger since some movement may cancel out (e.g. a rotation). \square