# Topological analysis of scalar fields with outliers

**Mickaël Buchet[1], Frédéric Chazal[2], Tamal K. Dey[3], Fengtao Fan[4], Steve Y. Oudot[5], and Yusu Wang[6]**

1 **Inria Saclay Île-de-France,**
**Palaiseau, France,**
mickael.buchet@m4x.org

2 **Inria Saclay Île-de-France,**
**Palaiseau, France,**
frederic.chazal@inria.fr

3 **Department of Computer Science and Engineering,**
**The Ohio State University,**
**Columbus, OH 43210, USA.**
tamaldey@cse.ohio-state.edu

4 **Department of Computer Science and Engineering,**
**The Ohio State University,**
**Columbus, OH 43210, USA.**
fanf@cse.ohio-state.edu

5 **Inria Saclay Île-de-France,**
**Palaiseau, France,**
steve.oudot@inria.fr

6 **Department of Computer Science and Engineering,**
**The Ohio State University,**
**Columbus, OH 43210, USA.**
yusu@cse.ohio-state.edu

## Abstract

Given a real-valued function $f$ defined over a manifold $\mathsf{M}$ embedded in $\mathbb{R}^d$, we are interested in recovering structural information about $f$ from the sole information of its values on a finite sample $P \subset \mathsf{M}$. Existing methods provide approximation to the persistence diagram of $f$ when the noise is bounded in both the functional and geometric domains. However, they fail in the presence of aberrant values, also called outliers, both in theory and practice.

We propose a new algorithm that deals with outliers. We handle aberrant functional values with a method inspired from the k-nearest neighbors regression and the local median filtering, while the geometric outliers are handled using the distance to a measure. Combined with topological results on nested filtrations, our algorithm performs robust topological analysis of scalar fields in a wider range of noise models than handled by current methods. We provide theoretical guarantees on the quality of our approximation and some experimental results illustrating its behavior.

**1998 ACM Subject Classification** Computational Geometry and Object Modeling

**Keywords and phrases** Persistent Homology, Topological Data Analysis, Scalar Field Analysis, Nested Rips Filtration, Distance to a Measure

**Digital Object Identifier** 10.4230/LIPIcs.xxx.yyy.p



Conference title on which this volume is based on.
Editors: Billy Editor and Bill Editors; pp. 1–24

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Introduction

Consider a network of sensors measuring a quantity such as the temperature, the humidity, or the elevation. These sensors also compute their positions and communicate these data to others. However, they are not perfect and can make mistakes such as providing some aberrant values. Can we still recover the topological structure of the measured quantity?

This is an instance of a scalar field analysis problem. Given a manifold $\mathsf{M}$ embedded in $\mathbb{R}^d$ and a scalar field $f : \mathsf{M} \to \mathbb{R}$, we want to extract the topological information of $f$, knowing only its values on a finite set of points $P$ sampled from $\mathsf{M}$. The topology of a function could refer to features such as peaks (local maxima) and pits (local minima). In addition, it is also interesting to be able to evaluate the prominence of these features which is the same notion geographers use to distinguish between a summit and a local maximum in its shadow. Such information can be captured by the so-called *topological persistence*, which studies the *sub-level sets* $f^{-1}(]-\infty, \alpha])$ of a function $f$ and the way their topology evolves with the parameter $\alpha$. In the case of geography, we can use the function minus-elevation to study the topography. Peaks will appear depending on their altitude and will merge into other topological features at saddle points. This provides a *persistence diagram* describing the lifespan of features where the prominent ones have the long lifespans.

When the domain $\mathsf{M}$ of the function $f$ is triangulated, one classical way of computing this diagram is to linearly interpolate the function $f$ on each simplex and then apply the standard persistence algorithm to this piecewise-linear function [18]. For cases where we only have pairwise distances between input points, one can build a family of complexes and infer the persistent homology of the input function $f$ from them [5] (this construction will be detailed in Section 2).

Both of these two approaches can provably infer correct topology when the input points admit a bounded noise model: in particular, the Haussdorf distance between $P$ and $\mathsf{M}$ is bounded and the error on the observed value of $f$ is also bounded. What happens if the noise is unbounded? A faulty sensor can provide completely wrong information or a bad position. Previous methods no longer work in this setting. Moreover, a sensor with a good functional value but a bad position can become an outlier in function value at its measured position (see Section 3.1 for an example). In this paper, we study the problem of scalar field analysis in the presence of unbounded noise both in the geometry and in the functional values. To the best of our knowledge, there is no other method to handle such combined unbounded geometric and functional noise with theoretical guarantees.

## Contributions

We consider a general noise model. Intuitively, a sample $(P, \tilde{f})$ of a function $f : \mathsf{M} \to \mathbb{R}$ respects our noise model if: (i) the domain $\mathsf{M}$ is sampled densely enough and there is no cluster of noisy samples outside $\mathsf{M}$ (roughly speaking, no area outside $\mathsf{M}$ has a higher sampling density than on $\mathsf{M}$), and (ii) for any point of $P$, at least half of its $k$ nearest neighbors have a functional value with an error less than a threshold $s$. This model allows functional outliers that may have a value arbitrarily far away from the true one. This noise model encompasses the previous bounded noise model as well as other noise models such as bounded Wasserstein distance for geometry, or generative models like convolution with a Gaussian. Connection to some of these classical noise models can be found in Appendices A and B.

We show how to infer the persistence diagram of $f$ knowing only $\tilde{f}$ on the set $P$. This comes with theoretical guarantees when the sampling respects the new noise model. We

achieve this goal through three main steps:

1. Using the observations $\tilde{f}$, we provide a new estimator $\hat{f}$ to approximate $f$. This estimator is inspired by the $k$-nearest neighbours regression technique but differs from it in an essential way.
2. We filter geometric outliers using a distance to a measure function.
3. We combine both techniques in a unified framework to estimate the persistence diagram of $f$.

The two sources of noise are not independent. The interdependency is first acknowledged by assuming appropriate noise models and then untangled by separate steps in our algorithm.

### Related work.

As mentioned earlier, a framework has been previously proposed in [5] for scalar field topology inference with theoretical guarantees. However, it is limited to a bounded noise assumption, which we aim to relax.

For handling the functional noise only, the traditional non-parametric regression mostly uses kernel-based or $k$-NN estimators. The $k$-NN methods are more versatile [13]. Nevertheless, the kernel-based estimators are preferred when there is structure in the data. However, the functional outliers destroys the structure on which kernel-based estimators rely. These functional outliers can arise as a result of geometric outliers (see Section 3.1). Thus, in a way, it is essential to be able to handle functional outliers when the input has geometric noise. Functional outliers can also introduce a bias that hampers the robustness of a $k$-NN regression. For example, if all outliers' values are greater than the target value, a $k$-NN regression will shift towards a larger value. Our approach leverages the $k$-NN regression idea while trying to avoid the sensitivity to this bias.

Various methods for geometric denoising have also been proposed in the literature. If the generative model for noise is known a priori, one can use de-convolution to remove noise. Some methods have been specifically adapted to using topological information for such denoising [14]. In our case where the generative model is unknown, we use a filtering by the value of the distance to a measure, which has been successfully applied to infer the topology of a domain under unbounded noise [4].

## 2 Preliminaries for Scalar Field Analysis

In [5], Chazal et al. presented an algorithm to analyze the scalar field topology using persistent homology which can handle bounded Hausdorff noise both in geometry and in observed function values. Our approach follows the same high level framework. Hence in this section, we introduce necessary preliminaries along with some of the results from [5].

### Riemannian manifold and its sampling.

Consider a compact Riemannian manifold $\mathsf{M}$. Let $d_{\mathsf{M}}$ denote the Riemannian metric on $\mathsf{M}$. Consider the open Riemannian ball $B_{\mathsf{M}}(x,r) := \{y \in \mathsf{M} \mid d_{\mathsf{M}}(x,y) < r\}$ centered at $x \in \mathsf{M}$. $B_{\mathsf{M}}(x,r)$ is *strongly convex* if for any pair $(y,y')$ in the closure of $B_{\mathsf{M}}(x,r)$, there exists a unique minimizing geodesic between $y$ and $y'$ whose interior is contained in $B_{\mathsf{M}}(x,r)$. Given any $x \in \mathsf{M}$, let $\varrho(x)$ denote the supremum of the value of $r$ such that $B_{\mathsf{M}}(x,r)$ is strongly convex. As $\mathsf{M}$ is compact, the infimum of all $\varrho(x)$ is positive and we denote it by $\varrho(\mathsf{M})$, which is called the *strong convexity radius* of $\mathsf{M}$.

A point set $P \subseteq \mathsf{M}$ is a *geodesic $\varepsilon$-sampling* of $\mathsf{M}$ if for any point $x$ of $\mathsf{M}$, the distance from $x$ to $P$ is less than $\varepsilon$ in the metric $d_{\mathsf{M}}$. Given a $c$-Lipschitz scalar function $f : \mathsf{M} \to \mathbb{R}$, we aim to study the topological structure of $f$. However, the scalar field $f : \mathsf{M} \to \mathbb{R}$ is only approximated by a discrete set of sample points $P$ and a function $\tilde{f} : P \to \mathbb{R}$. The goal of this paper is to retrieve the topological structure of $f$ from $\tilde{f}$ when some forms of noise are present both in the positions of $P$ and in the function values of $\tilde{f}$.

## Persistent homology.

As in [5], we infer the topology of $f$ using persistent homology of well-chosen *persistence modules*. A *filtration* $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ is a family of sets $F_\alpha$ totally ordered by inclusions $F_\alpha \subset F_\beta$. Following [3], a persistence module is a family of vector spaces $\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}$ with a family of homomorphisms $\phi_\alpha^\beta : \Phi_\alpha \to \Phi_\beta$ such that for all $\alpha \leq \beta \leq \gamma$, $\phi_\alpha^\gamma = \phi_\beta^\gamma \circ \phi_\alpha^\beta$. Given a filtration $\mathcal{F} = \{F_\alpha\}_{\alpha\in\mathbb{R}}$ and $\alpha \leq \beta$, the canonical inclusion $F_\alpha \hookrightarrow F_\beta$ induces a homomorphism at the homology level $H_*(F_\alpha) \to H_*(F_\beta)$. These homomorphisms and the homology groups of $F_\alpha$ form a persistence module called the *persistence module* of $\mathcal{F}$.

The persistence module of the filtration $\mathcal{F} = \{F_\alpha\}_{\alpha\in\mathbb{R}}$ is said to be *q-tame* when all the homomorphisms $H_*(F_\alpha) \to H_*(F_\beta)$ have finite rank [2]. Its algebraic structure can then be described by the *persistence diagram* $\mathrm{Dgm}(\mathcal{F})$, which is a multiset of points in $\mathbb{R}^2$ describing the lifespan of the homological features in the filtration $\mathcal{F}$. For technical reasons, $\mathrm{Dgm}(\mathcal{F})$ also contains the diagonal $y = x$ with infinite multiplicity. See [10] for a more formal discussion of the persistence diagrams.

Persistence diagrams can be compared using the *bottleneck distance* $d_B$ [7]. Given two multisets with the same cardinality, possibly infinite, $D$ and $E$ in $\mathbb{R}^2$, we consider the set $\mathcal{B}$ of all bijections between $D$ and $E$. The bottleneck distance (under $L_\infty$-norm) is then defined as:

$$d_B(D, E) = \inf_{b\in\mathcal{B}} \max_{x\in D} ||x - b(x)||_\infty. \tag{1}$$

Two filtrations $\{U_\alpha\}$ and $\{V_\alpha\}$ are said to be *$\varepsilon$-interleaved* if, for any $\alpha$, we have $U_\alpha \subset V_{\alpha+\varepsilon} \subset U_{\alpha+2\varepsilon}$. Recent work in [2, 3] shows that two "nearby" filtrations (as measured by the interleaving distance) will induce close persistence diagrams in the bottleneck distance.

▶ **Theorem 2.1.** *Let $U$ and $V$ be two q-tame and $\varepsilon$-interleaved filtrations. Then the persistence diagrams of these filtrations verify $d_B(\mathrm{Dgm}(U), \mathrm{Dgm}(V)) \leq \varepsilon$.*

## Nested filtrations.

The scalar field topology of $f : \mathsf{M} \to \mathbb{R}$ is studied via the topological structure of the sub-level sets filtration of $f$. More precisely, the sub-level sets of $f$ are defined as $F_\alpha = f^{-1}(]-\infty, \alpha])$ for any $\alpha \in \mathbb{R}$. The collection of sub-level sets form a filtration $\mathcal{F} = \{F_\alpha\}_{\alpha\in\mathbb{R}}$ connected by natural inclusions $F_\alpha \subseteq F_\beta$ for any $\alpha \leq \beta$. Our goal is to approximate the persistence diagram $\mathrm{Dgm}(\mathcal{F})$ from the observed scalar field $\tilde{f} : P \to \mathbb{R}$. We now describe the results of [5] for approximating $\mathrm{Dgm}(\mathcal{F})$ when $P$ is a geodesic $\varepsilon$-sampling of $\mathsf{M}$. These results will later be useful for our approach.

To simulate the sub-level sets filtration $\{F_\alpha\}$ of $f$, we introduce $P_\alpha = \tilde{f}^{-1}(]-\infty, \alpha]) \subset P$ for any $\alpha \in \mathbb{R}$. The points in $P_\alpha$ intuitively sample the sub-level set $F_\alpha$. To estimate the topology of $F_\alpha$ from these discrete samples $P_\alpha$, we consider the *$\delta$-offset* $P^\delta$ of the point set $P$ i.e. we grow geodesic balls of radius $\delta$ around the points of $P$. This gives us a union of balls that serves as a proxy for $f^{-1}(]-\infty, \alpha)$ and whose nerve is known as the *Čech complex*, $C_\delta(P)$.

It has many interesting properties but becomes difficult to compute in high dimensions. We consider an easier to compute complex called the *Vietoris-Rips complex* $R_\delta(P)$, defined as the maximal simplicial complex with the same 1-skeleton as the Čech complex. The Čech and Rips complexes are related in any metric space: $\forall \delta > 0,\ C_\delta(P) \subset R_\delta(P) \subset C_{2\delta}(P)$.

Even though no Vietoris-Rips complex might capture the topology of the manifold $\mathsf{M}$, it was shown in [6] that a structure of nested complexes can recover it from the filtration $\{P_\alpha\}$ using the inclusions $R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha)$. Specifically, for a fixed $\delta > 0$, consider the following commutative diagram induced by inclusions, for $\alpha \leq \beta$:

$$\begin{array}{ccc} H_*(R_{2\delta}(P_\alpha)) & \longrightarrow & H_*(R_{2\delta}(P_\beta)) \\ \uparrow & & \uparrow \\ H_*(R_\delta(P_\alpha)) & \longrightarrow & H_*(R_\delta(P_\beta)) \end{array}$$

As the diagram commutes for all $\alpha \leq \beta$, $\{\Phi_\alpha, \phi_\alpha^\beta\}$ defines a persistence module. We call it the persistent homology module of the filtration of the nested pairs $\{R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha)\}_{\alpha\in\mathbb{R}}$. This construction can also be done for any filtration of nested pairs. Using this construction, one of the main results of [5] is:

▶ **Theorem 2.2** (Theorems 2 and 6 of [5]). *Let $\mathsf{M}$ be a compact Riemannian manifold and let $f : \mathsf{M} \to \mathbb{R}$ be a $c$-Lipschitz function. Let $P$ be a geodesic $\varepsilon$-sampling of $\mathsf{M}$. If $\varepsilon < \frac{1}{4}\varrho(\mathsf{M})$, then for any $\delta \in \left[2\varepsilon, \frac{1}{2}\varrho(\mathsf{M})\right)$, the persistent homology modules of $f$ and of the filtration of nested pairs $\{R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha)\}$ are $2c\delta$-interleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta$.*

*Furthermore, the $k$-dimensional persistence diagram for the filtrations of nested pairs $\{R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha)\}$ can be computed in $O(|P|kN + N\log N + N^3)$ time, where $N$ is the number of simplices of $\{R_{2\delta}(P_\infty)\}$, and $|P|$ denotes the cardinality of the sample set $P$.*

It has been observed that in practice, the persistence algorithm often has a running time linear in the number of simplices, which reduces the above complexity to $O(|P| + N\log N)$ in a practical setting.

We say that $\tilde{f}$ has a precision of $\xi$ over $P$ if $|\tilde{f}(p) - f(p)| \leq \xi$ for any $p \in P$. We then have the following result for the case when only this Hausdorff-type functional noise is present:

▶ **Theorem 2.3** (Theorem 3 of [5]). *Let $\mathsf{M}$ be a compact Riemannian manifold and let $f : \mathsf{M} \to \mathbb{R}$ be a $c$-Lipschitz function. Let $P$ be a geodesic $\varepsilon$-sampling of $\mathsf{M}$ such that the values of $f$ on $P$ are known with precision $\xi$. If $\varepsilon < \frac{1}{4}\varrho(\mathsf{M})$, then for any $\delta \in \left[2\varepsilon, \frac{1}{2}\varrho(\mathsf{M})\right)$, the persistent homology modules of $f$ and of the filtration of nested pairs $\{R_\delta(P_\alpha) \hookrightarrow R_{2\delta}(P_\alpha)\}$ are $(2c\delta + \xi)$-interleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta + \xi$.*

Geometric noise was considered in the form of bounded noise in the estimate of the geodesic distances between points in $P$. It translated into a relation between the measured pairwise distances and the real ones. With only geometric noise, [5] provided the following stability result. It was stated in this form in the conference version of the paper.

▶ **Theorem 2.4** (Theorem 4 of [5]). *Let $\mathsf{M}$, $f$ be defined as previously and $P$ be an $\epsilon$-sample of $\mathsf{M}$ in its Riemannian metric. Assume that, for a parameter $\delta > 0$, the Rips complexes $R_\delta(\cdot)$ are defined with respect to a metric $\tilde{d}(\cdot,\cdot)$ which satisfies $\forall x, y \in P$, $\frac{d_\mathsf{M}(x,y)}{\lambda} \leq \tilde{d}(x,y) \leq \nu + \mu\frac{d_\mathsf{M}(x,y)}{\lambda}$, where $\lambda \geq 1$ is a sclaing factor, $\mu \geq 1$ is a relative error and $\nu \geq 0$ an additive error. Then, for any $\delta \geq \nu + 2\mu\frac{\epsilon}{\lambda}$ and any $\delta' \in [\nu + 2\mu\delta,\ \frac{1}{\lambda}\varrho(\mathsf{M})]$, the persistent homology modules of $f$ and of the filtration of nested pairs $\{R_\delta(P_\alpha) \hookrightarrow R_{\delta'}(P_\alpha)\}$ are $c\lambda\delta'$-interleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $c\lambda\delta'$.*

# 3 Functional Noise

In this section, we focus on the case where we have only functional noise in the observed function $\tilde{f}$. Suppose we have a scalar function $f$ defined on a manifold $\mathsf{M}$ embedded in a metric space $\mathbb{X}$ (such as the Euclidean space $\mathbb{R}^d$). We are given a geodesic $\varepsilon$-sample $P \subset \mathsf{M}$, and a noisy observed function $\tilde{f} : P \to \mathbb{R}$. Our goal is to approximate the persistence diagram $\mathrm{Dgm}(\mathcal{F})$ of the sub-level set filtration $\mathcal{F} = \{F_\alpha = f^{-1}((-\infty, \alpha])\}_\alpha$ from $\tilde{f}$. We assume that $f$ is $c$-Lipschitz with respect to the intrinsic metric of the manifold $\mathsf{M}$. Note that this does not imply a Lipschitz condition on $\tilde{f}$.

## 3.1 Functional noise model

Previous work on functional noise usually focuses on Hausdorff-type bounded noise (e.g, [5]) or statistical noise with zero-mean (e.g, [15]). However, we observe that there are many practical scenarios where the observed function $\tilde{f}$ may contain these previously considered types of noise mixed with *aberrant function values* in $\tilde{f}$. Hence, we propose below a more general noise model that allows such a mixture.

### Motivating examples.

First, we provide some motivating examples for the need of handling *aberrant* function values in $\tilde{f}$, where $\tilde{f}(p)$ at some sample point $p$ can be totally unrelated to the true value $f(p)$. Consider a sensor network, where each node returns some measures. Such measurements can be imprecise, and in addition to that, a sensor may experience failure and return a completely wrong measure that has no relation with the true value of $f$. Similarly, an image could be corrupted with white noise where there are random pixels with aberrant function values, such as random white or black dots.

More interestingly, outliers in function values can naturally appear as a result of (extrinsic) geometric noise present in the discrete samples. For example, imagine that we have a process that can measure the function value $f : \mathsf{M} \to \mathbb{R}$ with *no error*. However, the geometric location $\tilde{p}$ of a point $p \in \mathsf{M}$ can be wrong. In particular, $\tilde{p}$ can be close to other parts of the manifold, thereby although $\tilde{p}$ has the correct function value $f(p)$, it becomes a functional outlier among its neighbors (due to the wrong location of $\tilde{p}$). See Figure 1 for an illustration, where the two sides of the narrow neck of this bone-structure have very different function values. Now, suppose that the points are sampled uniformly on $\mathsf{M}$ and their position is then convolved with a Gaussian noise. Then points from one side of this neck can be sent closer to the other side, causing aberrant values in the observed function.

In fact, even if we assume that we have a "magic filter" that can project each sample back onto the underlying manifold $\mathsf{M}$, the result is a new set of samples where all points are on the manifold and thus can be seen as having **no** geometric noise; however, this point set now contains functional noise which is actually caused by the original geometric noise. Note that such a magic filter is the goal of many geometric denoising methods. This implies that a denoising algorithm perfect in the sense of geometric noise cannot remove or may even cause more aberrant functional noise. This motivates the need for handling functional outliers (in addition to traditional functional noise) as well as processing noise that combines geometric and functional noise together and that is not necessarily centered. Figure 1 shows a bone-like curve and a function defined as the curvilinear abscissa. The Gaussian noise applied to the example creates outliers even after applying a projection onto the original object.

Another case where our approach is useful concerns missing data. Assuming that some of the functional values are missing, we can replace them by anything and act as if they were outliers. Without modifying the algorithm, we obtain a way to handle the local loss of information.

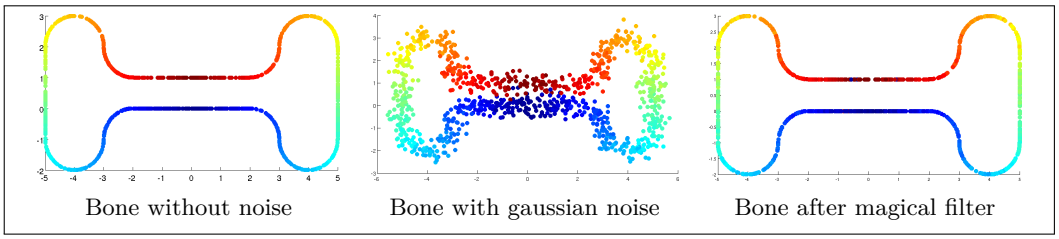



**Figure 1** Bone example after applying Gaussian perturbation and magical filter

### Functional noise model.

To allow both aberrant and more traditional functional noise, we introduce the following noise model. Let $P \subset \mathsf{M}$ be a geodesic $\varepsilon$-sample of the underlying manifold $\mathsf{M}$. Intuitively, our noise model requires that for any point $p \in P$, locally there is a sufficient number of sample points with reasonably good function values. Specifically, we fix two parameters $k$ and $k'$ with the condition that $k \geq k' > \frac{1}{2}k$. Let $\mathrm{NN}^k_P(p)$ denote the set of the $k$-nearest neighbors of $p$ in $P$ in the *extrinsic metric*. We say that a discrete scalar field $\tilde{f} : P \to \mathbb{R}$ is *a* $(k, k', \Delta)$*-functional-sample* of $f : \mathsf{M} \to \mathbb{R}$ if the following holds:

$$\forall p \in P, \ \left| \left\{ q \in \mathrm{NN}^k_P(p) \ \middle| \ |\tilde{f}(q) - f(p)| \leq \Delta \right\} \right| \geq k' \tag{2}$$

Intuitively, this noise model allows up to $k - k'$ samples around a point $p$ to be outliers (whose function values deviates from $f(p)$ by at least $\Delta$). In Appendix A, we consider two common functional noise models used in the statistical learning community and look at what they correspond to in our setting.

## 3.2 Functional Denoising

Given a scalar field $\tilde{f} : P \to \mathbb{R}$ which is a $(k, k', \Delta)$-functional-sample of $f : \mathsf{M} \to \mathbb{R}$, we now aim to compute a denoised function $\widehat{f} : P \to \mathbb{R}$ from the observed function $\tilde{f}$, and we will later use $\widehat{f}$ to infer the topology of $f : \mathsf{M} \to \mathbb{R}$. Below we describe two ways to denoise the noisy observation $\tilde{f}$: one of which is well-known, and the other one is new. As we will see later, these two treatments lead to similar theoretical guarantees in terms of topology inference. However, they have different characteristics in practice, which are discussed in the experimental illustration of Appendix C.

### $k$-median.

In the k-median treatment, we simply perform the following: given any point $p \in P$, we set $\widehat{f}(p)$ to be the median value of the set of $\tilde{f}$ values for the $k$-nearest neighbors $\mathrm{NN}^k_P(p) \subseteq P$ of $p$. We call $\widehat{f}$ the *k-median denoising of* $\tilde{f}$. The following observation is straightforward:

▶ Observation 1. If $\tilde{f} : P \to \mathbb{R}$ is a $(k, k', \Delta)$-functional-sample of $f : \mathsf{M} \to \mathbb{R}$ with $k' \geq k/2$, then we have $|\widehat{f}(p) - f(p)| \leq \Delta$ for any $p \in P$, where $\widehat{f}$ is the k-median denoising of $\tilde{f}$.

### Discrepancy.

In the k-median treatment, we choose a single value from the $k$-nearest neighbors of a sample point $p$ and set it to be the denoised value $\widehat{f}(p)$. This value, while within $\Delta$ distance to the true value $f(p)$ when $k' \geq k/2$, tends to have greater variability among neighboring sample points. Intuitively, taking the average (such as $k$-means) makes the function $\widehat{f}(p)$ smoother, but it is sensitive to outliers. We combine these ideas together, and use the following concept of discrepancy to help us identify a subset of points from the k-nearest neighbors of a sample point $p$ to estimate $\widehat{f}(p)$.

Given a set $Y = \{x_1, \ldots, x_m\}$ of $m$ sample points from $P$, we define its discrepancy w.r.t. $\tilde{f}$ as:

$$\phi(Y) = \frac{1}{m}\sum_{i=1}^{m}(\tilde{f}(x_i) - \mu(Y))^2, \quad \text{where } \mu(Y) = \frac{1}{m}\sum_{i=1}^{m}\tilde{f}(x_i).$$

$\mu(Y)$ and $\phi(Y)$ are respectively the average and the variance of the observed function values for points from $Y$. Intuitively, $\phi(Y)$ measures how tight the function values $(\tilde{f}(x_i))$ are clustered. Now, given a point $p \in P$, we define

$$\widehat{Y}_p = \operatorname*{argmin}_{Y \subseteq \mathrm{NN}_P^k(p), |Y|=k'} \phi(Y), \quad \text{and } \widehat{z}_p = \mu(\widehat{Y}_p).$$

That is, $\widehat{Y}_p$ is the subset of $k'$ points from the $k$-nearest neighbors of $p$ that has the smallest discrepancy and $\hat{z}_p$ is its mass center. It turns out that $\widehat{Y}_p$ and $\widehat{z}_p$ can be computed by the following sliding-window procedure: (i) Sort $\mathrm{NN}_P^k(p) = \{x_1, \ldots, x_k\}$ according to $\tilde{f}(x_i)$. (ii) For every $k'$ consecutive points $Y_i = \{x_i, \ldots, x_{i+k'-1}\}$ with $i \in [1, k-k'+1]$, compute its discrepancy $\phi(Y_i)$. (iii) Set $\widehat{Y}_p = \mathrm{argmin}_{Y_i, i\in[1,k-k']}\,\phi(Y_i)$, and return $\mu(\widehat{Y}_p)$ as $\widehat{z}_p$.

In the *discrepancy-based denoising* approach, we simply set $\widehat{f}(p) := \widehat{z}_p$ as computed above. The correctness of $\hat{f}$ to approximate $f$ is given by the following Lemma.

▶ **Lemma 3.1.** *If $\tilde{f} : P \to \mathbb{R}$ is a $(k, k', \Delta)$-functional-sample of $f : \mathsf{M} \to \mathbb{R}$ with $k' \geq \frac{k}{2}$, then we have $|\widehat{f}(p) - f(p)| \leq \left(1 + 2\sqrt{\frac{k-k'}{2k'-k}}\right)\Delta$ for any $p \in P$, where $\widehat{f}$ is the discrepancy-based denoising of $\tilde{f}$. In particular, if $k' \geq \frac{2}{3}k$, then $|\widehat{f}(p) - f(p)| \leq 3\Delta$ for any $p \in P$.*

**Proof.** Let $Y_\Delta = \{x \in \mathrm{NN}_P^k(p) : |\tilde{f}(x) - f(p)| \leq \Delta\}$ be the set of points in $\mathrm{NN}_P^k(p)$ whose observed function values are at most $\Delta$ distance away from $f(p)$. Since $\tilde{f}$ is a $(k, k', \Delta)$-functional-sample of $f$, it is clear that $|Y_\Delta| \geq k'$. Let $Y'_\Delta \subset Y_\Delta$ be a subset with $k'$ elements, $Y'_\Delta = \{x'_i\}_{i=1}^{k'}$. By the definitions of $Y_\Delta$ and $Y'_\Delta$, one can immediately check that $|\tilde{f}(x'_i) - \mu(Y'_\Delta)| \leq 2\Delta$ where $\mu(Y'_\Delta) = \frac{1}{k'}\sum_{i=1}^{k'}\tilde{f}(x'_i)$. This inequality then gives an upper bound of the discrepancy $\phi(Y'_\Delta)$,

$$\begin{array}{rcl} \phi(Y'_\Delta) & = & \frac{1}{k'}\sum_{i=1}^{k'}(\tilde{f}(x'_i) - \mu(Y'_\Delta))^2 \\ & \leq & \frac{1}{k'}\sum_{i=1}^{k'}(2\Delta)^2 \\ & = & 4\Delta^2 \end{array}.$$

Recall from the sliding window procedure that $\widehat{Y}_p = \mathrm{argmin}_{Y_i, i\in[1,k-k']}\,\phi(Y_i)$ and $\widehat{z}_p = \mu(\widehat{Y}_p)$. Denote $A_1 = \widehat{Y}_p \cap Y_\Delta$ and $A_2 = \widehat{Y}_p \setminus A_1$. Since $\tilde{f}$ is a $(k, k', \Delta)$-functional-sample of $f$, the size of $A_2$ is at most $k - k'$ and $|A_1| \geq 2k' - k$. If $|\widehat{z}_p - f(p)| \leq \Delta$, nothing

needs to be proved. Without loss of generality, one can assume that $f(p)+\Delta \leq \widehat{z}_p$. Denote $\delta = \widehat{z}_p - (f(p)+\Delta)$. The discrepancy of $\phi(\widehat{Y}_p)$ can be estimated as follows.

$$\begin{aligned}
\phi(\widehat{Y}_p) &= \tfrac{1}{k'}\left(\textstyle\sum_{x\in A_1}(\tilde{f}(x)-\widehat{z}_p)^2 + \sum_{x\in A_2}(\tilde{f}(x)-\widehat{z}_p)^2\right)\\
&\geq \tfrac{1}{k'}\left(|A_1|\delta^2 + \textstyle\sum_{x\in A_2}(\tilde{f}(x)-\widehat{z}_p)^2\right)\\
&\geq \tfrac{1}{k'}\left(|A_1|\delta^2 + \tfrac{1}{|A_2|}(\textstyle\sum_{x\in A_2}\tilde{f}(x) - |A_2|\widehat{z}_p)^2\right)\\
&= \tfrac{1}{k'}\left(|A_1|\delta^2 + \tfrac{1}{|A_2|}(\textstyle\sum_{x\in A_1}\tilde{f}(x) - |A_1|\widehat{z}_p)^2\right)\\
&\geq \tfrac{1}{k'}\left(|A_1|\delta^2 + \tfrac{1}{|A_2|}(|A_1|\delta)^2\right)\\
&\geq \tfrac{1}{k'}\delta^2\left(\tfrac{k'|A_1|}{|A_2|}\right)\\
&\geq \tfrac{2k'-k}{k-k'}\delta^2
\end{aligned}$$

where the third line uses the inequality $\sum_{i=1}^n a_i^2 \geq \frac{1}{n}(\sum_{i=1}^n a_i)^2$, and the fourth line uses the fact that $(|A_1|+|A_2|)\widehat{z}_p = \sum_{x\in\widehat{Y}_p}\tilde{f}(x)$. Since $\widehat{Y}_p = \operatorname{argmin}_{Y_i, i\in[1,k-k']}\phi(Y_i)$, it holds that $\phi(\widehat{Y}_p)\leq\phi(Y'_\Delta)$. Therefore,

$$\frac{2k'-k}{k-k'}\delta^2 \leq 4\Delta^2.$$

It then follows that $\delta \leq 2\sqrt{\frac{k-k'}{2k'-k}}\Delta$. Hence, $|\widehat{f}(p)-f(p)| \leq \left(1+2\sqrt{\frac{k-k'}{2k'-k}}\right)\Delta$ since $\widehat{z}_p = \widehat{f}(p)$. If $k' \geq \frac{2}{3}k$, then $1+2\sqrt{\frac{k-k'}{2k'-k}} \leq 1+2=3$, meaning that $|\widehat{f}(p)-f(p)|\leq 3\Delta$ in this case.

◀

▶ **Corollary 3.2.** *Given a $(k,k',\Delta)$-functional-sample of $f:\mathsf{M}\to\mathbb{R}$ with $k'\geq k/2$, we can compute a new function $\widehat{f}:P\to\mathbb{R}$ such that $|\widehat{f}(p)-f(p)|\leq\xi\Delta$ for any $p\in P$, where $\xi=1$ under $k$-median denoising, and $\xi=\left(1+2\sqrt{\frac{k-k'}{2k'-k}}\right)$ under the discrepancy-based denoising.*

Hence after the $k$-median denoising or the discrepancy-based denoising, we obtain a new function $\widehat{f}$ whose value at each sample point is within $\xi$ precision to the true function value. We can now apply the scalar field topology inference framework from [5] (as introduced in Section 2) using $\widehat{f}$ as input. In particular, set $L_\alpha = \{p\in P \mid \widehat{f}(p)\leq\alpha\}$, and let $R_\delta(X)$ denote the Rips complex over points in $X$ with parameter $\delta$. We approximate the persistence diagram induced by the sub-level sets filtration of $f:\mathsf{M}\to\mathbb{R}$ from the filtrations of nested pairs $\{R_\delta(L_\alpha)\hookrightarrow R_{2\delta}(L_\alpha)\}_\alpha$. It follows from Theorem 2.3 that:

▶ **Theorem 3.3.** *Let $\mathsf{M}$ be a compact Riemannian manifold and let $f:\mathsf{M}\to\mathbb{R}$ be a $c$-Lipschitz function. Let $P$ be a geodesic $\varepsilon$-sampling of $\mathsf{M}$, and $\tilde{f}:P\to\mathbb{R}$ a $(k,k',\Delta)$-functional-sample of $f$. Set $\xi=1$ if $P_\alpha$ is obtained via $k$-median denoising, and $\xi=\left(1+2\sqrt{\frac{k-k'}{2k'-k}}\right)$ if $P_\alpha$ is obtained via discrepancy-based denoising. If $\varepsilon<\frac{1}{4}\varrho(\mathsf{M})$, then for any $\delta\in\left[2\varepsilon,\frac{1}{2}\varrho(\mathsf{M})\right)$, the persistent homology modules of $f$ and the filtration of nested pairs $\{R_\delta(P_\alpha)\hookrightarrow R_{2\delta}(P_\alpha)\}$ are $(2c\delta+\xi\Delta)$-interleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta+\xi\Delta$.*

The above theoretical results are similar for $k$-median and discrepancy-based methods with a slight advantage for the $k$-median. However, interesting experimental results can be obtained when the Lipschitz condition on the function is removed, for example with images, where the discrepancy based method appear to be more resisilient to large amounts of noise, than the $k$-median denoising method. Illustrating examples can be found in Appendix C.

# 4 Geometric noise

In the previous section, we assumed that we have no geometric noise in the input. In this section, we deal with the case where there is only geometric noise in the input, but no functional noise of any kind. Specifically, for any point $p \in P$, we assume that the observed value $\tilde{f}(p)$ is equal to the true function value $f(\pi(p))$ where $\pi(p)$ is the orthogonal projection of $p$ to the manifold. If $p$ is on the medial axis of $\mathsf{M}$, the projection $\pi$ is arbitrary to one of the possible sites. As we have alluded before, general geometric noise implicitly introduces functional noise because the point $p$ might have become a functional aberration of its orthogonal projection $\pi(p)$. This error will be ultimately captured in Section 5 when we combine the results from the previous section on pure functional noise with the results in this section on pure geometric noise.

## 4.1 Noise model

### Distance to a measure.

The distance to a measure is a tool introduced to deal with geometrically noisy datasets, which are modelled as probability measures [4]. Given a probability measure $\mu$ we define the *pseudo-distance* $\delta_m(x)$ for any point $x \in \mathbb{R}^d$ and a mass parameter $m \in ]0,1]$ as $\delta_m(x) = \inf\{r \in \mathbb{R} | \mu(B(x,r)) \geq m\}$. The distance to a measure is then defined by averaging this quantity:

$$d_{\mu,m}(x) = \sqrt{\frac{1}{m}\int_0^m \delta_l(x)^2 \; dl}.$$

The *Wasserstein distance* is a standard tool to compare two measures. Given two probability measures $\mu$ and $\nu$ on a metric space $M$, a *transport plan* $\pi$ is a probability measure over $M \times M$ such that for any $A \times B \subset M \times M$, $\pi(A \times M) = \mu(A)$ and $\pi(M \times B) = \nu(B)$. Let $\Gamma(\mu,\nu)$ be the set of all transport plans between between measures $\mu$ and $\nu$. The Wassserstein distance is then defined as the minimum transport cost over $\Gamma(\mu,\nu)$:

$$W_2(\mu,\nu) = \sqrt{\min_{\pi \in \Gamma(\mu,\nu)} \int_{M\times M} d_M(x,y)^2 \; d\pi(x,y)},$$

where $d_M(x,y)$ is the distance between $x$ and $y$ in the metric space $M$. The distance to a measure is stable with respect to the Wasserstein distance as shown in [4]:

▶ **Theorem 4.1** (Theorem 3.5 of [4]). *Let $\mu$ and $\nu$ be two probability measures on $\mathbb{R}^d$ and $m \in ]0,1]$. Then, $||d_{\mu,m} - d_{\nu,m}||_\infty \leq \frac{1}{\sqrt{m}} W_2(\mu,\nu)$.*

We will mainly use the distance to empirical measures in this paper. (See [4] for more details on distance to a measure and its approximation.) Given a finite point set $P$, its associated *empirical measure* $\mu_P$ is defined as the sum of Dirac masses: $\mu_P = \frac{1}{|P|}\sum_{p\in P} \delta_p$. The distance to this empirical measure for a point $x$ can then be expressed as an average of its distances to the $k = m|P|$ nearest neighbors where $m$ is the parameter of mass. For the sake of simplicity, $k$ will be assumed to be an integer. The results also hold for other values of $k$ but the $k$-th nearest neighbor requires a specific treatment in every equation. Denoting by $p_i(x)$ the $i$-th nearest neighbors of $x$ in $P$, one can write:

$$d_{\mu_P,m}(x) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} d(p_i(x),x)^2}.$$

### Our geometric noise model.

Our noise model treats the input point data as a measure and relates it to the manifold (where input points are sampled from) via distance-to-measures with the help of two parameters.

▶ **Definition 4.2.** Let $P \subset \mathbb{R}^n$ be a discrete sample and $\mathsf{M} \subset \mathbb{R}^n$ a smooth manifold. Let $\mu$ denote the empirical measure of $P$. For a fixed mass parameter $m > 0$, we say that $P$ is an $(\varepsilon, r)$-sample of $\mathsf{M}$ if the following holds:

$$\forall x \in \mathsf{M}, d_{\mu,m}(x) \leq \varepsilon; \quad \text{and} \tag{3}$$

$$\forall x \in \mathbb{R}^n,\ d_{\mu,m}(x) < r \implies d(x, \mathsf{M}) \leq d_{\mu,m}(x) + \varepsilon. \tag{4}$$

The parameter $\varepsilon$ captures the distance to the empirical measure for points in $\mathsf{M}$ and intuitively tells us how dense $P$ is in relation to the manifold $\mathsf{M}$. The parameter $r$ intuitively indicates how far away we can deviate from the manifold, while keeping the noise sparse enough so as not to be mistaken for signal. We remark that if a point set is an $(\epsilon, r)$-sample of $\mathsf{M}$ then it is an $(\epsilon', r')$-sample of $\mathsf{M}$ for any $\epsilon' \geq \epsilon$ and $r' \leq r$. In general, the smaller $\varepsilon$ is and the bigger $r$ is, the better an $(\varepsilon, r)$-sample is.

For convenience, denote the distance function to the manifold $\mathsf{M}$ by $d_\pi : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto d(x, \mathsf{M})$. We have the following interleaving relation:

$$\forall \alpha < r - \varepsilon,\ d_\pi^{-1}(]-\infty, \alpha]) \subset d_{\mu,m}^{-1}(]-\infty, \alpha + \varepsilon]) \subset d_\pi^{-1}(]-\infty, \alpha + 2\varepsilon]) \tag{5}$$

To see why this interleaving relation holds, let $x$ be a point such that $d(x, \mathsf{M}) \leq \alpha$. Thus $d(\pi(x), x) \leq \alpha$. Using the hypothesis (3), we get that $d_{\mu,m}(\pi(x)) \leq \varepsilon$. Given that the distance to a measure is a 1-Lipschitz function we then obtain that $d_{\mu,m}(x) \leq \varepsilon + \alpha$.

Now let $x$ be a point such that $d_{\mu,m}(x) \leq \alpha + \varepsilon \leq r$. Using the condition on $r$ in (4) we get that $d(x, \mathsf{M}) \leq d_{\mu,m}(x) + \varepsilon \leq \alpha + 2\varepsilon$ which concludes the proof of Eqn (5).

Eqn (5) gives an interleaving between the sub-level sets of the distance to the measure $\mu$ and the offsets of the manifold $\mathsf{M}$. By Theorem 2.1, this implies the proximity between the persistence modules of their respective sub-level sets filtrations . Observe that this relation is in some sense analogous to the one obtained when two compact sets $A$ and $B$ have Hasudorff distance of at most $\varepsilon$:

$$\forall \alpha,\ d_A^{-1}(]-\infty, \alpha]) \subset d_B^{-1}(]-\infty, \alpha + \varepsilon]) \subset d_A^{-1}(]-\infty, \alpha + 2\varepsilon]). \tag{6}$$

### Relation to other noise models.

Our noise model encompasses several other existing noise models. While the parameter $\varepsilon$ is natural, the parameter $r$ may appear to be artificial. It bounds the distances at which we can observe the manifold through the scope of the distance to a measure. In most classical noise models, $r$ is equal to $\infty$ and thus we obtain a similar relation as for the classical Hausdorff noise model in Eqn (6).

One notable noise model where $r \neq \infty$ is when there is an uniform background noise in the ambient space $\mathbb{R}^d$, sometimes called *clutter noise*. In this case, $r$ will depend on the difference between the density of the relevant data and the density of the noise. For other noise models like Wassertein, Gaussian, Hausdorff noise models, $r$ equals to $\infty$. Detailed relations and proofs for the Wasserstein noise model can be found in Appendix B.

## 4.2 Scalar field analysis under geometric noise

In the rest of the paper, we assume that $\mathsf{M}$ is a manifold with positive reach $\rho_{\mathsf{M}}$ and whose curvature is bounded by $c_{\mathsf{M}}$. Assume that the input $P$ is an $(\varepsilon, r)$-sample of $\mathsf{M}$ for any value of $m$ satisfying the bound in Theorem 2.1, where

$$\varepsilon \leq \frac{\rho_{\mathsf{M}}}{6}\ , \text{ and } r > 2\varepsilon. \tag{7}$$

As discussed at the beginning of this section, we assume that there is no intrinsic functional noise in the sense that for any $p \in P$, the observed function value $\tilde{f}(p) = f(\pi(p))$ is the same as the true value for the projection $\pi(p) \in \mathsf{M}$ of this point. Our goal now is to show how to recover the persistence diagram induced by $f : \mathsf{M} \to \mathbb{R}$ from its observations $\tilde{f} : P \to \mathbb{R}$ on $P$.

Taking advantage of the interleaving (5), we can use the distance to the empirical measure to filter the points of $P$ to remove geometric noise. In particular, we consider the set

$$L = P \cap d_{\mu,m}^{-1}(]-\infty, \eta]) \text{ where } \eta \geq 2\epsilon. \tag{8}$$

We will then use a similar approach as the one from [5] for this set $L$. The optimal choice for the parameter $\eta$ is $2\epsilon$. However, any value with $\eta \leq r$ and $\eta + \epsilon < \rho_{\mathsf{M}}$ works as long as there exist $\delta$ and $\delta'$ satisfying the conditions stated in Theorem 2.4.

Let $\bar{L} = \{\pi(x) | x \in L\}$ denote the orthogonal projection of $L$ onto $\mathsf{M}$. To simulate sub-level sets $f^{-1}(]-\infty, \alpha])$ of $f : \mathsf{M} \to \mathbb{R}$, consider the restricted sets $L_\alpha := L \cap (f \circ \pi)^{-1}(]-\infty, \alpha])$ and let $\bar{L}_\alpha = \pi(L_\alpha)$. By our assumption on the observed function $\tilde{f} : P \to \mathbb{R}$, we have: $L_\alpha = \{x \in L | \tilde{f}(x) \leq \alpha\}$.

Let us first recall a result about the relation between Riemannian and Euclidian metrics [8]. For any two points $x, y \in \mathsf{M}$ with $d(x, y) \leq \frac{\rho_{\mathsf{M}}}{2}$ one has:

$$d(x, y) \leq d_{\mathsf{M}}(x, y) \leq \left(1 + \frac{4d(x, y)^2}{3\rho_{\mathsf{M}}^2}\right) d(x, y) \leq \frac{4}{3} d(x, y). \tag{9}$$

As a direct consequence of our noise model, for any point $x \in \mathsf{M}$, there exists a point $p \in L$ at distance less than $2\epsilon$: Indeed, for any $x \in \mathsf{M}$, since $d_{\mu,m}(x) \leq \varepsilon$, there must exist a point $p \in P$ such that $d(x, p) \leq \varepsilon$. On the other hand, since the distance to measure is 1-Lipschitz, we have $d_{\mu,m}(p) \leq d_{\mu,m}(x) + d(x, p) \leq 2\varepsilon$. Hence $p \in L$ as long as $\eta \geq 2\epsilon$. We will use the *extrinsic* Vietoris-Rips complex built on top points from $L$ to infer the scalar field topology. Using the previous relation Eqn (9), we obtain the following result which states that for points in $L$, the Euclidean distance for nearby points approximates the Riemannian metric on $\mathsf{M}$.

▶ Proposition 4.3. Let $\lambda = \frac{4}{3} \frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta + \epsilon)}$, and assume that $2\varepsilon \leq \eta \leq r$ and $\epsilon + \eta < \rho_{\mathsf{M}}$. Let $x, y \in L$ be two points from $L$ such that $d(x, y) \leq \frac{\rho_{\mathsf{M}}}{2} - \frac{\eta + \varepsilon}{2}$. Then,

$$\frac{d_{\mathsf{M}}(\pi(y), \pi(x))}{\lambda} \leq d(x, y) \leq 2(\eta + \varepsilon) + d_{\mathsf{M}}(\pi(x), \pi(y)).$$

**Proof.** Let $x$ and $y$ be two points of $L$ such that $d(x, y) \leq \frac{\rho_{\mathsf{M}}}{2} - \frac{\eta + \varepsilon}{2}$. As $d_{\mu,m}(x) \leq \eta \leq r$, Eqn (4) implies $d(\pi(x), x) \leq \eta + \epsilon$. Therefore $d(\pi(x), \pi(y)) \leq \frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta + \epsilon)} d(x, y)$ [11, Theorem 4.8,(8)]. This implies $d(\pi(x), \pi(y)) \leq \frac{\rho_{\mathsf{M}}}{2}$ and following (9), $d_{\mathsf{M}}(\pi(x), \pi(y)) \leq \frac{4}{3} d(\pi(x), \pi(y))$.

This proves the left inequality in the Proposition. The right inequality follows from

$$d(x, y) \leq d(\pi(x), x) + d(\pi(y), y) + d_{\mathsf{M}}(\pi(x), \pi(y)) \leq 2(\eta + \varepsilon) + d_{\mathsf{M}}(\pi(x), \pi(y)).$$

◀

▶ **Theorem 4.4.** *Let* $\mathsf{M}$ *be a compact Riemannian manifold and let* $f : \mathsf{M} \to \mathbb{R}$ *be a* $c$*-Lipschitz function. Let* $P$ *be an* $(\varepsilon, r)$*-sample of* $M$*, and* $L$ *introduced in Eqn (8). Assume* $\varepsilon \leq \frac{\rho_{\mathsf{M}}}{6}, r > 2\varepsilon$*, and* $2\varepsilon \leq \eta \leq r$*. Then, for any* $\delta \geq 2\eta + 6\varepsilon$ *and any* $\delta' \in \left[2\eta + 2\varepsilon + \frac{8}{3}\frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta+\varepsilon)}\delta,\ \frac{3}{4}\frac{\rho_{\mathsf{M}} - (\eta+\varepsilon)}{\rho_{\mathsf{M}}}\varrho(\mathsf{M})\right]$*,* $H_*(f)$ *and* $H_*(R_\delta(L_\alpha) \hookrightarrow R_{\delta'}(L_\alpha))$ *are* $\frac{4}{3}\frac{c\rho_{\mathsf{M}}\delta'}{\rho_{\mathsf{M}} - (\eta+\varepsilon)}$*-interleaved.*

**Proof.** First, note that $\bar{L}$ is a $2\epsilon$-samle of $\mathsf{M}$ in its Riemannian metric. This is because that for any point $x \in \mathsf{M}$, we know that there exists some $p \in L$ such that $d(x,p) \leq d_{\mu,m}(x) \leq \varepsilon$. Hence $d(x, \pi(p)) \leq d(x,p) + d(p, \pi(x)) \leq 2d(x,p) \leq 2\varepsilon$. Now we apply Theorem 2.4 to $\bar{L}$ by using $\tilde{d}(\pi(x), \pi(y)) := d(x,y)$; and setting $\lambda = \mu = \frac{4}{3}\frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta+\varepsilon)}$, $\nu = 2(\eta+\varepsilon)$: the requirement on the distance function $\tilde{d}$ in Theorem 2.4 is satisfied due to Proposition 4.3. The claim then follows. ◀

Since $\mathsf{M}$ is compact, $f$ is bounded due to the Lipschitz condition. We can look at the limit when $\alpha \to \infty$. There exists a value $T$ such that for any $\alpha \geq T$, $L_\alpha = L$ and $f^{-1}(]-\infty, \alpha]) = \mathsf{M}$. The above interleaving means that $H_*(\mathsf{M})$ and $H_*(R_\delta(L)) \hookrightarrow R_{\delta'}(L))$ are interleaved. However, both objects do not depend on $\alpha$ and this gives the following inference result:

▶ **Corollary 4.5.** $H_*(\mathsf{M})$ *and* $H_*(R_\delta(L)) \hookrightarrow R_{\delta'}(L))$ *are isomorphic under conditions specified in Theorem 4.4.*

## 5 Scalar Field Topology Inference under Geometric and Functional Noise

Our constructions can be combined to analyze scalar fields in a more realistic setting. Our *combined noise model* follows conditions (3) and (4) for the geometry. We adapt condition (2) to take into account the geometry and we assume that there exist $\eta \geq 2\varepsilon$ and $s$ such that:

$$\forall p \in d_{\mu,m}^{-1}(]-\infty, \eta, ]),\ |\{q \in NN_k(p) |\ |\tilde{f}(q) - f(\pi(p))| \leq s\}| \geq k' \tag{10}$$

Note that in (10), we are using $f(\pi(p))$ as the "true" function value at a sample $p$ which is off the manifold $\mathsf{M}$. The condition on the functional noise is only for points close to the manifold (under the distance to a measure). Combining the methods from the previous two sections, we obtain the *combined noise algorithm* where $\eta$ is a parameter greater than $2\epsilon$.

We propose the following 3-steps algortihm. It starts by handling outliers in the geometry then it makes a regression on the function values to obtain a smoothed function $\hat{f}$ before running the existing algorithm for scalar field analysis [5] on the filtration $\hat{L}_\alpha = \{p \in L | \hat{f}(p) \leq \alpha\}$.

**Combined noise algorithm**

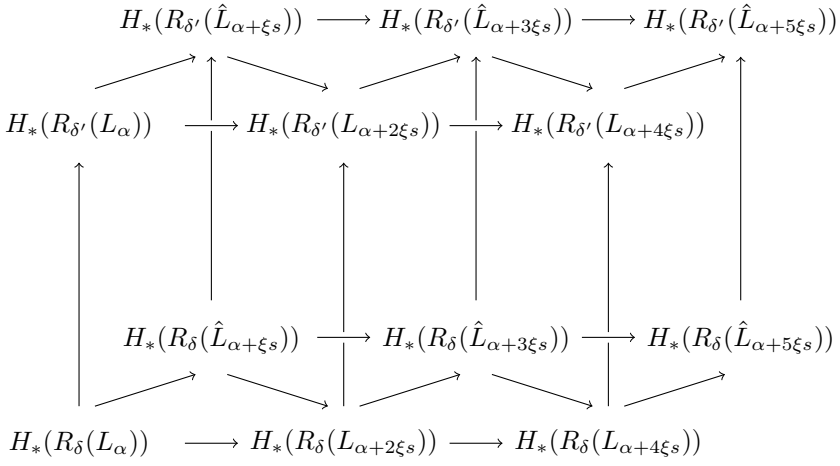
1. Compute $L = P \cap d_{\mu,m}^{-1}(]-\infty, \eta])$.
2. Replace functional values $\tilde{f}$ by $\hat{f}$ for points in $L$ using either k-median or discrepancy based method.
3. Run the scalar field analysis algorithm from [5] on $(L, \hat{f})$.

▶ **Theorem 5.1.** *Let* $\mathsf{M}$ *be a compact smooth manifold embedded in* $\mathbb{R}^d$ *and* $f$ *a* $c$*-Lipschitz function on* $\mathsf{M}$*. Let* $P \subset \mathbb{R}^d$ *be a point set and* $\tilde{f} : P \to \mathbb{R}$ *observed function values such that hypotheses (3), (4), (7) and (10) are satisfied. For* $\eta \geq 2\epsilon$*, the combined noise algorithm has the following guarantees:*

*For any* $\delta \in \left[2\eta + 6\epsilon, \frac{\varrho(\mathsf{M})}{2}\right]$ *and any* $\delta' \in \left[2\eta + 2\epsilon + \frac{8}{3}\frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta+\epsilon)}\delta, \frac{3}{4}\frac{\rho_{\mathsf{M}} - (\eta+\epsilon)}{\rho_{\mathsf{M}}}\varrho(\mathsf{M})\right]$*,* $H_*(f)$ *and* $H_*(R_\delta(\hat{L}_\alpha) \hookrightarrow R_{\delta'}(\hat{L}_\alpha))$ *are* $\left(\frac{4}{3}\frac{c\rho_{\mathsf{M}}\delta'}{\rho_{\mathsf{M}} - (\eta+\epsilon)} + \xi s\right)$*-interleaved where* $\xi = 1$ *if we use the* $k$*-median and* $\xi = \left(1 + 2\sqrt{\frac{k-k'}{2k'-k}}\right)$ *if we use the discrepancy method for Step 2.*

**Proof.** First, consider the filtration induced by $L_\alpha = \{x \in L | f(\pi(x)) \leq \alpha\}$; that is, we first imagine that all points in $L$ have correct function value (equals to the true value of their projection on $\mathsf{M}$). By Theorem 4.4, for $\delta \in \left[2\eta + 6\epsilon, \frac{\varrho(\mathsf{M})}{2}\right]$ and $\delta' \in \left[2\eta + 2\epsilon + \frac{8}{3}\frac{\rho_{\mathsf{M}}}{\rho_{\mathsf{M}} - (\eta+\epsilon)}\delta, \frac{3}{4}\frac{\rho_{\mathsf{M}} - (\eta+\epsilon)}{\rho_{\mathsf{M}}}\varrho(\mathsf{M})\right]$, $H_*(f)$ and $H_*(R_\delta(L_\alpha) \hookrightarrow R_{\delta'}(L_\alpha))$ are $\frac{4}{3}\frac{c\rho_{\mathsf{M}}\delta'}{\rho_{\mathsf{M}} - (\eta+\epsilon)}$-interleaved.

Next, consider $\hat{L}_\alpha = \{p \in L | \hat{f}(p) \leq \alpha\}$, which leads to a filtration based on the smoothed function values $\hat{f}$ (not observed values). Recall that our algorithm returns $H_*(R_\delta(\hat{L}_\alpha) \hookrightarrow R_{\delta'}(\hat{L}_\alpha))$. We aim to relate this persistence module with $H_*(R_\delta(L_\alpha) \hookrightarrow R_{\delta'}(L_\alpha))$. Specifically, fix $\alpha$ and let $(x, y)$ be an an edge of $R_\delta(L_\alpha)$. This means that $d(x, y) \leq 2\delta$, $f(\pi(x)) \leq \alpha$, $f(\pi(y)) \leq \alpha$. Corollary 3.2 can be applied to the function $f \circ \pi$ due to hypothesis (10). Hence $|\hat{f}(x) - f(\pi(x))| \leq \xi s$ and $|\hat{f}(y) - f(\pi(y))| \leq \xi s$. Thus $(x, y) \in R_\delta(\hat{L}_{\alpha+\xi s})$. One can reverse the role of $\hat{f}$ and $f$ and get an $\xi s$-interleaving of $\{R_\delta(L_\alpha)\}$ and $\{R_\delta(\hat{L}_\alpha)\}$. This gives rise to the following commutative diagram since all arrows are induced by inclusions.




Thus the two persistence modules induced by filtrations of nested pairs $\{R_\delta(L_\alpha) \hookrightarrow R_{\delta'}(L_\alpha)\}$ and $\{R_\delta(\hat{L}_\alpha) \hookrightarrow R_{\delta'}(\hat{L}_\alpha)\}$ are $\xi s$-interleaved. Combining this with the interleaving between $H_*(R_\delta(L_\alpha) \hookrightarrow R_{\delta'}(L_\alpha))$ and $H_*(f)$, the theorem follows. ◀

We note that while this theorem assumes a setting where we can ensure theoretical guarantees, the algorithm can be applied in a more general setting and still produce good results.

## References


1. Dana Angluin and Leslie G Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and system Sciences*, 18(2):155–193, 1979.
2. F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules, 2013. arXiv:1207.3674.
3. Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th ACM Sympos. on Comput. Geom.*, pages 237–246, 2009.
4. Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.
5. Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743–775, 2011.
6. Frédéric Chazal and Steve Yann Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, pages 232–241. ACM, 2008.
7. David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.
8. Tamal K Dey, Jian Sun, and Yusu Wang. Approximating cycles in a shortest basis of the first homology group from point data. *Inverse Problems*, 27(12):124004, 2011.
9. Yiqiu Dong and Shufang Xu. A new directional weighted median filter for removal of random-valued impulse noise. *Signal Processing Letters, IEEE*, 14(3):193–196, 2007.
10. H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Amer. Math. Soc., Providence, Rhode Island, 2009.
11. Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, pages 418–491, 1959.
12. Alfred Gray. The volume of a small geodesic ball of a riemannian manifold. *The Michigan Mathematical Journal*, 20(4):329–344, 1974.
13. László Györfi. *A distribution-free theory of nonparametric regression*. Springer, 2002.
14. Jennifer Kloke and Gunnar Carlsson. Topological de-noising: Strengthening the topological signal. *arXiv preprint arXiv:0910.5947*, 2009.
15. Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *arXiv preprint arXiv:1110.4300*, 2011.
16. Ching-Ta Lu and Tzu-Chun Chou. Denoising of salt-and-pepper noise corrupted image using modified directional-weighted-median filter. *Pattern Recognition Letters*, 33(10):1287–1295, 2012.
17. Shuenn-Shyang Wang and Cheng-Hao Wu. A new impulse detection and filtering method for removal of wide range impulse noises. *Pattern Recognition*, 42(9):2194–2202, 2009.
18. Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

# A Relations between our functional noise model and classical noise models

## Bounded noise model.

The standard "bounded noise" model assumes that all observed function values are within some $\delta$ distance away from the true function values: that is, $|\tilde{f}(p) - f(p)| \leq \delta$ for all $p \in P$. Hence this bounded noise model simply corresponds to a $(1, 1, \delta)$-functional-sample.

## Gaussian noise model.

Under the popular Gaussian noise model, for any $x \in \mathsf{M}$, its observed function value $\tilde{f}(x)$ is drawn from a normal distribution $\mathcal{N}(f(x), \sigma)$, that is a probability measure with density $g(y) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{(y-f(x))^2}{\sigma^2}}$. We say that a point $q \in P$ is $a$-accurate if $|\tilde{f}(q) - f(q)| \leq a$. For the Gaussian noise model, we will first bound the quantity $\mu(k, k')$ defined as the smallest value such that at least $k'$ out of the $k$ nearest neighbors of $p$ in $\mathrm{NN}^k_P(p)$ are $\mu(k, k')$-accurate. We claim the following statement.

▶ Claim 1.1. With probability at least $1 - e^{-\frac{k-k'}{6}}$, $\mu(k, k') \leq \sigma\sqrt{\ln \frac{2k}{k-k'}}$.

**Proof.** First note that for $\frac{b}{\sigma} \geq 1$, we have that:

$$\int_b^{+\infty} e^{-\frac{t^2}{\sigma^2}} dt \leq \int_b^{+\infty} \frac{t}{\sigma} e^{-\frac{t^2}{\sigma^2}} dt = \frac{1}{\sigma} \int_b^{+\infty} t e^{-\frac{t^2}{\sigma^2}} dt = -\frac{\sigma}{2} e^{-\frac{t^2}{\sigma^2}} \Big|_b^{\infty} = \frac{\sigma}{2} e^{-\frac{b^2}{\sigma^2}}.$$

Now we introduce $I(a) = \frac{1}{\sigma\sqrt{\pi}} \int_{-a}^{a} e^{-\frac{x^2}{\sigma^2}} dx$. Since $\frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{\sigma^2}} dx = 1$, we thus obtain that for $a \geq \sigma$:

$$1 - \frac{1}{\sqrt{\pi}} e^{-(\frac{a}{\sigma})^2} < 1 - e^{-(\frac{a}{\sigma})^2} \leq I(a) \; (= 1 - \frac{2}{\sigma\sqrt{\pi}} \int_a^{+\infty} e^{-\frac{x^2}{\sigma^2}} dx). \tag{11}$$

Now set $\delta = \frac{k-k'}{k} \leq \frac{1}{2}$ and $s = \sigma\sqrt{\ln \frac{2k}{k-k'}} \geq \sigma$. Let $p_1, \ldots, p_k$ denote the $k$ nearest neighbors of some point, say $p_1$. For each $p_i$, let $Z_i = 1$ if $p_i$ is **not** $s$-accurate, and $Z_i = 0$ otherwise. Hence $Z = \sum_{i=1}^{k} Z_i$ denotes the total number of points from these $k$ nearest neighbors that are not $s$-accurate. By Equation (11), we know that

$$\mathrm{Prob}[Z_i = 1] = 1 - I(s) \leq e^{-(\frac{s}{\sigma})^2}.$$

It then follows that the expected value of $Z$ satisfies:

$$E(Z) \leq k e^{-(\frac{s}{\sigma})^2} = \frac{\delta k}{2}.$$

Now set $\rho = \frac{\delta k}{2E(Z)}$. Since $E(Z) \leq \frac{\delta k}{2}$, it follows that $(1 + \rho)E(Z) \leq \delta k$. Using Chernoff's bound [1], we obtain

$$\begin{aligned} \mathrm{Prob}\,[Z \geq k - k'] = \mathrm{Prob}\,[Z \geq \delta k] &\leq \mathrm{Prob}\,[Z \geq (1+\rho)E(Z)] \\ &\leq e^{-\frac{\rho^2 E(Z)}{2+\rho}} = e^{-\frac{\delta^2 k^2}{4E(Z)} \cdot \frac{1}{2 + \frac{\delta k}{2E(Z)}}} \leq e^{-\frac{\delta^2 k^2}{6\delta k}} = e^{-\frac{k-k'}{6}}. \end{aligned}$$

The claim then follows, that is, with probability at least $1 - e^{-\frac{k-k'}{6}}$, at least $k'$ number of points out of any $k$ points are $s = \sigma\sqrt{\ln \frac{2k}{k-k'}} \geq \sigma$-accurate. ◀

Next, we convert the value $\mu(k,k')$ to the value $\Delta$ as in Equation (2). In particular, being a $(k,k',\Delta)$-functional-sample means that for any $p \in P$, there are at least $k'$ samples $q$ from $\mathrm{NN}_P^k(p)$ such that $|\tilde{f}(q) - f(p)| \leq \Delta$. Now assume that the furthest geodesic distance from any point in $\mathrm{NN}_P^k(p)$ to $p$ is $\lambda$. Then since $f$ is a $c$-Lipschitz function, we have $\max_{q \in \mathrm{NN}_P^k(p)} |f(q) - f(p)| \leq c\lambda$.

We note that Claim 1.1 is valid for any point $p$ of $P$. Using the union bound, the relation holds for all points in $P$ with probability at least $1 - ne^{-\frac{k-k'}{6}}$. Note that if $k - k' \geq 12 \ln n$, then this probability is at least $1 - \frac{1}{n}$, that is, the relation holds with high probability. Thus, with probability at least $1 - ne^{-\frac{k-k'}{6}}$, the input function $\tilde{f} : P \to \mathbb{R}$ under Gaussian noise model is a $(k,k',\Delta)$-functional-sample with $\Delta = \sigma\sqrt{\ln \frac{2k}{k-k'}} + c\lambda$.

## B Relations between our geometric noise model and the Wasserstein noise model

The Wasserstein noise model assumes that the empirical measure $\mu = \mu_P$ for $P$ is close to the uniform measure $\mu_{\mathsf{M}}$ on $\mathsf{M}$ under the Wasserstein distance. Let $\mathsf{M}$ be a $d'$-Riemannian manifold whose curvature is bounded from above by $c_{\mathsf{M}}$ and has a positive strong convexity radius $\varrho(\mathsf{M})$. Let $V_{\mathsf{M}}$ denote the volume of $\mathsf{M}$. Writing, $\Gamma$ the Gamma function, let us set $\mathcal{C}_{d'}^{c_{\mathsf{M}}}$ to be the following constant:

$$\mathcal{C}_{d'}^{c_{\mathsf{M}}} = \frac{4}{d'}\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} \left(\frac{\sqrt{c_{\mathsf{M}}}}{\pi}\right)^{d'-1}, \tag{12}$$

▶ **Theorem 2.1.** *Let $P$ be a set of points whose empirical measure $\mu$ satisfies $W_2(\mu, \mu_{\mathsf{M}}) \leq \sigma$, where $\mu_{\mathsf{M}}$ is the uniform measure on $\mathsf{M}$. Then, for any $m \leq \frac{\mathcal{C}_{d'}^{c_{\mathsf{M}}} \left(\frac{\pi}{c_{\mathsf{M}}}\right)^{d'}}{V_{\mathsf{M}}}$, $P$ is an $(\varepsilon, r)$-sample under our noise model for*

$$\varepsilon \geq \frac{1}{\sqrt{1 + \frac{2}{d'}}} \left(\frac{mV_{\mathsf{M}}}{\mathcal{C}_{d'}^{c_{\mathsf{M}}}}\right)^{\frac{1}{d'}} + \frac{\sigma}{\sqrt{m}}, \quad \text{and} \quad r = \infty.$$

**Proof.** Fixing a point $x \in \mathsf{M}$, we can lower bound the volume of the Riemannian ball of radius $a$, centered at $x$, using the Günther-Bishop Theorem:

▶ **Theorem 2.2** (Günther-Bishop). *Assuming that the sectional curvature of a manifold $\mathsf{M}$ is always less than $c_{\mathsf{M}}$ and $a$ is less than the strong convexity radius of $\mathsf{M}$, then for any point $x \in \mathsf{M}$, the volume $\mathcal{V}(x,a)$ of the geodesic ball centred on $x$ and of radius $a$ is greater than $V_{d'}^{c_{\mathsf{M}}}(a)$ where $d'$ is the intrinsic dimension of $\mathsf{M}$ and $V_{d'}^{c_{\mathsf{M}}}(a)$ is the volume of the Riemannian ball of radius $a$ on a surface with constant curvature $c_{\mathsf{M}}$.*

We explicitly bound the value of $\mathcal{V}(x,a)$, with the following technical lemma:

▶ **Lemma 2.3.** *Let $\mathsf{M}$ be a Riemannian manifold with curvature upper bounded by $c_{\mathsf{M}}$, then for any $x \in \mathsf{M}$ and $a \leq \min(\varrho(\mathsf{M}); \frac{\pi}{\sqrt{c_{\mathsf{M}}}})$, the volume $\mathcal{V}(x,a)$ of the geodesic ball centred at $x$ and of radius $a$ verifies:*

$$\mathcal{V}(x,a) \geq \mathcal{C}_{d'}^{c_{\mathsf{M}}} a^{d'}$$

*where $\mathcal{C}_{d'}^{c_{\mathsf{M}}}$ is a constant independent of $x$ and $a$.*

**Proof.** Given $a \leq \min(\varrho(\mathsf{M}), \frac{\pi}{\sqrt{c_\mathsf{M}}})$, we want to bound the volume $V_{d'}^{c_\mathsf{M}}(a)$. Consider the sphere of dimension $d'$ and curvature $c_\mathsf{M}$. The surface $S_{c_\mathsf{M}}^{d'-1}$ of the border of a ball of radius $a \leq \frac{\pi}{\sqrt{c_\mathsf{M}}}$ on this sphere is given by [12]:

$$S_{c_\mathsf{M}}^{d'-1}(a) = 2\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} c_\mathsf{M}^{-\frac{1}{2}(d'-1)} \sin^{d'-1}(c_\mathsf{M} a)$$

We can bound the value of $V_{d'}^{c_\mathsf{M}}(a)$ :

$$\begin{aligned}
V_{d'}^{c_\mathsf{M}}(a) &= \int_0^a S^{d'-1}(l)dl \\
&= \int_0^a 2\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} c_\mathsf{M}^{-\frac{1}{2}(d'-1)} \sin^{d'-1}(c_\mathsf{M} l)dl \\
&\geq 2\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} c_\mathsf{M}^{-\frac{1}{2}(d'-1)} 2\int_0^{\frac{a}{2}} \left(\frac{2c_\mathsf{M} l}{\pi}\right)^{d'-1} dl \\
&= 4\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} c_\mathsf{M}^{-\frac{1}{2}(d'-1)} \frac{\pi}{2c_\mathsf{M}} \int_0^{\frac{c_\mathsf{M} a}{\pi}} u^{d'-1} du
\end{aligned}$$

Writing

$$\mathcal{C}_{d'}^{c_\mathsf{M}} = \frac{4}{d'}\Gamma\left(\frac{1}{2}\right)^{d'} \Gamma\left(\frac{d'}{2}\right)^{-1} \left(\frac{\sqrt{c_\mathsf{M}}}{\pi}\right)^{d'-1},$$

and using the Günther-Bishop Theorem, we have for any $a \leq \min(\varrho(\mathsf{M}); \frac{\pi}{\sqrt{c_\mathsf{M}}})$ and any $x \in \mathsf{M}$,

$$\mathcal{V}(x, a) \geq \mathcal{C}_{d'}^{c_\mathsf{M}} a^{d'}.$$

489 ◀

We next prove that the empirical measure $\mu$ of $P$ satisfies the two conditions in Eqns (3) and (4) for the value of $\varepsilon$ and $r$ specified in Theorem 2.1. Specifically, recall that $\mu_\mathsf{M}$ be the uniform measure on $\mathsf{M}$ and $\mu$ is a measure such that $W_2(\mu, \mu_\mathsf{M}) \leq \sigma$. Now consider a point $x \in \mathsf{M}$ and the Euclidean ball $B(x, a)$ centred in $x$ and of radius $a$. By definition of $\mu_\mathsf{M}$, for any $a \leq \frac{\pi}{c_\mathsf{M}}$:

$$\mu_\mathsf{M}(B(x, a)) = \frac{\mathcal{V}ol(x, a)}{V_\mathsf{M}} \geq \frac{\mathcal{C}_{d'}^{c_\mathsf{M}} a^{d'}}{V_\mathsf{M}}$$

By the definition of the pseudo-distance $\delta_m(x)$, we can then bound it, for any $m \leq \frac{\mathcal{C}_{d'}^{c_\mathsf{M}} \left(\frac{\pi}{c_\mathsf{M}}\right)^{d'}}{V_\mathsf{M}}$, as follows:

$$\delta_m(x) \leq \left(\frac{m\, V_\mathsf{M}}{\mathcal{C}_{d'}^{c_\mathsf{M}}}\right)^{\frac{1}{d'}}.$$

This in turn produces an upper bound on the distance to the measure $\mu_\mathsf{M}$:

$$d_{\mu_\mathsf{M},m}(x) \leq \frac{1}{\sqrt{m}} \sqrt{\int_0^m \left(\frac{V_\mathsf{M}}{\mathcal{C}_{d'}^{c_\mathsf{M}}} l\right)^{\frac{2}{d'}} dl} \leq \frac{1}{\sqrt{1+\frac{2}{d'}}} \left(\frac{V_\mathsf{M} m}{\mathcal{C}_{d'}^{c_\mathsf{M}}}\right)^{\frac{1}{d'}}$$

By Theorem 4.1, it then follows that for any $x \in \mathsf{M}$:

$$d_{\mu,m}(x) \leq \frac{1}{\sqrt{1+\frac{2}{d'}}} \left(\frac{V_\mathsf{M} m}{\mathcal{C}_{d'}^{c_\mathsf{M}}}\right)^{\frac{1}{d'}} + \frac{\sigma}{\sqrt{m}}$$

The first part of our noise model (i.e., Eqn (3)) is hence verified for any $\epsilon \geq \frac{1}{\sqrt{1+\frac{2}{d'}}} \left( \frac{V_{\mathsf{M}} m}{\mathcal{C}^{\epsilon}_{\mathsf{M}}} \right)^{\frac{1}{d'}} + \frac{\sigma}{\sqrt{m}}$. Moreover, for any $x \in \mathbb{R}^d$, $d_{\mu_{\mathsf{M}},m}(x) \geq d(x,\mathsf{M})$ because $\mathsf{M}$ is the support of $\mu_{\mathsf{M}}$. Thus:

$$d(x,\mathsf{M}) \leq d_{\mu_{\mathsf{M}},m}(x) \leq d_{\mu,m}(x) + \frac{\sigma}{\sqrt{m}} \leq d_{\mu,m}(x) + \epsilon$$

holds with no constraints on the value of $d_{\mu,m}(x)$. That is, for $r = \infty$, $\mu$ verifies the second part of our noise model (Eqn (4). This completes the proof of Theorem 2.1. ◀

## C Experimental illustration for functional noise

Here, we present results obtained by applying our methods to cases where there is only functional noise. Our goals are to demonstrate the denoising power of both the $k$-median and the discrepancy-based approaches and to illustrate the differences between the practical performances of the $k$-median and discrepancy-based denoising methods. We compare our denoising results with the popular k-NN algorithm, which simply sets the function at point $p$ to be the mean of the observed function values of its $k$ nearest neighbours. Note that, when $k' = k$, our discrepancy-based method is equivalent to the k-NN algorithm.

Going back to the bone example from section 3.1, we apply our algorithm to the 10-nearest neighbours and $k' = 8$. Using 100 sampling of the Bone with 1000 points each, we compute the average maximal error made by the various methods. The discrepancy-based method commits a maximal error of 10% on average, while the median-based method recovers the values with an error of 2% and the simple $k$-NN regression gives a maximal error of 16%, with most error concentrated around the neck region, see Figure 2. These results translate into the persistence diagrams that are more robust with the use of the discrepancy (blue squares) or the $k$-median (red diamond) instead of the $k$-NN regression (green circles), see Figure 3. Both methods retrieve the 1-dimensional topological feature. The $k$-NN regression keeps some prominent 0-dimensional feature through the diagram instead of having a unique component, result obtained by using the discrepancy or the median. The persistence diagram of the original bone is given in red and contains only one feature.

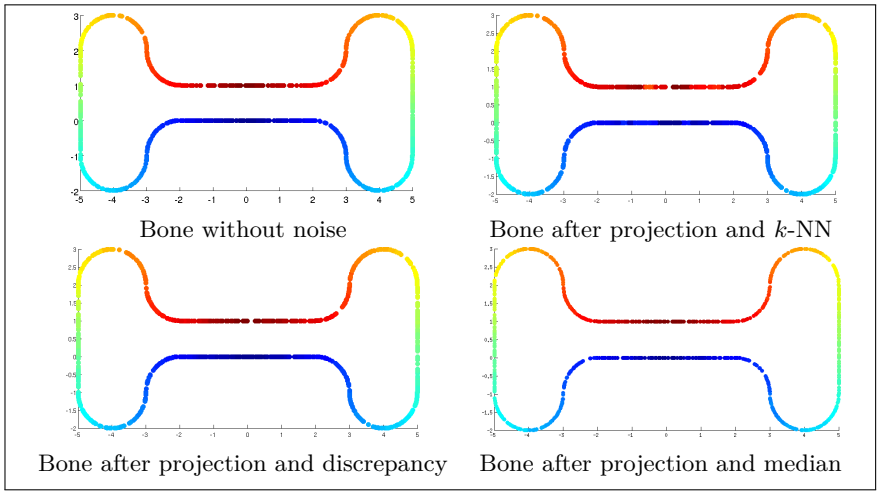



**Figure 2** Bone example after applying Gaussian perturbation, magical filter and a regression

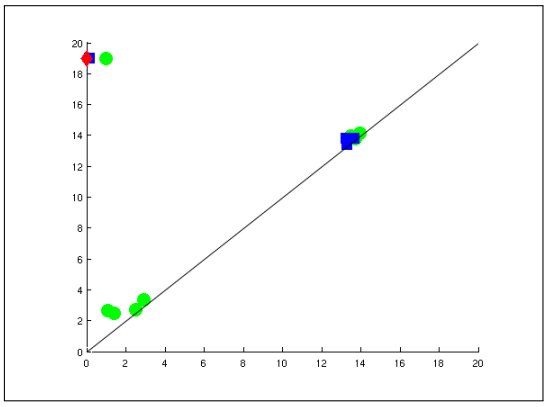



**Figure 3** Persistence diagrams in dimension 0 for the Bone example: red, green and blue points constitute the 0-th persistence diagram produced from clean (noise-less) data, from the denoised data by using $k$-NN regression, and from the denoised data by using discrepancy method, respectively.

As indicated by the theoretical results, the discrepancy-based method improves the classic $k$-NN regression but the median-based algorithm performs slightly better. The discrepancy however displays a better empirical behaviour when the Lipschitz condition on the input scalar field is relaxed, and/or the amount of noise becomes large. Additional illustrations can be found in the appendix.

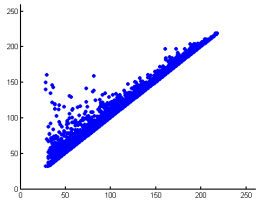
## Image denoising

We use a practical application: image denoising. We take the greyscale image Lena as the target scalar field $f$. In Figure 4, we use two ways to generate a noisy input scalar field $\tilde{f}$. The first type of noisy input is generated by adding uniform random noise as follows: with probability $p$, each pixel will receive a uniformly distributed random value in range $[0, 255]$ as its function value; otherwise, it is unchanged. Results under random noises are in the second and third rows of Figure 4. We also consider what we call *outlier noise*: with probability $p$, each pixel will be a outlier meaning that its function value is a fixed constant, which is set to be 200 in our experiments. This outlier noise is to simulate the aberrant function values caused by say, a broken sensor. The denoising results under the outlier-noise are shown in the last row of Figure 4.

First, we note that kNN approach tends to smooth out function values. In addition to the blurring artifact, its denoising capability is limited when the amount of noise is high (where imprecise values become dominant). As expected, both k-median and discrepancy based methods outperform the kNN approach. Indeed, they demonstrate robust recovery of the input image even with 50% amount of random noise are added.

While both k-median and discrepancy based methods are more resilient against noise, there are interesting difference between their practical performances. From a theoretical point of view, when the input scalar field is indeed a $(k, k', \Delta)$-functional-sample, k-median method gives a slightly better error bound (Observation 1) as compared to the discrepancy based method (Lemma 3.1). However, when $(k, k', \Delta)$-sampling condition is not satisfied, the median value can be quite arbitrary. By taking the average of a subset of points, the discrepancy method, on the other hand, is more robust against large amount of noise. This difference is evident in the third and last row of Figure 4.

Original Lena



The 0-th persistence diagram

10% random noise

kNN: $k = 9$

k-median, $k = 9$

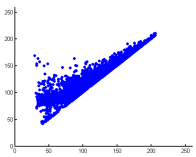
discrepancy, $k = 9, k' = 5$

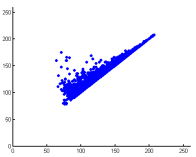
50% random noise

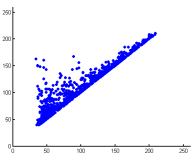
kNN: $k = 25$

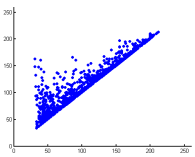
k-median, $k = 25$

discrepancy, $k = 25, k' = 13$









40% outlier noise

kNN: $k = 25$

k-median, $k = 25$

discrepancy, $k = 25, k' = 13$

**Figure 4** The denoised images after kNN, k-median, and discrepancy denoising approaches. The first row shows the original image and its 0-th persistence diagram. Second and third rows are under random noise of input, while fourth row are under outlier-noise as described in the text. The fifth row provides the 0-th persistence diagrams on images in the fourth row, which are computed by the scalar field analysis algorithm from [5] .

Moreover, the application to persistent homology which was our primary goal is much cleaner after the discrepancy-based method. The structure of the beginning of the diagrams is almost perfectly retrieved by both the median and discrepancy-based methods. However, the median induces a shrinking phenomenon to the diagram. This means that the width of the diagram is reduced ans so are the lifespans of topological features, making it more difficult to distinguish between noise and relevant information. We remark that the classic $k$-NN approach shrinks the diagram even more, to the point that it is very hard to distinguish the information from the noise.

The standard indicator to measure the quality of a denoising is the *Peak Signal over Noise Ratio* (PSNR). Given a grey scale input image $I$ and an output image $O$ with the grey scale between 0 and 255, it is defined by

$$\mathrm{PSNR}(I,O) = 10\log_{10}\left(\frac{256^2}{\frac{1}{ij}\sum_i\sum_j (I[i][j]-O[i][j])^2}\right).$$

Figure 5 shows the quality of the denoising for a set of Lena images with increasing quantity of noise. The curves are obained using the median ($M$) and different values of $k'$ in the discrepancy while $k$ is fixed at 25. The median is better when the noise ratio is small but as we increase the number of outliers, the discrepancy obtains better results. This also shows that the optimal $k'$ depends on the noise ratio. It also depends on the image we consider and thus makes it difficult to find an easy way to choose it automatically. Heuristically, it is better to take $k'$ around $\frac{2}{3}k$, especially when there is a lot of noise.

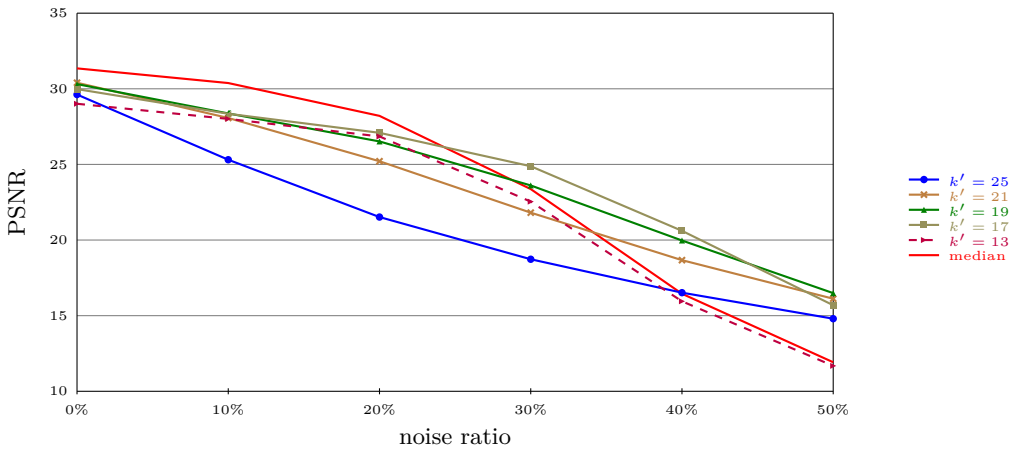



**Figure 5** PSNR for Lena images depending on the choice of $k'$ and the quantity of noise

State of the art results in computer vision obtain better experimental results (e.g. [9, 16, 17]). However, these results assume that the noise model is known and they can start by detecting and removing noisy points before rebuilding the image. Our methods are free from assumptions on the generative model of the image. The algorithms do not change depending on the type of noise.

## Persistence diagram computation

We consider a more topological example from real data. We consider an elevation map of an area near Corte in the French island of Corsica. The true measures of elevation are given

in the left image of Figure 6. The topography can be analysed by looking at the function minus-altitude. We add random faulty sensors that give false results with a 20% probability to simulate malfunctioning equipments. The area covers a square of 2 minutes of arc in both latitude and longitude. We apply our algorithm with the following parameters: $k = 9$, $k' = 7$, $\eta = .05$ minute and $\delta = .025$ minute. We show the recovered persistence diagrams in Figure 7, where the prominent peaks of the original elevation map are highlighted. The "gap" stands for the ratio between the shortest living relevant feature, highlighted in red, and the longest feature created by the noise.

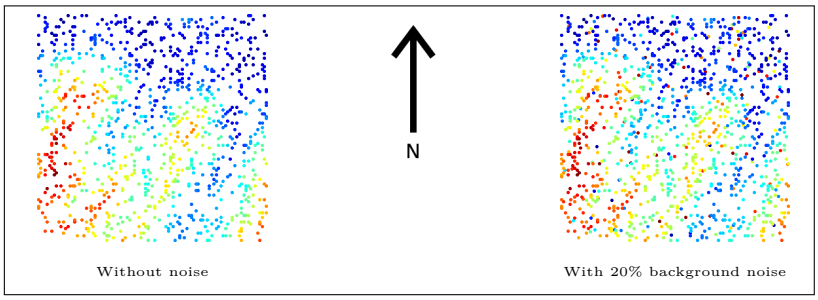



**Figure 6** Elevation map around Corte

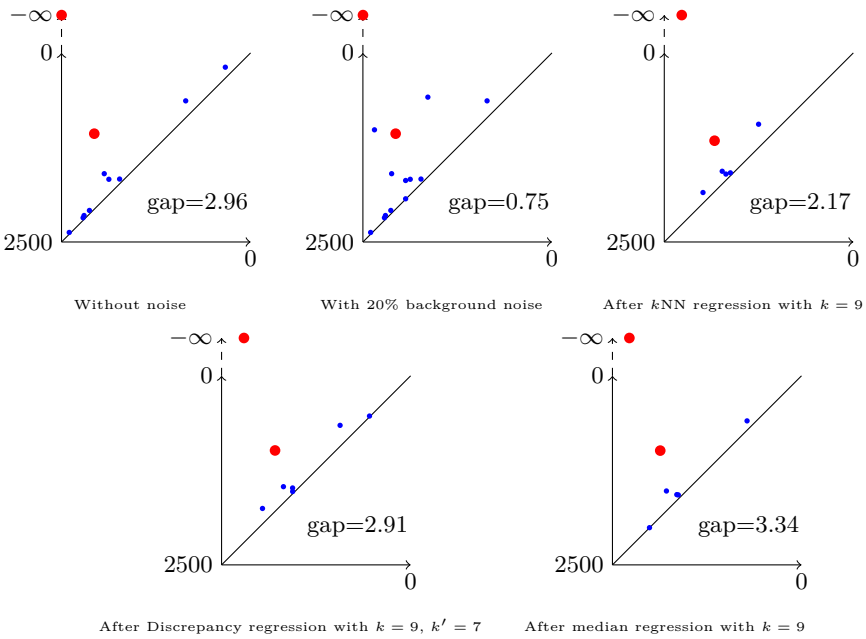



**Figure 7** Persistence diagrams of Corte Elevation map

We note that the gap in the case of the noisy point cloud (before denoising) is less than 1. This means that some relevant topological feature has a shorter lifespan than one caused by noise. Intuitively, this means that it is difficulty to tell true features from noise from this persistence diagram, without performing denoising. We also show the persistence diagrams, as well as the "gap" values, for the denoised data after the three denoising method: $k$-NN regression, $k$-median and our discrepancy based method. In the case of the $k$-NN regression, the topological feature are in the right order. However, the prominence given by the gap is significantly smaller than the one from the original point cloud. Both the discrepancy based

method and the median provides gaps on par with the non-noisy input and thus allow a good recovery of the correct topology.