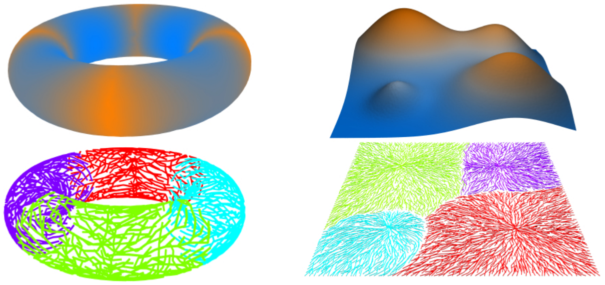
# Analysis of Scalar Fields over Point Cloud Data

Frédéric Chazal* Leonidas J. Guibas† Steve Y. Oudot* Primoz Skraba†



## Abstract

Given a real-valued function $f$ defined over some metric space $\mathbb{X}$, is it possible to recover some structural information about $f$ from the sole information of its values at a finite set $L \subseteq \mathbb{X}$ of sample points, whose pairwise distances in $\mathbb{X}$ are given? We provide a positive answer to this question. More precisely, taking advantage of recent advances on the front of stability for persistence diagrams, we introduce a novel algebraic construction, based on a pair of nested families of simplicial complexes built on top of the point cloud $L$, from which the persistence diagram of $f$ can be faithfully approximated. We derive from this construction a series of algorithms for the analysis of scalar fields from point cloud data. These algorithms are simple and easy to implement, have reasonable complexities, and come with theoretical guarantees. To illustrate the generality of the approach, we present some experimental results obtained in various applications, ranging from clustering to sensor networks (see the electronic version of the paper for color pictures).

## 1 Introduction

Given an unknown domain $\mathbb{X}$ and a scalar field $f : \mathbb{X} \to \mathbb{R}$ whose values are known only at a finite set $L \subseteq \mathbb{X}$ of sample points, our goal is to extract structural information about $f$ from the sole information of the pairwise geodesic distances between the data points and of their function values. No parametrization of $\mathbb{X}$ is assumed — in other words, no coordinates are needed for the points of $L$. This problem finds applications in many scientific fields, including sensor networks, where the data points correspond to sensor nodes and their function values to measurements of some physical quantity (temperature, humidity, etc.), or unsupervised learning, where the data points are sampled from some unknown density distribution, a rough estimate of which is computed at each sample.

The nature of the sought-for information is highly application-dependent. In the aforementioned examples, one is mainly interested in finding the peaks and valleys of the function $f$, together with their respective basins of attraction[1]. In addition, it is desirable to have a mechanism for distinguishing between significant and insignificant peaks or valleys of $f$, which requires to introduce some notion of prominance for the critical points of a function. This is where *topological persistence* comes into play. Inspired from Morse theory, this framework describes the evolution of the topology of the sublevel-sets of $f$, *i.e.* the sets of type $f^{-1}((-\infty, \alpha])$, as parameter $\alpha$ ranges from $-\infty$ to $+\infty$. Topological changes occur only at critical points of $f$, which can be paired in some natural way. The outcome is a set of intervals (called a *persistence barcode* [3]), each of which gives the birth and death times of a topological feature in the sublevel-sets of $f$ — see Figure 1. An equivalent representation is by a multiset of points in the plane, called a *persistence diagram*, where the coordinates of each point correspond to the endpoints of some interval in the barcode. Such representations are used to guide the simplification of scalar fields by iterative cancellations of critical pairs [16, 17]. As such, they provide meaningful information about the prominence of the critical points of a scalar field.

Thus, our goal becomes to approximate the persistence diagram of an unknown scalar field $f$ from its values at a finite set $L$ of sample points, and from the pairwise distances between these points. We provide a theoretically sound solution to this problem; in Section 3 we exhibit a novel algebraic construction, based on a pair of nested families of simplicial complexes, from which the persistence diagram of $f$ is approximated (Theorem 3.1). This construction is provably robust to noise both in pairwise distances and function values (Theorem 3.2). From these structural results, we derive algorithms (Section 4) for approximating the persistence diagram of $f$ from its values at the points of $L$, both in static (fixed $f$) and in dynamic (time-varying $f$)

*INRIA, Geometrica group, 4 rue Jacques Monod, 91893 ORSAY, France. {frederic.chazal, steve.oudot}@inria.fr

†Department of Computer Science, Stanford University, Stanford, CA 94305. guibas@cs.stanford.edu, primoz@stanford.edu

[1] In the context of clustering, this approach to the problem is reminiscent of Mean Shift [13].

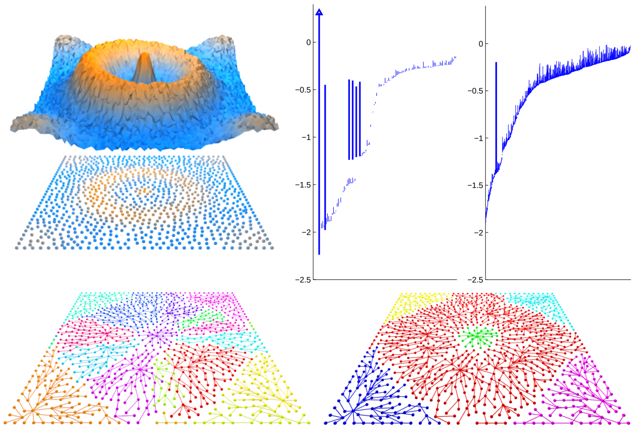


Figure 1: *Top row, left: a noisy scalar field $f$ defined over a sampled planar square domain $\mathbb{X}$; center and right: approximations of the 0- and 1-dimensional persistence barcodes of $(-f)$ generated by our method from the values of $f$ at the sample points and from their approximate pairwise geodesic distances in $\mathbb{X}$. The six long intervals in the 0-dimensional barcode correspond to the six prominent peaks of $f$ (including the top of the crater), while the long interval in the 1-dimensional barcode reveals the ring shape of the basin of attraction of the top of the crater. Bottom row: approximate basins of attraction of the peaks of $f$, before (left) and after (right) merging non-persistent clusters, thus revealing the intuitive structure of $f$.*

settings. We also show how to find the basins of attraction of the peaks of $f$ inside the point cloud $L$, and how to merge them according to the persistence information, as shown in Figure 1 (right). Our algorithms are based on variants [11, 12] of the celebrated persistence algorithm [16, 26]. They can be easily implemented, have reasonable complexities, and are provably correct. Finally, we show experimental results in a variety of applications (Section 5): while we do not provide definitive solutions to these problems, the results demonstrate the potential of our method and its possible interest for the community.

**Related work.** Topological persistence has already been used in the past for the analysis and simplification of scalar fields. The original persistence paper [16] showed how to simplify the graph of a piecewise-linear (PL) real-valued function $f$ defined over a simplicial complex $\mathbb{X}$ in $\mathbb{R}^3$, by iteratively cancelling the pairs of critical points provided by the persistence barcode of $f$. This approach was later refined, in the special case where $\mathbb{X}$ is a triangulated 2-manifold, to only cancel the pairs corresponding to short intervals in the barcode, thus removing the topological noise up to a certain prescribed amplitude [17]. In parallel, others have considered computing complete or simplified representations of Morse-Smale complexes, which capture important information about the structure of scalar fields. Building upon the idea of iterative cancellations of pairs of critical points, it is possible to construct hierarchies of increasingly coarse Morse-Smale complexes from PL functions defined over triangulated 2- or 3-manifolds [1, 5, 15, 20, 21]. All these methods are restricted to the low-dimensional PL setting, and in this respect our work suggests a way of extending the approach to a more general class of spaces via finite sampling and modulo some (controlled) errors in the output. Although finding and merging the basins of attraction of the peaks of a scalar field $f$ is simpler than computing a full hierarchy of Morse-Smale complexes, it is already a challenge in our context, where the knowledge of $f$ is very weak, and where the potentially high dimensionality of the data makes PL approximations prohibitively costly.

Another line of work in which persistence has played a prominent role is homology inference from point cloud data, where the goal is to recover the homological type of some unknown compact set $\mathbb{X} \subset \mathbb{R}^d$ from a finite set $L$ of sample points. Under a sufficient sampling density, the distance to $L$ in $\mathbb{R}^d$ approximates the distance to $\mathbb{X}$, therefore their persistence diagrams are close, by a stability result of [10]. This makes the inference of the homology of $\mathbb{X}$ from the persistence of the distance to $L$ theoretically possible [8, 10]. In practice, computing this distance at every point of the ambient space $\mathbb{R}^d$ is prohibitively expensive. It is then necessary to resort to auxiliary algebraic constructions, such as the *Rips complex* $R_\delta(L)$, defined as the abstract simplicial complex whose simplices correspond to non-empty subsets of $L$ of diameter less than $\delta$. As proved in [9], a pair of nested Rips complexes $R_\delta(L) \subseteq R_{\delta'}(L)$ can provably capture the homology of the underlying space $\mathbb{X}$, though the the individual complexes do not. Our algebraic construction (see Section 3) is directly inspired from this property, and in fact our theoretical analysis is articulated in the same way as in [9], namely: we first work out structural properties of unions of geodesic balls, which we prove to also hold for their nerves (also called *Čech complexes*); then, using the strong relationship that exists between Čech and Rips complexes, we derive structural properties for families of Rips complexes. Note that the core of our analysis differs significantly from [9], because our families of complexes are built differently. In particular, the classical notion of stability of persistence diagrams, as introduced in [10], is not broad enough to encompass our setting, where it is replaced by a generalized version recently proposed by Chazal *et al.* [6].

## 2 Background

Throughout the paper we use singular homology with coefficients in a commutative ring $R$, assumed to be a field and omitted in the notations. We also use elements of Riemannian geometry and of Morse theory (in Section 4.2). Thorough introductions to these topics may be found in [4, 22, 23].

### 2.1 Persistence modules and filtrations.

The main algebraic objects under study here are persistence modules. A persistence module is a family $\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}$ of $R$-modules together with a family $\{\phi_\alpha^\beta : \Phi_\alpha \to \Phi_\beta\}_{\alpha\leq\beta\in\mathbb{R}}$ of homomorphisms such that $\forall\alpha \leq \beta \leq \gamma$, $\phi_\alpha^\gamma = \phi_\beta^\gamma \circ \phi_\alpha^\beta$ and $\phi_\alpha^\alpha = \mathrm{id}_{\Phi_\alpha}$. For simplicity of notation, we remove the ranges of indices when they are obvious. Persistence modules are often derived from *filtrations*, *i.e.* families $\{F_\alpha\}$ of topological spaces that are nested with respect to inclusion. For all $\alpha \leq \beta$, the canonical inclusion map $F_\alpha \hookrightarrow F_\beta$ induces homomorphisms between the homology groups $H_k(F_\alpha) \to H_k(F_\beta)$ of all dimensions $k \in \mathbb{N}$. Thus, for any fixed $k$ the family $\{H_k(F_\alpha)\}$, together with the homomorphisms induced by inclusions, form a persistence module, called the *$k$th persistent homology module of* $\{F_\alpha\}$. An important class of filtrations are the *sublevel-sets filtrations*. Given a topological space $\mathbb{X}$ and a function $f : \mathbb{X} \to \mathbb{R}$, the sublevel-sets filtration of $f$ is the family $\{F_\alpha\}$ of subspaces of $\mathbb{X}$ of type $F_\alpha = f^{-1}((-\infty, \alpha])$. This family forms a filtration because $f^{-1}((-\infty, \alpha]) \subseteq f^{-1}((-\infty, \beta])$ whenever $\alpha \leq \beta$. By default, the *$k$th persistent homology module of $f$* refers to the one of its sublevel-sets filtration $\{F_\alpha\}$.

**2.2 Persistence diagrams, tameness, stability.** Following [6], we say that a persistence module $(\{\Phi_\alpha\}, \{\phi_\alpha^\beta\})$ is 0-*tame* (or simply *tame*) if the rank of $\phi_\alpha^\beta$ is finite for all $\alpha < \beta$. Observe that persistent homology modules of nested families of finite simplicial complexes are always tame. Therefore, all the persistence modules introduced in the following sections of the paper will be tame.

The persistence diagram of a tame persistence module $(\{\Phi_\alpha\}, \{\phi_\alpha^\beta\})$ is defined as a multiset of points in the extended plane $\bar{\mathbb{R}}^2$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. This multiset is obtained as the limit of the following iterative process: given arbitrary values $a, \varepsilon > 0$, we discretize the persistence module over the integer scale $a + \varepsilon\mathbb{Z}$, considering the subfamily $\{\Phi_{a+k\varepsilon}\}_{k\in\mathbb{Z}}$ of vector spaces together with the subfamily $\{\phi_{a+k\varepsilon}^{a+l\varepsilon}\}_{k\leq l\in\mathbb{Z}}$ of linear maps. Its persistence diagram is defined naturally from [10] as the set of vertices of the regular grid $(a + \varepsilon\mathbb{Z}) \times (a + \varepsilon\mathbb{Z})$ in $\bar{\mathbb{R}}^2$, plus the diagonal $\Delta = \{(x, x),\ x \in \bar{\mathbb{R}}\}$, where each grid vertex $(a + k\varepsilon, a + l\varepsilon)$ is given the (finite) multiplicity[2] $\mathrm{mult}(a + k\varepsilon, a + l\varepsilon) = \mathrm{rank}\ \phi_{a+k\varepsilon}^{a+(l-1)\varepsilon} - \mathrm{rank}\ \phi_{a+k\varepsilon}^{a+l\varepsilon} + \mathrm{rank}\ \phi_{a+(k-1)\varepsilon}^{a+l\varepsilon} - \mathrm{rank}\ \phi_{a+(k-1)\varepsilon}^{a+(l-1)\varepsilon}$, while each point of $\Delta$ is given infinite multiplicity. Then, the persistence diagram of $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}}, \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ is the limit multiset obtained as $\varepsilon \to 0$. It is known to be independent of the choice of $a$ [6].

An important property of persistence diagrams is that they are stable under small perturbations of the persistence modules. The first result in this vein was devised for the special case of persistent homology modules of continuous functions [10]. Recently, Chazal *et al.* [6] dropped the functional setting and proposed instead the following generalized notion of proximity between persistence modules:

DEFINITION 2.1. *Two persistence modules $(\{\Phi_\alpha\}_{\alpha\in\mathbb{R}},\ \{\phi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ and $(\{\Psi_\alpha\}_{\alpha\in\mathbb{R}},\ \{\psi_\alpha^\beta\}_{\alpha\leq\beta\in\mathbb{R}})$ are (strongly) $\varepsilon$-*interleaved *if there exist two families of homomorphisms, $\{\mu_\alpha : \Phi_\alpha \to \Psi_{\alpha+\varepsilon}\}_{\alpha\in\mathbb{R}}$ and $\{\nu_\alpha : \Psi_\alpha \to \Phi_{\alpha+\varepsilon}\}_{\alpha\in\mathbb{R}}$, such that $\phi_\alpha^\beta = \nu_{\beta-\varepsilon} \circ \psi_{\alpha+\varepsilon}^{\beta-\varepsilon} \circ \mu_\alpha$ and $\psi_\alpha^\beta = \mu_{\beta-\varepsilon} \circ \phi_{\alpha+\varepsilon}^{\beta-\varepsilon} \circ \nu_\alpha$ for all $\beta \geq \alpha + 2\varepsilon$.*

Under this condition, they proved the following generalized stability result [6]:

THEOREM 2.1. *If two tame persistence modules are $\varepsilon$-interleaved, then, in the extended plane $\bar{\mathbb{R}}^2$ endowed with the $l^\infty$ norm, the bottleneck distance between their persistence diagrams is at most $\varepsilon$.*

Recall that the bottleneck distance $\mathrm{d}_B^\infty(A, B)$ between two multisets in $(\bar{\mathbb{R}}^2, l^\infty)$ is the quantity $\min_\gamma \max_{p\in A} \|p - \gamma(p)\|_\infty$, where $\gamma$ ranges over all bijections from $A$ to $B$. An important special case is when the persistence modules are the $k$th persistent homology modules of two filtrations $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ and $\{G_\alpha\}_{\alpha\in\mathbb{R}}$ such that $F_\alpha \subseteq G_{\alpha+\varepsilon}$ and $G_\alpha \subseteq F_{\alpha+\varepsilon}$ for all $\alpha \in \mathbb{R}$ (we then say that $\{F_\alpha\}$ and $\{G_\alpha\}$ are $\varepsilon$-interleaved, by analogy). In this case, the maps $\mu_\alpha$ and $\nu_\alpha$ induced at homology level by the inclusions $F_\alpha \hookrightarrow G_{\alpha+\varepsilon}$ and $G_\alpha \hookrightarrow F_{\alpha+\varepsilon}$ make the two persistence modules $\varepsilon$-interleaved, and Theorem 2.1 guarantees that their persistence diagrams are $\varepsilon$-close, if they are tame. To simplify the exposition, we call a filtration or function *tame* if all its persistent homology modules are tame. Then, its *$k$th persistence diagram* is the persistence diagram of its $k$th persistent homology module.

**2.3 Geodesic $\varepsilon$-samples on Riemannian manifolds.** From now on, and unless otherwise stated, $\mathbb{X}$ denotes a compact Riemannian manifold, possibly with boundary, and $\mathrm{d}_\mathbb{X}$ denotes its geodesic distance. We also let $L$ be a finite set of points of $\mathbb{X}$ that form a *geodesic $\varepsilon$-sample* of $\mathbb{X}$, that is: $\forall x \in \mathbb{X}$, $\mathrm{d}_\mathbb{X}(x, L) < \varepsilon$. Our theoretical claims will assume $\varepsilon$ to be at most a fraction of the *strong convexity radius* of $\mathbb{X}$, denoted $\varrho_c(\mathbb{X})$. It is defined as the largest value such that every open geodesic ball $B_\mathbb{X}(x, r) = \{y \in \mathbb{X},\ \mathrm{d}_\mathbb{X}(x, y) < r\}$ of center $x \in \mathbb{X}$ and radius $r < \varrho_c(\mathbb{X})$ is *strongly convex*, namely: for every pair of points $y, y'$ in the closure of $B_\mathbb{X}(x, r)$, there exists a unique shortest path in $\mathbb{X}$ between $y$ and $y'$, and the interior of this path is included in $B_\mathbb{X}(x, r)$. The strong convexity radius plays an important role in the paper because strongly convex sets are contractible[3] and intersections of strongly convex sets are also strongly convex.

[2] Roughly speaking, the multiplicity of point $(a+k\varepsilon, a+l\varepsilon)$ corresponds to the number of homological features that are born between times $a+(k-1)\varepsilon$ and $a+k\varepsilon$ and that die between times $a+(l-1)\varepsilon$ and $a+l\varepsilon$.

[3] A topological space is contractible if it can be continuously deformed to a point within itself.

**2.4 Offsets, nerves, Rips complexes.** For all $\delta > 0$, let $L^\delta$ denote the $\delta$-*offset* of $L$, defined as the union of the open geodesic balls of same radius $\delta$ about the points of $L$, namely: $L^\delta = \bigcup_{p \in L} B_{\mathbb{X}}(p, \delta)$. Its nerve, also called *Čech complex* and noted $C_\delta(L)$, is the abstract simplicial complex of vertex set $L$ whose simplices correspond to non-empty subsets of the family of open balls $\{B_{\mathbb{X}}(p, \delta)\}_{p \in L}$ whose elements have a non-empty common intersection. The topology of $C_\delta(L)$ is related to the topology of its dual union of balls through the Nerve Theorem [22, §4G]. This makes the Čech complex a good candidate data structure in theory. However, it can be difficult to compute in practice, where it is often replaced by the *(Vietoris-)Rips complex* $R_\delta(L)$, whose simplices correspond to non-empty subsets of $L$ of geodesic diameter less than $\delta$. This condition only involves distance comparisons, which are much easier to check than the emptiness of an intersection of geodesic balls. In addition, the Rips complex is related to the Čech complex through the following sequence of inclusions [9]:

$$\forall \delta > 0,\ C_{\frac{\delta}{2}}(L) \subseteq R_\delta(L) \subseteq C_\delta(L) \tag{2.1}$$

In the literature, Čech and Rips complexes are usually turned into filtrations by letting parameter $\delta$ vary from $0$ to $+\infty$. In contrast, the following sections present algebraic constructions where $\delta$ remains fixed to some constant value while the vertex set grows from $\emptyset$ to $L$.

### 3 Structural properties

Let $\mathbb{X}, L$ be defined as in Section 2, and let $f : \mathbb{X} \to \mathbb{R}$ be tame and $c$-Lipschitz. Assuming $\mathbb{X}$ and $f$ to be unknown, our goal is to approximate the $k$th persistence diagram of $f$ from its values at the points of $L$. The main result of the section (Theorem 3.1 below) claims that this is possible using an algebraic construction based on Rips complexes. From now on, $L_\alpha$ denotes the set $L \cap f^{-1}((-\infty, \alpha])$.

Our construction is inspired from [9], where it is shown that a pair of nested Rips complexes can provably-well capture the homology of a domain even though none of the individual Rips complexes does. Given a fixed parameter $\delta > 0$, we use two Rips-based filtrations simultaneously, $\{R_\delta(L_\alpha)\}_{\alpha \in \mathbb{R}}$ and $\{R_{2\delta}(L_\alpha)\}_{\alpha \in \mathbb{R}}$, and we consider the persistence modules formed at homology level by the images of the homomorphisms induced by the inclusions $R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)$. Specifically, for all $k \in \mathbb{N}$ and all $\alpha \leq \beta$ we have the following induced commutative diagram at homology level:

$$\begin{tikzcd}
H_k(R_\delta(L_\beta)) \arrow[r] & H_k(R_{2\delta}(L_\beta)) \\
H_k(R_\delta(L_\alpha)) \arrow[r] \arrow[u] & H_k(R_{2\delta}(L_\alpha)) \arrow[u]
\end{tikzcd}$$

Letting $\Gamma_\alpha^k$ be the image of $H_k(R_\delta(L_\alpha)) \to H_k(R_{2\delta}(L_\alpha))$, we get that the above commutative diagram induces a map $\gamma_\alpha^\beta : \Gamma_\alpha^k \to \Gamma_\beta^k$. Since this is true for all $\alpha \leq \beta$, the family $\{\Gamma_\alpha^k\}_{\alpha \in \mathbb{R}}$ of vector spaces, together with the family $\{\gamma_\alpha^\beta\}_{\alpha \leq \beta}$ of linear maps, forms a persistence module. By analogy with the terminology of Section 2, we call it the $k$th persistent homology module of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha \in \mathbb{R}}$, and its persistence diagram the $k$th persistence diagram of the nested pair. In fact, this construction is not specific to families of Rips complexes, and it allows a persistence module to be defined $(\{\Gamma_\alpha^k\}, \{\gamma_\alpha^\beta\})$ from the $k$-dimensional homology groups of any pair of filtrations $\{G_\alpha\}$ and $\{G'_\alpha\}$ that are nested with respect to inclusion: $\forall \alpha \in \mathbb{R}$, $G_\alpha \subseteq G'_\alpha$.

THEOREM 3.1. *Let $\mathbb{X}$ be a compact Riemannian manifold possibly with boundary, $L$ a geodesic $\varepsilon$-sample of $\mathbb{X}$, and $f : \mathbb{X} \to \mathbb{R}$ a tame $c$-Lipschitz function. If $\varepsilon < \frac{1}{4}\varrho_c(\mathbb{X})$, then for any $\delta \in [2\varepsilon,\ \frac{1}{2}\varrho_c(\mathbb{X}))$ and any $k \in \mathbb{N}$, the $k$th persistent homology modules of $f$ and of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha \in \mathbb{R}}$ are $2c\delta$-interleaved. Therefore, the bottleneck distance between their persistence diagrams is at most $2c\delta$, by Theorem 2.1.*

The proof, based on a technique of algebraic topology called *diagram chasing*, is given in Section 3.1. Its robustness with respect to noise in the distances or function values is addressed in Section 3.2.

**3.1 Proof of Theorem 3.1.** We begin with some preliminary results about unions of geodesic balls and their nerves, whose proofs can be found in the full version of the paper [7]. Let $\delta > 0$ be a fixed parameter.

LEMMA 3.1. *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. Then, for any $\delta \geq \varepsilon$, the sublevel-sets filtration $\{F_\alpha\}$ of $f$ is $c\delta$-interleaved with the $\delta$-offsets filtration $\{L_\alpha^\delta\}_{\alpha \in \mathbb{R}}$. Hence, $\forall k \in \mathbb{N}$, the bottleneck distance between their $k$th persistence diagrams is at most $c\delta$, by Theorem 2.1.*

Using a variant of the Nerve Theorem introduced in [9], we can extend the above result to the nerves $C_\delta(L_\alpha)$ of the $\delta$-offsets $L_\alpha^\delta$:

LEMMA 3.2. *Let $\mathbb{X}, f, L$ be as in Theorem 3.1. If $\varepsilon < \varrho_c(\mathbb{X})$, then, $\forall k \in \mathbb{N}$, there exists a family of isomorphisms $\{H_k(C_\delta(L_\alpha)) \to H_k(L_\alpha^\delta)\}_{\alpha \in \mathbb{R},\ \varepsilon \leq \delta < \varrho_c(\mathbb{X})}$ such that the following diagrams (where horizontal homomorphisms are induced by inclusions) commute: $\forall \alpha \leq \alpha' \in \mathbb{R}, \forall \delta \leq \delta' \in [\varepsilon,\ \varrho_c(\mathbb{X}))$,*

$$\begin{tikzcd}
H_k(C_\delta(L_\alpha)) \arrow[r] \arrow[d] & H_k(C_{\delta'}(L_{\alpha'})) \arrow[d] \\
H_k(L_\alpha^\delta) \arrow[r] & H_k(L_{\alpha'}^{\delta'})
\end{tikzcd}$$

*Hence, $\forall k \in \mathbb{N}$, $\forall \delta \in [\varepsilon, \varrho_c(\mathbb{X}))$, the $k$th persistent homology modules of $\{C_\delta(L_\alpha)\}_{\alpha \in \mathbb{R}}$ and $\{L_\alpha^\delta\}_{\alpha \in \mathbb{R}}$ are $0$-interleaved, which implies that their persistence diagrams are identical, by Theorem 2.1.*

With these preliminary results at hand, we can now proceed to the proof of Theorem 3.1:

**Proof of Theorem 3.1.** For convenience, we let $\varepsilon' = \delta$, $\varepsilon'' = 2\delta$, $G_\alpha = R_\delta(L_\alpha)$, and $G'_\alpha = R_{2\delta}(L_\alpha)$. Given some fixed $k \in \mathbb{N}$, let $(\{\Gamma^k_\alpha\}, \{\gamma^\beta_\alpha\})$ be the $k$th persistent homology module of the nested pair of filtrations $\{G_\alpha \hookrightarrow G'_\alpha\}_{\alpha \in \mathbb{R}}$. Since $2\varepsilon \leq \varepsilon' \leq \frac{\varepsilon''}{2}$, Eq. (2.1) induces the following commutative diagram at $k$th homology level: $\forall \alpha \leq \beta$,

(3.2)
$$\begin{tikzcd}
H_k(C_\varepsilon(L_\alpha)) \arrow[r, "i^\beta_\alpha"] \arrow[d, "a_\alpha"] & H_k(C_\varepsilon(L_\beta)) \arrow[d, "a_\beta"] \\
H_k(G_\alpha) \arrow[r, "j^\beta_\alpha"] \arrow[d, "b_\alpha"] & H_k(G_\beta) \arrow[d, "b_\beta"] \\
H_k(C_{\varepsilon'}(L_\alpha)) \arrow[r, "l^\beta_\alpha"] \arrow[d, "d_\alpha"] & H_k(C_{\varepsilon'}(L_\beta)) \arrow[d, "d_\beta"] \\
H_k(G'_\alpha) \arrow[r, "m^\beta_\alpha"] \arrow[d, "e_\alpha"] & H_k(G'_\beta) \arrow[d, "e_\beta"] \\
H_k(C_{\varepsilon''}(L_\alpha)) \arrow[r, "n^\beta_\alpha"] & H_k(C_{\varepsilon''}(L_\beta))
\end{tikzcd}$$

This diagram emphasizes the relationship between the persistence module $(\{\Gamma^k_\alpha\}, \{\gamma^\beta_\alpha\})$ and the homology of a family of Čech complexes. In addition, Lemma 3.1 tells us that the homology of the sublevel-sets filtration $\{F_\alpha\}$ of $f$ is related to the homology of the $\varepsilon$-, $\varepsilon'$- and $\varepsilon''$-offsets of $L_\alpha$ through the following sequence of homomorphisms induced by inclusions: $\forall \alpha, \beta$ s.t. $\beta - \alpha \geq c(\varepsilon + \varepsilon'')$,

(3.3)
$$\begin{tikzcd}
H_k(F_{\alpha - c\varepsilon}) \arrow[d, "t_\alpha"] & H_k(F_{\beta + c\varepsilon''}) \\
H_k(L^\varepsilon_\alpha) \arrow[d, "u_\alpha"] & H_k(L^{\varepsilon''}_\beta) \arrow[u, "w_\beta"] \\
H_k(L^{\varepsilon'}_\alpha) \arrow[d, "v_\alpha"] & H_k(L^{\varepsilon'}_\beta) \arrow[u, "v_\beta"] \\
H_k(L^{\varepsilon''}_\alpha) \arrow[d, "w_\alpha"] & H_k(L^{\varepsilon}_\beta) \arrow[u, "u_\beta"] \\
H_k(F_{\alpha + c\varepsilon''}) \arrow[r, "s^\beta_\alpha"] & H_k(F_{\beta - c\varepsilon}) \arrow[u, "t_\beta"]
\end{tikzcd}$$

Now, for all $\alpha \in \mathbb{R}$ we let $h_\alpha : H_k(C_\varepsilon(L_\alpha)) \to H_k(L^\varepsilon_\alpha)$, $h'_\alpha : H_k(C_{\varepsilon'}(L_\alpha)) \to H_k(L^{\varepsilon'}_\alpha)$ and $h''_\alpha : H_k(C_{\varepsilon''}(L_\alpha)) \to H_k(L^{\varepsilon''}_\alpha)$ be the isomorphisms provided by Lemma 3.2 — which are well-defined since $\varepsilon \leq \varepsilon' \leq \varepsilon'' < \varrho_c(\mathbb{X})$. Combining these isomorphisms with Eqs. (3.2) and (3.3), we get a full diagram relating $(\{\Gamma^k_\alpha\}, \{\gamma^\beta_\alpha\})$ to the $k$th persistent homology module of $\{F_\alpha\}$. This diagram may not commute: for instance, there is no particular reason why $m^\beta_\alpha$ should coincide with $d_\beta \circ b_\beta \circ a_\beta \circ h^{-1}_\beta \circ t_\beta \circ s^\beta_\alpha \circ w_\alpha \circ h''_\alpha \circ e_\alpha$. Nevertheless, the subdiagram of Eq. (3.2) commutes because it is induced by inclusions. Furthermore, Lemma 3.2 ensures that the following subdiagrams (where the new homomorphisms ${l'}^\beta_\alpha$ and ${n'}^\beta_\alpha$ are induced by inclusions) also commute: $\forall \alpha \leq \beta$, $\forall \gamma \in \{\alpha, \beta\}$,

(3.4)
$$\begin{tikzcd}
H_k(C_\varepsilon(L_\gamma)) \arrow[r, "h_\gamma"] \arrow[d, "b_\gamma \circ a_\gamma"] & H_k(L^\varepsilon_\gamma) \arrow[d, "u_\gamma"] \\
H_k(C_{\varepsilon'}(L_\gamma)) \arrow[r, "h'_\gamma"] \arrow[d, "e_\gamma \circ d_\gamma"] & H_k(L^{\varepsilon'}_\gamma) \arrow[d, "v_\gamma"] \\
H_k(C_{\varepsilon''}(L_\gamma)) \arrow[r, "h''_\gamma"] & H_k(L^{\varepsilon''}_\gamma)
\end{tikzcd}$$

(3.5)
$$\begin{tikzcd}
H_k(C_{\varepsilon'}(L_\alpha)) \arrow[r, "l^\beta_\alpha"] \arrow[d, "h'_\alpha"] & H_k(C_{\varepsilon'}(L_\beta)) \arrow[d, "h'_\beta"] \\
H_k(L^{\varepsilon'}_\alpha) \arrow[r, "{l'}^\beta_\alpha"] & H_k(L^{\varepsilon'}_\beta).
\end{tikzcd}$$

(3.6)
$$\begin{tikzcd}
H_k(C_{\varepsilon''}(L_\alpha)) \arrow[r, "n^\beta_\alpha"] \arrow[d, "h''_\alpha"] & H_k(C_{\varepsilon''}(L_\beta)) \arrow[d, "h''_\beta"] \\
H_k(L^{\varepsilon''}_\alpha) \arrow[r, "{n'}^\beta_\alpha"] & H_k(L^{\varepsilon''}_\beta)
\end{tikzcd}$$

For all $\alpha \in \mathbb{R}$, let $\phi_\alpha : \Gamma^k_\alpha \to H_k(F_{\alpha + c\varepsilon''})$ be the restriction of the map $w_\alpha \circ h''_\alpha \circ e_\alpha$ to the subspace $\Gamma^k_\alpha = \mathrm{im}\ d_\alpha \circ b_\alpha \subseteq H_k(G'_\alpha)$. Symmetrically, let $\psi_{\alpha - c\varepsilon} : H_k(F_{\alpha - c\varepsilon}) \to \Gamma^k_\alpha$ be the map $d_\alpha \circ b_\alpha \circ a_\alpha \circ h^{-1}_\alpha \circ t_\alpha$. Its image is indeed included in the subspace $\Gamma^k_\alpha = \mathrm{im}\ d_\alpha \circ b_\alpha \subseteq H_k(G'_\alpha)$. To prove that the persistence module $(\{\Gamma^k_\alpha\}, \{\gamma^\beta_\alpha\})$ is $c\varepsilon''$-interleaved with the $k$th persistent homology module of $\{F_\alpha\}$, it suffices to show that (a.) the map $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha$ is equal to $m^\beta_\alpha$ over the subspace $\Gamma^k_\alpha \subseteq H_k(G_\alpha)$ for all $\beta \geq \alpha + c(\varepsilon + \varepsilon'')$, and (b.) the map $\phi_\beta \circ m^\beta_\alpha \circ \psi_{\alpha - c\varepsilon}$ is equal to the homomorphism $s^{\beta + c\varepsilon''}_{\alpha - c\varepsilon} : H_k(F_{\alpha - c\varepsilon}) \to H_k(F_{\beta + c\varepsilon''})$ induced by inclusion for all $\beta \geq \alpha$.

(a.) Consider the map $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha$. Since by definition we have $\Gamma^k_\alpha = \mathrm{im}\ d_\alpha \circ b_\alpha \subseteq \mathrm{im}\ d_\alpha$, the fact that $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha$ coincides with $m^\beta_\alpha$ over $\Gamma^k_\alpha$ is a direct consequence of the fact that the map $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha \circ d_\alpha$ equals $m^\beta_\alpha \circ d_\alpha$ over $H_k(C_{\varepsilon'}(L_\alpha))$, which we will now prove. Replacing $\phi_\alpha$ and $\psi_{\beta - c\varepsilon}$ by their definitions, we get $d_\beta \circ (b_\beta \circ a_\beta \circ h^{-1}_\beta) \circ t_\beta \circ s^\beta_\alpha \circ w_\alpha \circ ({h''}_\alpha \circ e_\alpha \circ d_\alpha)$, which by commutativity of (3.4) is equal to $d_\beta \circ {h'}^{-1}_\beta \circ u_\beta \circ t_\beta \circ s^\beta_\alpha \circ w_\alpha \circ v_\alpha \circ {h'}_\alpha$. Now, observe that $u_\beta \circ t_\beta \circ s^\beta_\alpha \circ w_\alpha \circ v_\alpha$ is nothing but the homomorphism ${l'}^\beta_\alpha$ induced by the inclusion $L^{\varepsilon'}_\alpha \hookrightarrow L^{\varepsilon'}_\beta$. Therefore, we have $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha \circ d_\alpha = d_\beta \circ ({h'}^{-1}_\beta \circ {l'}^\beta_\alpha \circ {h'}_\alpha)$, which is equal to $d_\beta \circ l^\beta_\alpha$ by commutativity of (3.5). Finally, we have $d_\beta \circ l^\beta_\alpha = m^\beta_\alpha \circ d_\alpha$ by commutativity of (3.2). Thus, $\psi_{\beta - c\varepsilon} \circ s^\beta_\alpha \circ \phi_\alpha$ coincides with $m^\beta_\alpha$ over $\Gamma^k_\alpha$.

(b.) Consider now the map $\phi_\beta \circ m^\beta_\alpha \circ \psi_{\alpha - c\varepsilon}$. Replacing $\phi_\beta$ and $\psi_{\alpha - c\varepsilon}$ by their definitions, we get $w_\beta \circ h''_\beta \circ (e_\beta \circ m^\beta_\alpha) \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h^{-1}_\alpha \circ t_\alpha$, which by commutativity of (3.2) is equal to $w_\beta \circ h''_\beta \circ (n^\beta_\alpha \circ e_\alpha) \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h^{-1}_\alpha \circ t_\alpha$. Now, the commutativity of (3.4) implies that $e_\alpha \circ d_\alpha \circ b_\alpha \circ a_\alpha \circ h^{-1}_\alpha = {h''}^{-1}_\alpha \circ v_\alpha \circ u_\alpha$, therefore $\phi_\beta \circ m^\beta_\alpha \circ \psi_{\alpha - c\varepsilon}$ is equal to

$w_\beta \circ (h''_\beta \circ n_\alpha^\beta \circ {h''}_\alpha^{-1}) \circ v_\alpha \circ u_\alpha \circ t_\alpha$, which by commutativity of (3.6) is equal to $w_\beta \circ {n'}_\alpha^\beta \circ v_\alpha \circ u_\alpha \circ t_\alpha = s_{\alpha - c\varepsilon}^{\beta + c\varepsilon''}$. □

**3.2 Robustness with respect to noise in the data.** Theorem 3.1 assumes that exact geodesic distances and function values are used in the construction of the Rips complexes. In practice however, function values from physical measurements are inherently noisy, while geodesic distances are estimated through some neighborhood graph distance. We claim that our framework is generic enough to handle these practical situations:

THEOREM 3.2. *Let $\mathbb{X}, f, L$ be as in Theorem 3.1, and let $k \in \mathbb{N}$.*

(i) *Suppose $f$ is known within a precision of $\zeta$ at the points of $L$. Then, for any $\delta \in [2\varepsilon, \frac{1}{2}\varrho_c(\mathbb{X}))$, the $k$th persistent homology modules of $f$ and of the nested pair of filtrations $\{R_\delta(L_\alpha) \hookrightarrow R_{2\delta}(L_\alpha)\}_{\alpha \in \mathbb{R}}$ are $(2c\delta + \zeta)$-interleaved, hence their persistence diagrams are $(2c\delta + \zeta)$-close.*

(ii) *Suppose the Rips complexes are now defined with respect to some distance $\tilde{\mathrm{d}}_\mathbb{X}$ that is related to $\mathrm{d}_\mathbb{X}$ through some scaling factor $\lambda \geq 1$, relative error $\mu \geq 1$, and additive error $\nu \geq 0$: $\forall p, q \in L$, $\frac{\mathrm{d}_\mathbb{X}(p,q)}{\lambda} \leq \tilde{\mathrm{d}}_\mathbb{X}(p,q) \leq \nu + \mu \frac{\mathrm{d}_\mathbb{X}(p,q)}{\lambda}$. Then, for any $\delta \geq \nu + 2\mu\frac{\varepsilon}{\lambda}$ and any $\delta' \in [\nu + 2\mu\delta, \ \frac{1}{\lambda}\varrho_c(\mathbb{X}))$, the $k$th persistent homology modules of $f$ and of the nested pair of filtrations $\{\tilde{R}_\delta(L_\alpha) \hookrightarrow \tilde{R}_{\delta'}(L_\alpha)\}_{\alpha \in \mathbb{R}}$ are $c\lambda\delta'$-interleaved, hence their persistence diagrams are $c\lambda\delta'$-close.*

While the meaning of assertion (i) is self-explanatory, assertion (ii) deserves a few words of context. We use the notations $\tilde{R}_\delta(L_\alpha)$ and $\tilde{R}_{\delta'}(L_\alpha)$ to emphasize that the Rips complexes are now defined with respect to $\tilde{\mathrm{d}}_\mathbb{X}$. The latter is usually taken to be the distance $\mathrm{d}_G$ in some neighborhood graph $G$ built on top of the point cloud $L$, whose edges can be either weighted or unweighted, depending on the application. Both classes of graphs are known to provide distances $\mathrm{d}_G$ satisfying the hypotheses of (ii) for some quantities $\lambda, \mu, \nu$ [18, 24]. In addition, parameters $\delta, \delta'$ are required to be somewhat relaxed: specifically, $\delta$ must be sligthly larger than $2\frac{\varepsilon}{\lambda}$ and $\delta'$ slightly larger than $2\delta$.

**Proof of Theorem 3.2.** Assume first that for each point $p \in L$ we are given a value $\tilde{f}(p) \neq f(p)$, and let $\zeta = \max_{p \in L} |\tilde{f}(p) - f(p)|$. For convenience, for all $\alpha \in \mathbb{R}$ we introduce the set $\tilde{L}_\alpha$ of points of $L$ whose $\tilde{f}$-values are at most $\alpha$. Note that $\tilde{L}_\alpha$ may neither contain nor be contained in $L_\alpha$ in general. However, we have $\tilde{L}_\alpha \subseteq L_{\alpha+\zeta}$, which, plugged into the proof of Lemma 3.1, implies that the sublevel-sets filtration of $f$ is $(c\delta + \zeta)$-interleaved with $\{\tilde{L}_\alpha^\delta\}_{\alpha \in \mathbb{R}}$. The rest of the analysis of Section 3.1 carries through, with $L_\alpha$ replaced by $\tilde{L}_\alpha$ for all $\alpha \in \mathbb{R}$ and $c\varepsilon$ and $c\varepsilon''$ replaced respectively by $c\varepsilon + \zeta$ and $c\varepsilon'' + \zeta$ in Eq. (3.3) and in the rest of the proof of Theorem 3.1.

Assume now that $\mathrm{d}_\mathbb{X}$ is replaced by some distance $\tilde{\mathrm{d}}_\mathbb{X}$ satisfying the inequalities of (ii), and let $\varepsilon' = \lambda\delta$ and $\varepsilon'' = \lambda\delta'$. Under this assumption, we prove in the full version of the paper [7] that the following sequence of inclusions holds for all values $\alpha \in \mathbb{R}$:

$$C_\varepsilon(L_\alpha) \subseteq \tilde{R}_\delta(L_\alpha) \subseteq C_{\varepsilon'}(L_\alpha) \subseteq \tilde{R}_{\delta'}(L_\alpha) \subseteq C_{\varepsilon''}(L_\alpha). \tag{3.7}$$

Letting $G_\alpha = \tilde{R}_\delta(L_\alpha)$ and $G'_\alpha = \tilde{R}_{\delta'}(L_\alpha)$, we can then override Eq. (2.1) with Eq. (3.7) in the proof of Theorem 3.1, which gives the desired result. □

## 4 Algorithms

We only provide brief descriptions of the core algorithm (Section 4.1), of its variants (Sections 4.2, and 4.3), and of their guarantees. A thorough treatment is done in the full version of the paper [7].

**4.1 Core algorithm.** The algorithm takes as input a $n$-dimensional vector $v$, a $n \times n$ distance matrix $D$, and a parameter $\delta \geq 0$. The entries of $v$ give the function values at the data points, while the entries of $D$ give their pairwise distances. No geographic coordinates are required, so that the algorithm can virtually be applied in any metric space. For simplicity, we assume the entries of $v$ to be sorted ($v_1 \leq v_2 \leq \cdots \leq v_n$), although they are not in our implementation. The algorithm proceeds in two steps:

1. It builds two families of nested Rips complexes: $R_\delta(\{1\}) \subseteq R_\delta(\{1,2\}) \subseteq \cdots \subseteq R_\delta(\{1,2,\cdots,n\})$ and $R_{2\delta}(\{1\}) \subseteq R_{2\delta}(\{1,2\}) \subseteq \cdots \subseteq R_{2\delta}(\{1,2,\cdots,n\})$. The $i$th complex in each family is computed from the sub-matrix of $D$ spanned by the rows and columns of indices $1, \cdots, i$. The time of appearance of its simplices that are not in the $(i-1)$th complex is set to $v_i$.
2. For $k$ ranging from zero to the dimension of the $n$th complex, the algorithm computes the $k$th persistence diagram of the nested pair of filtrations $\{R_\delta(\{1,\cdots,i\}) \hookrightarrow R_{2\delta}(\{1,\cdots,i\})\}_{1 \leq i \leq n}$.

Upon termination, the algorithm returns the persistence diagrams computed at step 2.

**Quality of the output.** Observe that the filtrations built at step 1. are the same as the ones considered in Theorem 3.1, which therefore provides the following theoretical guarantee: if the data points form a geodesic $\varepsilon$-sample of some Riemannian manifold $\mathbb{X}$, with $\varepsilon < \frac{1}{4}\ \varrho_c(\mathbb{X})$, and if the input distance matrix $D$ gives the exact geodesic distances between the data points, then, for any tame $c$-Lipschitz function $f : \mathbb{X} \to \mathbb{R}$ whose values at the data points are given exactly by the input vector $v$, the $k$th persistence diagram output by the algorithm lies at bottleneck distance at most $2c\delta$ of the $k$th persistence diagram of $f$, provided that the input parameter $\delta$ satisfies $2\varepsilon \leq \delta < \frac{1}{2}\varrho_c(\mathbb{X})$.

Similar theoretical guarantees are obtained from Theorem 3.2 in cases where the entries of the input vector $v$ or matrix $D$ are noisy, provided that $2\delta$ is replaced by $\delta' \gtrsim 2\delta$ in the algorithm. This is quite useful in practice, when geodesic distances are estimated using some neighborhood graph.

**Implementation and complexity.** A useful property of the Rips complex is that its simplices are in bijection with the cliques of its 1-skeleton graph [9]. At step 1. of the algorithm, we build the 1-skeleton graph of $R_{2\delta}(\{1, \cdots, n\})$ in $O(n^2)$ time, then we enumerate all its cliques using version 1 of the Bron-Kerbosch algorithm[4] [2] in $O(Nn \log n)$ time, where $N$ is the total number of simplices of $R_{2\delta}(\{1, \cdots, n\})$. Finally, we encode the filtrations of parameters $\delta$ and $2\delta$ as two orderings on the simplices of $R_{2\delta}(\{1, \cdots, n\})$, which takes $O(N \log N)$ time. At step 2., we apply the modified persistence algorithm of [11] on our two filtrations, which raises the overall running time to $O(N^3)$.

In principle, $R_{2\delta}(\{1, \cdots, n\})$ could span the full $(n-1)$-simplex and have as many as $2^n$ simplices. In practice however, data points are often drawn from manifolds of low dimensions $m$. Then, under a uniform sampling condition, a packing argument shows that the size of the complex is at most $2^{2^m} n$, and that it even drops down to $2^{O(m^2)}n$ if a reasonable upper bound on $m$ is known [9]. This reduces the running time of the algorithm to $2^{O(m^2)}n^3$ and thus makes the approach tractable. Sampling uniformity is a stringent condition, but it is achieveable by a landmarking strategy [19].

**4.2 Extracting spatial information.** The input is the same as in Section 4.1, and we call $\mathbb{X}$ the underlying space, $f$ the unknown scalar field, and $L$ the set of data points. We want to partition $L$ into clusters, each of which consists of the points that flow down to a same minimum of $f$ when moving opposite to the gradient vector field of $f$ in $\mathbb{X}$. Stated in Morse-theoretic terms, our goal is to compute the trace in $L$ of the descending regions of the minima[5] of $f$. We assume for simplicity that the values of $f$ at the points of $L$ are all different, which is easily ensured by an infinitesimal perturbation of $f$.

After building the 1-skeleton graph $G_{2\delta}$ of $R_{2\delta}(L)$ as in Section 4.1, we proceed in two steps:

1. at each point $p \in L$ we compute a rough estimate of the direction of steepest descent of $f$, by connecting $p$ to its neighbor in $G_{2\delta}$ with lowest $f$-value (and to itself if $p$ is a local minimum); the local minima of $f$ in $G_{2\delta}$ are then promoted to the status of cluster centers, and $L$ is partitioned according to their descending regions in $G_{2\delta}$;
2. we apply the core algorithm to approximate of the 0th persistence diagram of $f$, which is then used to merge clusters of lifespan less than some input threshold $\tau$ into longer-lasting clusters.

The procedure is illustrated in Figure 1. The clusters construction at step 1. is inspired from [14], and we refer the reader to [25] for a detailed implementation in our point cloud setting. This construction is known to be quite unstable under small perturbations of $L$ or $f$, and the novelty of our approach resides in the way we use persistence to merge clusters and regain some stability at step 2. Specifically, the approximate persistence diagram consists of a set of critical pairs $(v, e)$, where $v$ is a local minimum of $f$ in $G_{2\delta}$ and $e$ is an edge of $G_{2\delta}$ that links the connected component created by $v$ in $G_{2\delta}$ to the one created by some lower minimum $u$. If the lifespan[6] of the connected component of $v$ is shorter than $\tau$, then the algorithm merges the cluster of $v$ into the cluster of $u$.

Our implementation makes a single pass through the 1-skeleton graph $G_{2\delta}$, creating and merging clusters on the fly using the modified union-find data structure of [16]. Once $G_{2\delta}$ has been built, the remaining running time is $O(NA^{-1}(N))$, where $N$ is the size of $G_{2\delta}$ and $A$ the Ackermann function.

The number of clusters output by the algorithm is guaranteed by Theorem 3.1 to coincide with the number of descending regions of minima of $f$ of lifespan at least $\tau$ in $\mathbb{X}$, provided that the 0th persistence diagram of $f$ satisfies some well-separatedness condition made explicit in the full version of the paper. In addition to this stability guarantee, it would be desirable to have an approximation result[7] bounding the distances in $\mathbb{X}$ between the computed clusters and their corresponding descending regions of minima of $f$ in $\mathbb{X}$. This question remains open for now.

**4.3 Time-varying functions.** Suppose now that $\mathbb{X}$ and $L$ remain fixed while $f$ varies with time. More precisely, our input is now a finite sequence $(v_1, \cdots, v_s)$ of $n$-dimensional vectors, each of which is a *snapshot* of $f$ at the points of $L$ and at a certain time $t_i \leq t_{i+1}$. For convenience, we let $f_i : \mathbb{X} \to \mathbb{R}$ denote $f$ at time $t_i$, and we call $c_i$ its Lipschitz constant. Our aim is to approximate the persistence diagrams of the $f_i$.

The naive approach applies the core algorithm at each time step separately. By Theorem 3.1, the output persistence

[4] This algorithm was designed to enumerate maximal cliques, but it actually enumerates all cliques.

[5] Symmetrically, we can approximate the ascending regions of the maxima of $f$ by using $-f$ in our computations.

[6] Defined as the difference between the times at which $e$ and $v$ appear in the 1-skeleton graph $G_{2\delta}$.

[7] Our experimental results, detailed in Section 5, suggest the existence of such a guarantee.

| data set | dimension | # vertices | # edges | Rips graph (sec.) | clustering (sec.) | total (sec.) |
|---|---|---|---|---|---|---|
| crater | 2 | 1,048 | 7,095 | 0.01 | 0.00 | 0.01 |
| torus | 3 | 2,034 | 7,650 | 0.01 | 0.00 | 0.01 |
| four Gaussians | 2 | 6,354 | 51,946 | 0.07 | 0.02 | 0.09 |
| hand | 2 | 19,470 | 158,395 | 0.27 | 0.05 | 0.32 |
| double spiral | 2 | 114,563 | 2,116,035 | 2.43 | 0.61 | 3.04 |

Table 1: *Timings on an Intel Core 2 Duo T7500 @ 2.20GHz with 2GB of RAM. The pacing phase of the approach is the construction of the Rips graph, which performs a linear number of proximity queries, implemented using the C++ library ANN [27]. The* clustering *phase comprises both steps of the algorithm of Section 4.2, which are performed simultaneously.*

diagrams at time $t_i$ approximate the ones of $f_i$ within a bottleneck distance of $2\delta c_i$, under a sampling condition and with a choice of parameter $\delta$ that are time-independent. The total running time is $O(sN^3)$, where $N$ is the number of simplices of the Rips complex $R_{2\delta}(L)$.

A more elaborate variant inspired from [12] exploits the fact that $R_{2\delta}(L)$ remains fixed throughout the process, and that the sole orderings of its simplices corresponding to the filtrations of parameters $\delta$ and $2\delta$ have to be updated between $t_{i-1}$ and $t_i$. For each filtration, we permute the order of the simplices of $R_{2\delta}(L)$ using *insertion sort*, which decomposes the permutation into a sequence of $m_i$ simple transpositions of simplices. This sequence is given as input to the *vineyards* algorithm [12], which updates the persistence diagrams in $O(m_iN)$ time. In the worst case, this bound is no better than $O(N^3)$. However, it is sensitive to the timewise variations of $f$, which can be small. A more formal discussion on this point can be found in the full version of the paper.

## 5 Applications and discussion

We illustrate the relevance and generality of our approach through three specific applications. For each application, we describe the context and present some experimentation validation. Timings on our data sets are given in Table 1.

**5.1 Sensor networks.** Our approach was originally designed with the sensor network framework in mind, where physical quantities such as temperature or humidity are measured by a collection of communicating sensors, and where the goal is to answer qualitative queries such as how many significant *hot spots* are being sensed. Purely geometric approaches cannot be applied in this setting, since geographic location is usually unavailable. Rough pairwise geodesic distances however are available, in the form of graph distances in the communication network. With this data at hand, the algorithms of Section 4 can find the number of hot spots, provide an estimation of their prominance and of their size in the network, and track them as the quantity being measured changes. As a by-product, they also compute the homology of the underlying domain as the linear span of the infinitely-persistent homological features in the output barcodes. The computations are done in a centralized way, after a data aggregation step. A sample result is shown in Figure 1.

**5.2 Clustering.** Clustering attempts to group points by assuming they are drawn from some unknown probability distribution. Our approach is inspired by Mean-Shift clustering [13]. Given an input point cloud $L$, we use a simple density estimator to approximate the local density at the points of $L$. As Figure 2 shows, our estimator can be quite noisy. However, our emphasis is not on accurate density estimation, but rather on clustering with noisy density estimates. Our estimator is provided together with $L$ as input to the algorithm of Section 4.2, which clusters the points of $L$ according to the basins of attraction of the local maxima of the estimator in the Rips graph $G_{2\delta}$ built over $L$. Due to the noisy nature of the estimator, we get a myriad of small clusters before the merging phase. The novelty of our approach is to provide visual feedback to the user in the form of an approximate persistence barcode of the estimator, from which the user can choose a relevant merging parameter $\tau$. For instance, the example of Figure 2 is highly non-linear and noisy, yet the barcode clearly shows two long intervals, suggesting that there are two main clusters.

Another important feature of our approach is to make a clear distinction between the merging criterion, governed by $\tau$ and based solely on persistence information, and the approximation accuracy of the basins of attraction of the maxima, governed by the Rips parameter $\delta$ and based solely on spatial information. In the example of Figure 2, reducing $\delta$ while keeping $\tau$ fixed enabled us to separate the two spirals from the background while keeping them separate and integral.

**5.3 Shape Segmentation.** The goal of shape segmentation is to partition a given shape into *meaningful* segments, such as fingers on a hand. This problem is ill-posed by nature, as meaningfulness is a subjective notion. Given a sampled shape $\mathbb{X}$, our approach is to apply the algorithm of Section 4.2 on some *segmentation function* $f : \mathbb{X} \to \mathbb{R}$ derived from the geometric features of $\mathbb{X}$. The output is a partition of the point cloud into clusters corresponding roughly to the basins of attraction of the significant peaks of $f$. Thus, we cast the segmentation problem into another problem, namely the

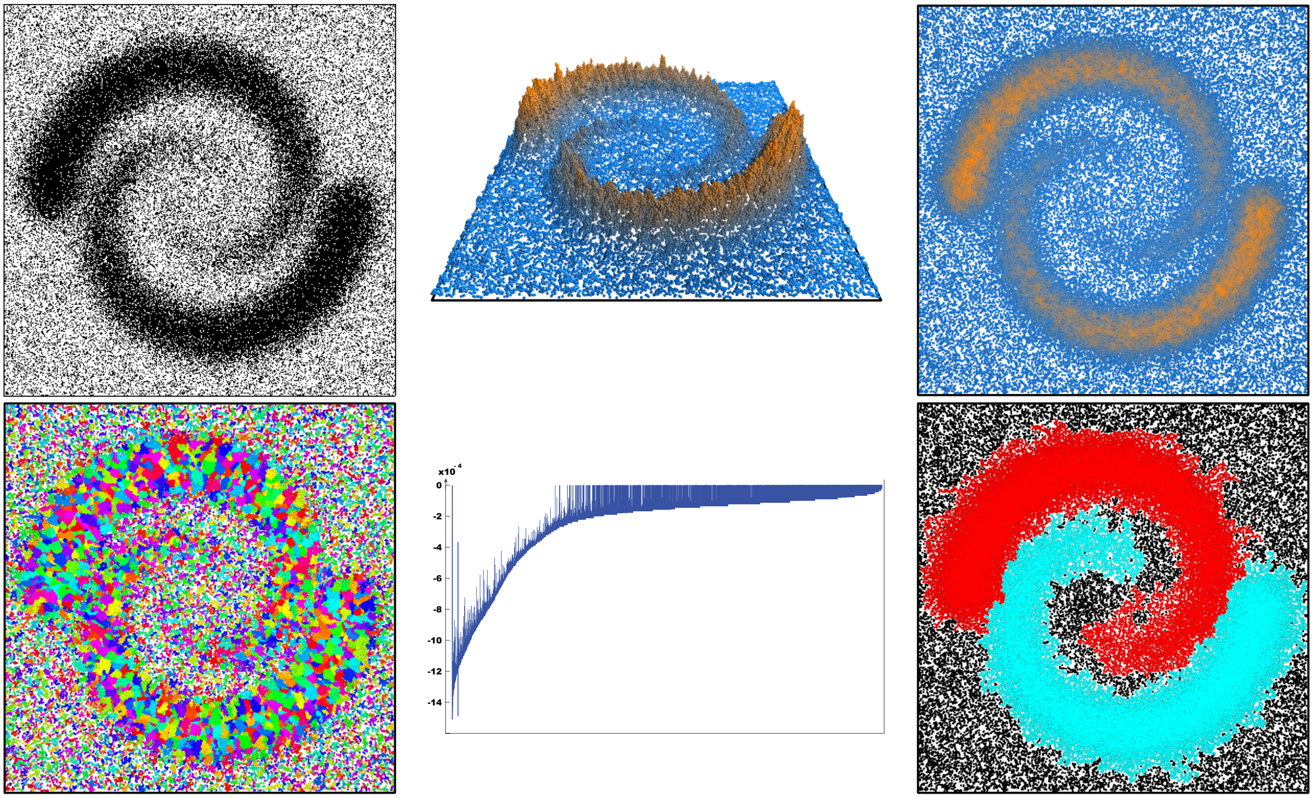


Figure 2: *A result in clustering. The top row shows the input provided to the algorithm of Section 4.2: the data points (left), or rather their pairwise Euclidean distances, and the estimated density function $f$ (center and right). The 3-d view of $f$ illustrates how noisy this function can be in practice, thereby emphasizing the importance of our robustness result (Theorem 3.2). The bottom row shows the estimated basins of attraction of the peaks of $f$, before (left) and after (right) merging non-persistent clusters. The 0-dimensional persistence barcode of $(-f)$ (center) contains two prominent intervals corresponding to the two main clusters. Since the estimated density is everywhere non-negative, the barcode has been thresholded at $0$. Thus, intervals reaching $0$ correspond to independent connected components in the Rips graph. Among those, the ones that appear lately are treated as noise and their basins of attraction shown in black, since their corresponding density peaks are low.*

one of finding a relevant segmentation function $f$ for a given class of data.

In our experiments, we defined $f(x)$ to be the diameter of the set of sample points on the boundary of the shape that are closest to $x$, normalized by their distance to $x$. We chose this particular function as a demonstration, but our method can be applied virtually with any segmentation function. The barcode computed by the algorithm measures the stability of the various segments, as illustrated in Figure 3. As such, it provides feedback on the relevance of the chosen segmentation function on the considered class of shapes.

## 6 Final remarks

The potential of our approach stems from the observation that many problems can be reduced to the analysis of some scalar field defined over a given point cloud data. With the theoretical and algorithmic tools developed in this paper at hand, the user can cast each of these problems into the one of finding the scalar field that is most suitable for his particular purpose. Thus, clustering is turned into a density estimation problem, while shape segmentation is turned into finding a relevant segmentation function for a given class of shapes. Many application-specific questions arise from this paradigm, which we do not pretend to solve in the paper. Some of them, related to the above scenarios, will be addressed in subsequent work.

**Acknowledgements.** The authors wish to thank the anonymous referees for their helpful comments. This work was carried out within the Associate Team *TGDA: Topological and Geometric Data Analysis* involving the Geometrica group at INRIA and the Geometric Computing group at Stanford University. It was supported by ANR project 05-JCJC-0246 (*GeoTopAl*), DARPA grant HR0011-05-1-0007, and NSF grants ITR-0205671, FRG-0354543, and CCF-0634803.

## References

[1] P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. Topological hierarchy for functions on triangulated surfaces. *IEEE Trans. Vis. Comput. Graphics*, 10:385–396, 2004.

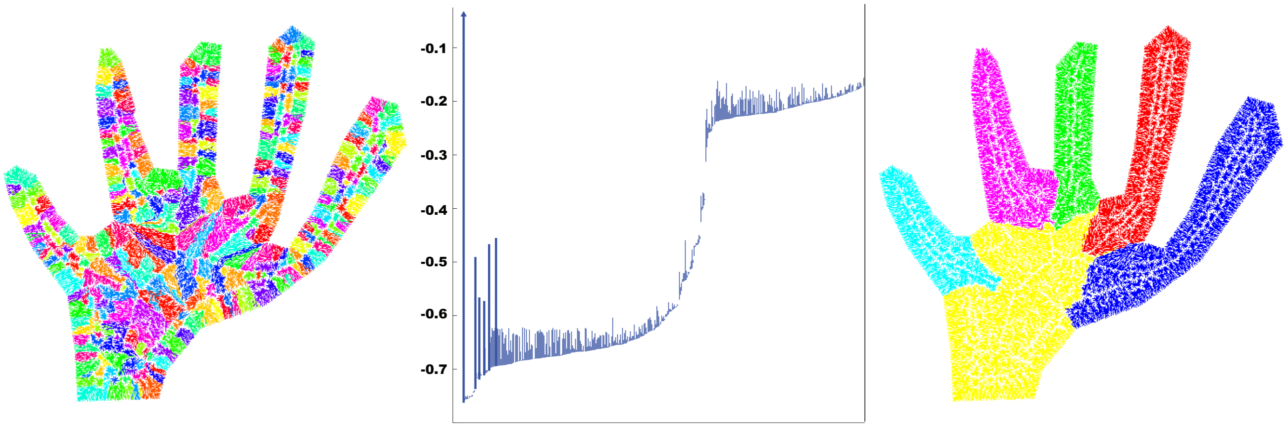


Figure 3: *Segmentation result on a sampled hand-shaped 2-D domain. The segmentation function is the (normalized) diameter of the set of nearest boundary points. The barcode shows six long intervals, corresponding to the palm of the hand and to the five fingers. The results before and after merging non-persistence clusters are shown respectively to the left and to the right of the barcode.*

[2] C. Bron and J. Kerbosch. Finding all cliques of an undirected subgraph. *Commun. ACM*, 16:575–577, 1973.

[3] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas. Persistence barcodes for shapes. *Interational Journal of Shape Modeling*, 11:149–187, 2005.

[4] M. Do Carmo. *Riemannian Geometry*. Birkhäuser, Boston, Basel, Berlin, 1992.

[5] F. Cazals, F. Chazal, and T. Lewiner. Molecular shape analysis based upon the Morse-Smale complex and the Connolly function. In *Proc. 19th Annu. ACM Sympos. Comput. Geom.*, pages 237–246, 2003.

[6] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. Research Report 6568, INRIA, November 2008.

[7] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. Research Report 6576, INRIA, July 2008.

[8] F. Chazal and A. Lieutier. Stability and computation of topological invariants of solids in $\mathbb{R}^n$. *Discrete Comput. Geom.*, 37(4):601–617, 2007.

[9] F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Proc. 24th ACM Sympos. Comput. Geom.*, pages 232–241, 2008.

[10] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *Proc. 21st ACM Sympos. Comput. Geom.*, pages 263–271, 2005.

[11] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Persistent homology for kernels and images. Preprint, 2008.

[12] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Proc. 22nd Sympos. on Comput. Geom.*, pages 119–126, 2006.

[13] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[14] T. K. Dey and R. Wenger. Stability of critical points with interval persistence. *Discrete Comput. Geom.*, 38:479–512, 2007.

[15] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse complexes for piecewise linear 2-manifolds. In *Proc. 17th Annu. Sympos. Comput. Geom.*, pages 70–79, 2001.

[16] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[17] H. Edelsbrunner, D. Morozov, and V. Pascucci. Persistence-sensitive simplification of functions on 2-manifolds. In *Proc. 22nd Sympos. on Comput. Geom.*, pages 127–134, 2006.

[18] J. Gao, L. Guibas, S. Oudot, and Y. Wang. Geodesic Delaunay triangulation and witness complex in the plane. Full version, partially published in *Proc. 18th ACM-SIAM Sympos. on Discrete Algorithms*, pages 571–580, 2008. Full draft available at: `http://graphics.stanford.edu/projects/lgl/papers/ggow-gtwcp-08/ggow-gdtwcp-08-full.pdf`.

[19] L. G. Guibas and S. Y. Oudot. Reconstruction using witness complexes. In *Proc. 18th Sympos. on Discrete Algorithms*, pages 1076–1085, 2007.

[20] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. Topology-based simplification for feature extraction from 3d scalar fields. In *Proc. IEEE Conf. Visualization*, pages 275–280, 2005.

[21] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. A topological approach to simplification of three-dimensional scalar fields. *IEEE Trans. Vis. Comput. Graphics*, 12(4):474–484, 2006.

[22] A. Hatcher. *Algebraic Topology*. Cambridge Univ. Press, 2001.

[23] John W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.

[24] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[25] X. Zhu, R. Sarkar, and J. Gao. Shape segmentation and applications in sensor networks. In *Proc. INFOCOM*, pages 1838–1846, 2007.

[26] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.

[27] `http://www.cs.umd.edu/~mount/ANN/`.